INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Bayesian HMM based x-vector clustering for Speaker Diarization

Mireia Diez[1], *Lukáš Burget*[1], *Shuai Wang*[1,2], *Johan Rohdin*[1], *Honza Černocký*[1]

[1]Brno University of Technology, Faculty of Information Technology, IT4I Centre of Excellence
[2]Speechlab, Department of Computer Science and Engineering, Shanghai Jiao Tong University

{mireia,burget,rohdin,cernocky}@fit.vutbr.cz; feixiang121976@sjtu.edu.cn

Abstract

This paper presents a simplified version of the previously proposed diarization algorithm based on Bayesian Hidden Markov Models, which uses Variational Bayesian inference for very fast and robust clustering of x-vector (neural network based speaker embeddings). The presented results show that this clustering algorithm provides significant improvements in diarization performance as compared to the previously used Agglomerative Hierarchical Clustering. The output of this system can be further employed as an initialization for a second stage VB diarization system, using frame-wise MFCC features as input, to obtain optimal results.

Index Terms: Speaker Diarization, Variational Bayes, HMM, x-vector, DIHARD

1. Introduction

Diarization is the task of determining speaker turns in an audio conversation. That is, given an audio conversation, a diarization system must infer the number of speakers speaking in the audio and find when each of them is speaking. In last years, the interest for diarization tasks has grown in the community. After a rather long break since the last editions of diarization evaluations [1], new diarization challenges are being organized like the DIHARD series [2, 3], the Fearless Step Challenge [4] or the VoxCeleb Evaluation [5]. Even the last edition of the NIST Speaker Recognition Evaluation included conditions that required diarization systems [6].

Driven by the success of x-vector embeddings in the related Speaker Recognition (SR) task [7], x-vector based diarization works keep emerging in the community [8, 9, 10]. Usually, these systems make a coarse segmentation of the input conversation into 1.5-2 second chunks and extract an x-vector for each of the segments, which are then clustered using Agglomerative Hierarchical Clustering (AHC) [8] or other clustering methods [9]. The x-vector based systems have proven to be very robust for the diarization task. Nevertheless, the segmentation step needed for the x-vector extraction sets the granularity (or time resolution) of the system outputs, which calls for an extra re-segmentation step to refine the timing of speaker changes.

In our previous works, we have presented the diarization system based on Variational inference in Bayesian Hidden Markov Model (HMM) with Eigenvoice priors [11, 12]. This diarization system (which is often referred to as *Variational Bayes (VB) diarization*) has been established as the state-of-the-art method. In fact, in the last DIHARD challenge [2], which was designed to foster research on "hard" diarization conditions, the two best performing systems were based on a cascade of two diarization systems: In the first stage, x-vectors were clustered using AHC with Probabilistic Linear Discriminant Analysis (PLDA) metric [13]. In the second stage, the output of the x-vectors based system was used as an initialization for our Bayesian HMM based system, which offers a principled way of robustly clustering the standard speech features (MFCC) with better time resolution.

The x-vector based initialization allows to benefit from the discriminative power of the NNs based embeddings. However the simple AHC used in previous works for the x-vector clustering might be sub-optimal, as it cannot recover from hard decision mistakes made during the clustering process. In this paper, we propose to use the Bayesian HMM also for the x-vector clustering. The VB inference used for this model avoids making any hard decisions. Instead, it iteratively refines the soft probabilistic alignment of x-vectors to speakers and re-estimates the speaker specific x-vector distributions (i.e. speaker models). The inference is able to determine the number of speakers in the recording. It also takes into account the uncertainty in the speaker model estimates (i.e. we cannot be very certain about speaker distributions estimated from only few x-vectors), which also contributes to the robustness of the resulting x-vector clustering. Experiments to show the effectiveness of the clustering method are carried out on the DIHARD dataset [14].

The Bayesian HMM used for the x-vector clustering (i.e. the first stage of the cascade described above) can be seen as a simplified version of our original Bayesian HMM [11] as applied to the MFCC features in the second stage: In the original model, the speaker distributions were modeled using i-vector-like subspace constrained Gaussian Mixture Models (GMMs) [15]. The simplified version presented here derives the speaker models directly from a PLDA model pretrained on x-vectors (i.e. speakers specific x-vector distributions are assumed to be Gaussian). The work [16] also used a Bayesian GMM to cluster i-vectors, but it did not used pretrained PLDA to facilitate the speaker clustering. Our model can also be seen as a simpler and more practical variant of the VB-GMM introduced in [17], where we further introduce scaling parameters controlling the VB inference that are important for good performance. Further, our model is a VB-HMM which allows for modeling speaker turn duration. Also, we use it to cluster x-vectors rather than i-vectors.

Finally, another motivation for replacing the AHC with the Bayesian HMM clustering is that having both stages of our diarization system implemented using the same framework will open up the possibility to integrate both stages into single probabilistic model, which could benefit from jointly modeling both the discriminative x-vectors and fine grained MFCC features.

2. The VB diarization model

This section provides an overview of the Bayesian HMM diarization Model introduced in our previous works [11, 12]. A short summary of the complete model is given to provide the foundation for the next section, where a simplified variant of the model suitable for the x-vector clustering is proposed. For a more complete description of the model, we refer the reader to [11, 12].

http://dx.doi.org/10.21437/Interspeech.2019-2813

The VB diarization model is a Bayesian HMM, where the states corresponds to speakers, the transition between states represent the speaker turns, and the speaker distributions are modeled by GMMs with parameters constrained by eigenvoice priors like in i-vector [15] and JFA [18] models. This allows us to represent the distribution of speaker s by means of a low dimensional latent vector $\mathbf{y}_s$.

The HMM has a one-to-one correspondence between the HMM states and speakers.[1] The HMM model is ergodic (i.e. transitions between all speakers are possible), where the (initial) number of states is chosen to be at least the highest number of speakers we would expect to appear in a conversation.

The HMM topology and transition probabilities model the speaker turn durations (see Figure 1 for an example with 3 speaker states): P_{loop} is a tunable parameter, which corresponds to the probability of staying on the same state (speaker). For high frame rate features such as MFCCs, this is typically set to a value close to one to naturally model the speaker turns. For each frame, we leave the current state with probability $1-P_{\text{loop}}$ and we transition to new state of speaker s with probability π_s. The probabilities π_s also control the selection of the initial HMM state. The probabilities π_s are inferred from the input conversation using the iterative VB updates and, thanks to the automatic relevance determination (ARD) [19] principle, the values π_s for any redundant speaker states tend to converge to zero. This allows to infer the number of speakers in the input conversation.

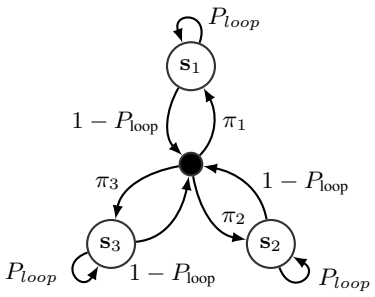

Figure 1: *HMM model for 3 speakers with 1 state per speaker, with a dummy non-emitting (initial) state.*

Each speaker (or HMM state) specific distribution is modeled using a subspace constrained GMM as in the i-vector extractor model [15] used for speaker recognition: The model assumes that the speaker specific GMMs are all related to a single UBM model, and they share the component weights w_c^{ubm} and covariance matrices Σ_c^{ubm}. For a speaker s, its super-vector of concatenated means $\boldsymbol{\mu}_s = [\boldsymbol{\mu}_{s1}^T \boldsymbol{\mu}_{s2}^T \dots \boldsymbol{\mu}_{sC}^T]^T$ is constrained to live in a low-dimensional subspace

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

around the origin given by the UBM mean super-vector $\boldsymbol{\mu}^{ubm}$, spanned by the Total Variability matrix $\mathbf{V}$. The matrix $\mathbf{V}$ is pre-trained on a large amount of data in the same way as for the i-vector extraction [15] and is shared among all the speaker models. To estimate the speaker specific distributions, we only need to infer the low dimensional vectors $\mathbf{y}_s$, which can be seen as coordinates of the speaker models in the low-dimensional space. This allows us to robustly estimate speaker models even from small amounts of speech. In our Bayesian model, $\mathbf{y}_s$ are treated as latent random variables with standard normal prior

$$p(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s; \mathbf{0}, \mathbf{I}). \tag{2}$$

Let us state the diarization problem formally for our model. Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_S\}$ be the set of all speaker vectors. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ be the sequence of observed vectors (i.e. MFCC features) and $\mathbf{Z} = \{z_1, z_2, ..., z_T\}$ the corresponding sequence of latent variables defining the hard alignment of speech frames to HMM states. To address the SD task, the speaker distributions $\mathbf{y}_s$ and latent variable $\mathbf{Z}$ are jointly estimated (together with the π_s probabilities) given the input sequence $\mathbf{X}$. In order to find the most likely alignment $\mathbf{Z}$, we need to infer the posterior distribution $p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) d\mathbf{Y}$. Since the evaluation of this integral is intractable, we use Variational Bayes and mean field approximation [19] to approximate the posterior $p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \approx q(\mathbf{Z}, \mathbf{Y}) = q(\mathbf{Z})(\mathbf{Y})$ and the solution to the diarization task is taken as the most likely alignment according to inferred distribution $q(\mathbf{Z})$. We search for such $q(\mathbf{Z}, \mathbf{Y})$ that minimizes the Kullback-Leibler divergence $D_{KL}(q(\mathbf{Z}, \mathbf{Y}) \| p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}))$. This is equivalent to maximizing the standard VB objective – the Evidence Lower Bound Objective (ELBO), which takes the following form for our model:

$$\begin{aligned}\hat{\mathcal{L}}\left(q(\mathbf{X}, \mathbf{Y})\right) = F_A E_{q(\mathbf{Y},\mathbf{Z})}\left[\ln p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})\right] \\ + F_B E_{q(\mathbf{Y})}\left[\ln \frac{p(\mathbf{Y})}{q(\mathbf{Y})}\right] + E_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})}\right].\end{aligned} \tag{3}$$

The term $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ is the likelihood of the observation sequence $\mathbf{X}$ evaluated using the HMM model described above given fixed speaker models $\mathbf{Y}$ and alignment $\mathbf{Z}$. The prior on possible alignments $p(\mathbf{Z})$ is defined as in standard HMMs in terms of the transition probabilities and the prior on the speaker modes $p(\mathbf{Y}) = \prod_s p(\mathbf{y}_s)$.

The VB inference iteratively estimates the distributions $q(\mathbf{Y})$ and $q(\mathbf{Z})$ and parameters π_s using the update formulas, which can be derived from ELBO (3) using (variational) derivatives as detailed in [12]. For the space constraints, we do not provide the update formulas here. Instead, we kindly refer the reader to [11, 12] or directly to our python implementation of this inference [20].

In (3), we modified the ELBO by scaling the first two terms by constant factors F_A and F_B. The theoretically correct values for these factors are $F_A = F_B = 1$. However, choosing different values gives us finer control over the inference, which can be used to improve diarization performance: The *Acoustic scaling factor* F_A is introduced to counteract the assumption of statistical independence between observations by scaling down the log likelihood of the observations.[2] The *speaker regularization coefficient* F_B weighs the second term of the ELBO in (3), the Kullback-Leibler divergence between the approximate speaker posterior and the speaker prior $D_{KL}(q(\mathbf{Y}) \| p(\mathbf{Y}))$. This term can be seen as a regularization term penalizing the complexity of the speaker models, which allows us to control the number of speakers inferred from the input utterance (a high value of F_B results in the VB inference dropping more speaker models). For a more detailed interpretation and analysis of these parameters we refer the reader to [12].

3. Simplified Variational Bayes Diarization

The Bayesian HMM described in the previous section was designed to be applied on the fine-grained MFCCs. In this section we will focus on using the same model for clustering x-vectors.

[1] An extension with multiple states per speaker is described in [11], which allows for minimum speaker turn duration modeling.

[2] Note that in [11] this factor corresponds to the *statScale* parameter

Since the PLDA was found effective for modeling x-vectors in speaker recognition, we will use the same model for modeling the speaker distributions in our VB HMM model. Let us consider the simplified PLDA model [21], which assumes that the distribution of speaker means

$$\mathbf{m}_s = \mathbf{m} + \mathbf{V}\mathbf{y}_s, \tag{4}$$

where $\mathbf{y}_s$ is again a latent vector with standard normal prior and where the speaker specific distribution of x-vectors

$$p(\mathbf{x}_t|\mathbf{y}_s) = \mathcal{N}(\mathbf{x}_t; \mathbf{m}_s, \mathbf{\Sigma}_{wc}), \tag{5}$$

where $\mathbf{\Sigma}_{wc}$ is the within speaker covariance matrix shared by all speaker models.

This model is equivalent to the one used to model speaker distributions in our VB diarization system under the assumption that the speaker specific distribution can be modeled using only a Gaussian distribution (in contrast to the GMM). In other words, we can use exactly the same model and the same inference as described in the previous section, with the following simplified setup: the number of GMM components for the speaker models is set to one. The PLDA mean $\mathbf{m}$ is used in place of the UBM mean $\boldsymbol{\mu}^{ubm}$. The $\mathbf{V}$ matrix from the PLDA model is used as the eigenvoice matrix, and the covariance matrix of the single UBM component is set to $\mathbf{\Sigma}_{wc}$.

Further, since the x-vectors come from a coarse segmentation in our experiments, we found that the optimal value for $P_{loop} = 0$, which effectively degrades the Bayesian HMM into a Bayesian GMM. Nevertheless, treating the model as an HMM may be useful in the future when working with x-vectors of higher frame-rate.

4. x-vector extraction

Our x-vector system is built based on the Kaldi recipe [22], but with some modifications on the data preparation. Voxceleb [23] training and Voxceleb2 [24] development sets are combined to generate the training set for the x-vector extractor. The data augmentation procedure described in [7] is adopted to increase the amount and diversity of the training data. The final training set contains around 6 million speech sessions from 7146 speakers. The utterances are further cut into segments of 2s for the neural network training. 64-dimensional filter banks (Fbanks) are used for the x-vector system, with an energy-based voice activity detector (VAD) to remove silence.

The standard time-delay neural network (TDNN) described in [7] is employed, which consists of 5 time delay layers and two dense layers. The embeddings are extracted after the first dense layer with a dimensionality of 512.

The PLDA model is trained using the same Voxceleb training set as the NN, but with the length of segments for x-vector extraction set to 3s. x-vectors are centered and whitened using DIHARD 2018 development data, and then length normalized.

The x-vectors for the speaker diarization clustering stage are extracted with a 1.5s sliding window and a window shift of 0.75s, then centered and whitened and length normalized in the same way as for the PLDA training data.

5. Experimental setup

5.1. Evaluation datasets

The experiments are evaluated on the DIHARD I dataset. This dataset was created for the DIHARD challenge [2], the first of a yearly series of challenges designed to foster research on diarization in hard conditions. The dataset includes utterances coming from several sources (YouTube, court rooms, meetings, etc.) [14]. The corpus consists of 164 development and 172 evaluation recordings, containing around 14h and 17h of speech, respectively.

5.2. Baseline AHC

The baseline clustering approach follows the method described in [25] as implemented in [22]. The method is based on AHC, which uses PLDA metric as the similarity measure between the embeddings. The full clustering process goes as follows: For each input conversation, x-vectors are extracted for 1.5s windows with 0.75s overlap. A conversation dependent PCA is estimated and the x-vectors are projected so as to keep only a 10% of the total variability. PLDA similarity measure (i.e. log likelihood ratio between same-speaker and different-speaker hypotheses) is computed between the projected x-vectors. For each clustering step, the weighted average of the PLDA scores is used as similarity measure between clusters. The threshold used as stopping criteria for the AHC is fine-tuned on the development set.

5.3. Setup for the VB Diarization systems

5.3.1. VB clustering of x-vectors

In [11, 12], we have also introduced a parameter controlling the VB inference called *downsamplingFactor*, which effectively constrains the model in such a way that each consecutive sequence of *downsamplingFactor* observations has to be generated by the same HMM state. Given that the x-vectors already come from a coarse segmentation, this parameter is set to one (effectively not used) for the x-vector clustering. As described in Section 3, the P_{loop} parameter is set to 0 for this version of the model. The PLDA model used in this clustering is exactly the same trained for the AHC system.

To initialize the inference, the speaker labels were generated by simply grouping N number of consecutive x-vectors and assigning the same speaker label to each group. The N value, as well as the other parameters of the VB algorithm, namely F_A and F_B, were tuned for best performance on the development set.

5.3.2. Framewise MFCC VB diarization

For the VB diarization system used in the second stage of the cascade of systems, the following configuration is used: The Weighted Prediction Error (WPE) [26] method was used to remove late reverberation from the audio signal. As features, standard MFCCs are extracted from 16kHz speech. We employ 19 MFCC plus Energy plus first order deltas. Neither mean nor variance normalization are applied in the feature extraction. We use a gender-independent UBM-GMM, with 1024 diagonal-covariance Gaussian components. The dimensionality of the speaker latent variable $\mathbf{y}_s$ is 400. The UBM-GMMs and the total variability matrix are trained using the VoxCeleb2 dataset, which amounts to 2025h of speech [24].

The parameters for the diarization system, namely: *downsamplingFactor*, P_{loop}, F_A and F_B were tuned for best performance on the development set.

5.4. Evaluation metric

Diarization Error Rate (DER) as defined by NIST [1] is used to evaluate the system. As is the standard practice, we use the oracle speech activity labels – we drop the silence parts from the signal– so that only speaker errors are accounted for in the DER,

(missed speech and false alarm speech errors are not taken into account). We evaluate the system with no collar and we evaluate the overlap speech regions, as it was done in the DIHARD challenge [2].

6. Results

6.1. Clustering of x-vectors (first stage)

Table 1 shows results attained with the different clustering approaches on the DIHARD development and evaluation sets. Note that the results for the dev set are overoptimistic, as the threshold for stopping the AHC (for AHC) and the parameters of the VB inference (for VB clustering) are optimized on the same development set. The numbers show that the simplified VB performs significantly better than the AHC approach.

Table 1: *DER results attained with the different clustering methods on the dev and eval sets of the DIHARD I dataset.*

Clustering	Dev	Eval
AHC	18.88	25.23
VB clustering	17.10	24.26

To further evaluate the quality of both clusterings, we also analyze the number of speakers found per recording. Figures 2 and 3 show the histograms for the real number of speakers per conversation on the dev and eval sets, and the histograms for the number of speakers found by both clustering approaches.

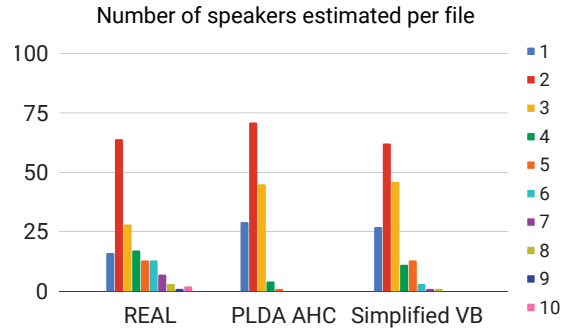

Figure 2: *Histogram of the real number of speakers per recording on the DIHARD development dataset, and the histograms for the number of speakers found with the clustering approaches*

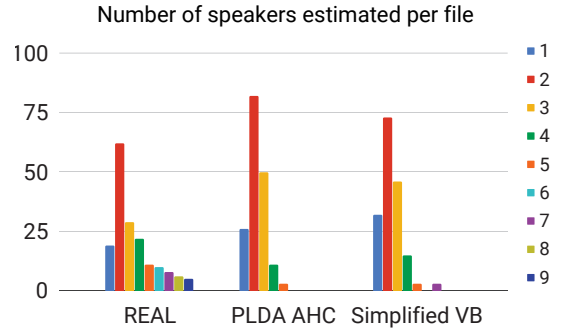

Figure 3: *Histogram of the real number of speakers per recording on the DIHARD evaluation dataset, and the histograms for the number of speakers found with the clustering approaches*

The figures reveal that both systems tend to over-cluster, ending with less number of speakers per conversation than what the real distribution shows. When analyzing these results, it was found that for some of the very challenging domains (i.e. SEEDLINGS, VAST) the systems perform so *poorly* that better DER rates are often obtained when reporting a single speaker per conversation. This suggests that the over-clustering tendency might come as a side effect of tuning the systems for optimal DER. Nevertheless, the over-clustering is more pronounced for the PLDA AHC than for the simplified VB clustering approach, which denotes that the latter uses a better criterion for inferring the number of speakers per conversation.

6.2. Framewise MFCC VB diarization (second stage)

For the second stage of our diarization pipeline, we use the output of the previous clustering as initialization for the VB diarization model. To build the input, the assignment of x-vectors to speakers obtained from the first clustering stage is taken, and upsampled to build a matrix of frame-wise speaker assignments.

When trying to optimize results with the final VB diarization, a similar trend as the one described in [25] was observed: when letting the VB diarization fully converge, the DER would get worse. This denotes that the x-vectors are in fact more discriminative as features than the frame by frame MFCCs used as input for the VB diarization. As suggested in the mentioned work [25], early stopping was employed, performing a single iteration of the VB diarization.

Table 2: *DER results attained with the different clustering methods on the dev and eval sets of the DIHARD I dataset.*

Clustering	Dev	Eval
AHC + second stage VB	18.09	24.55
VB clustering + second stage VB	16.29	23.43

Table 2 shows the results obtained after the final diarization stage. The full VB diarization framework obtains, as far as we know, the best results published for the DIHARD dataset.

7. Conclusions

In this paper we have introduced a simplified version of the VB diarization system to perform fast and effective x-vector clustering. The proposed approach has proven to be useful, outperforming the state-of-the-art AHC clustering approach.

The fact that both stages of our diarization system can be now implemented using the same Bayesian HMM framework opens up the possibility to integrate both stages into single probabilistic model. Such model, which we plan to investigate in future, could benefit from jointly modeling both the discriminative x-vectors and fine grained MFCC features.

8. Acknowledgements

The work was supported by European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 748097, the Marie Sklodowska-Curie cofinanced by the South Moravian Region under grant agreement No. 665860, Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X, Czech Ministry of Interior project No. VI20152020025 "DRAPAK", Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602" and by research contract with Ericsson. Shuai is supported by the Shanghai International Science and Technology Cooperation Fund (No. 16550720300) and the China NSFC project (No. U1736202).

9. References

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

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD Challenge Evaluation Plan," https://zenodo.org/record/1199638, 2018.

[3] N. R. et. al., "The Second DIHARD Diarization Challenge: Dataset, task, and baselines." in *Proceedings of Interspeech 2016*, 2019.

[4] "The FEARLESS STEPS Challenge," http://fearlesssteps.exploreapollo.org/, accessed: 2019-03-30.

[5] "The VoxCeleb Speaker Recognition Challenge ," http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition.html, accessed: 2019-03-30.

[6] "NIST 2018 Speaker Recognition Evaluation Plan," https://www.nist.gov/sites/default/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf.

[7] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH 2017*, August 2017, pp. 999–1003.

[8] 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.

[9] A. Zhang, C. Wang, J. Paisley, Q. Wang, and Z. Zhu, "Fully supervised speaker diarization," in *Arxiv*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04719

[10] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker Recognition for Multi/Spekaer Conversatiosn using x-vectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[11] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on bayesian hmm with eigenvoice priors," in *Proceedings of Odyssey 2018, The speaker and Language Recognition Workshop*, 2018.

[12] M. Diez, L. Burget, F. Landini, and H. Černocký, "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors," *Submitted to: IEEE Transactions on Audio, Speech, and Language Processing*, 2019. [Online]. Available: http://www.fit.vutbr.cz/~mireia/Diez19_Submitted_IEEE.pdf

[13] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Proc. ˜Odyssey-10*, Brno, Czech Republic, Jun. 2010.

[14] N. Ryant and et al., "DIHARD Corpus. Linguistic Data Consortium." 2018.

[15] 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.

[16] 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.

[17] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational bayesian plda for speaker diarization in the mgb challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 667–674.

[18] 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.

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

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

[21] N. Brummer, "EM for Simple PLDA," 2010, technical report.

[22] Kaldi, "dihard_2018 v2," https://github.com/kaldi-asr/kaldi/tree/master/egs/dihard_2018/v2, [Downloaded: 2019-02].

[23] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proceedings of Interspeech 2018*, 2018, pp. 1086–1090. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1929

[25] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Interspeech*. ISCA, 2018, pp. 2808–2812.

[26] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.