BAYESIAN ANALYSIS OF SIMILARITY MATRICES FOR SPEAKER DIARIZATION

Alexey Sholokhov[2,3], *Timur Pekhovsky*[1,3], *Oleg Kudashev*[1,3], *Andrei Shulipa*[1], *Tomi Kinnunen*[2]

[1]Speech Technology Center Ltd., St. Petersburg, Russia
[2]School of Computing, University of Eastern Finland
[3] ITMO University, Russia*

{sholohov, tim, kudashev, shulipa}@speechpro.com, tkinnu@cs.uef.fi

ABSTRACT

Inspired by recent success of speaker clustering in Total Variability space we propose a new probabilistic model for speaker diarization based on Bayesian modeling of pairwise similarity scores. The recordings are represented by symmetric similarity matrices of likelihood ratio scores from probabilistic linear discriminant analysis (PLDA) trained on short-term i-vectors. We employ Bayesian approach to address the problem of unknown number of speakers in conversation. Diarization error rates on the NIST 2008 SRE telephone data indicate comparable performance with state-of-the-art eigenvoice-based diarization. But unlike the eigenvoice approach, our method finds the number of speakers automatically, making the proposed model more viable for practical applications.

Index Terms— Speaker diarization, variational Bayesian inference, similarity matrix

1. INTRODUCTION

The goal of *speaker diarization* [1, 2] is to determine how many speakers there are in a given speech recording, and to segment the recording so that each part corresponds to one speaker. There are two dominant approaches to speaker diarization. The first and the most common approach consist of two stages: speaker change point detection followed by speaker clustering. The second approach assumes that the signal is divided beforehead into short (typically half-second) segments that are directly clustered. We adopt the latter approach.

In recent years, rapid progress in text-independent speaker verification has influenced other areas of speech technology, including speaker diarization. In particular, the idea of using an *eigenvoice* (EV) prior for speaker diarization was first proposed in [3]. In that paper, as well as in the subsequent studies using a weak eigenvoice prior [4, 5], the diarization procedure is reduced to indexing short segments of speech, which circumvents the problem of detecting the speaker change points. The *weak* EV prior implies that the *maximum a posteriori* (MAP) estimate of speaker latent variable is independently calculated from short segments, causing the number of eigenvoices to be bounded from above by 10–20 [4, 5]. In contrast to these studies, in [3] latent identity variables are estimated from the full recording, which increases the number of eigenvoices to 200–300, therefore, implementing a *strong* eigenvoice prior. It yields a considerable improvement of the performance for NIST 2008 SRE telephone dialogue diarization among systems using this kind of prior.

*This work was partially financially supported by the Academy of Finland and the Government of the Russian Federation, Grant 074-U01

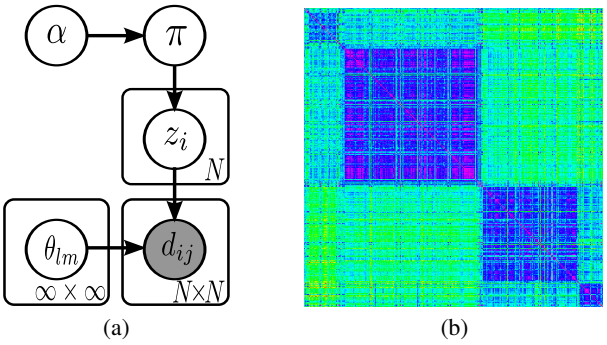

Fig. 1. (a) A graphical model of the proposed generative model (hyperparameters are not shown) (b) An example of similarity matrix calculated for conversation of four speakers. Rows and colums are ordered according to ground truth labels.

Unfortunately, as demonstrated by our experiments, the EV approach [3] has a major drawback: it is incapable for automatic selection of the number of speakers within the recording (i.e. model selection). More specifically, we have observed that the variational lower bound increases monotonically with the number of speakers in most cases. In the system of [4], the authors address this problem in a standard way by applying the classical *Bayesian information criterion* (BIC), though only at the cost of switching to a weak EV prior. As a consequence, their diarization result is behind that of [3] on 2008 NIST dialog database, but *only* if the number of speakers is known beforehand.

The idea of using strong prior information obtained from large datasets was further developed in studies that first adopted i-vectors as features for speaker diarization [6, 7, 8]. For example, [6] not only showed that diarization can be successfully carried out directly in the i-vector space, but it also achieves state-of-the-art performance on conversational telephone data from the 2008 NIST SRE corpus. Unlike [4], model selection in [8] was performed in terms of iterative variational lower bound maximization, defined for the full recording.

Proposed method and its relation to earlier studies: In this paper we propose a similarity based clustering method utilizing i-vectors. In contrast to [8], we do not carry out diarization in the i-vector space directly, but use a similarity matrix constructed from pairwise segment scores in the given recording (Fig. 1). Thus, our model does not impose an explicit model for the data distribution, but instead analyzes the structure of relations between data points. A well-known related method is *spectral clustering* [9], which analyzes the eigenstructure of a similarity matrix. For example, [7] applies

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

spectral clustering to diarization using the matrix of i-vector cosine similarities as an input.

Unfortunately, although spectral clustering methods can detect non-Gaussian clusters of complex shape, they usually fail in identifying noisy or partially overlapping clusters. These problems were addressed in [10] by introducing a probabilistic model for handling noise in data and employing maximum likelihood estimation of the model parameter. Inspired by [10], we present a generative probabilistic model for similarity matrices that tackles the problems of the spectral clustering method and uses fully Bayesian inference to automatically determine the number of clusters (speakers) in the data.

In [8], Gaussian mixture model (GMM) based clustering was applied to speaker diarization, treating i-vectors as points in Eucludean space regardless of their nature. Compared to [8], the main advantage of the proposed approach is that the model uses additional prior knowledge from learned speaker verification metric. We adopt the current state-of-the-art speaker verification model, *probabilistic linear discriminant analysis* (PLDA) [11], as a similarity measure between pairs of speech segments.

Finally, our model is related to the problem of detecting communities in social networks. Typically the graph of such a network is represented by an adjacency matrix whose elements can be either binary (presence of the link) or real-valued (strength of the connections). Authors of [12] developed a generative model for such matrices in order to find communities (densely-connected sub graphs of the original graph) in sparsely connected networks.

2. PRELIMINARIES

2.1. Total Variability

Contemporary state-of-the-art speaker recognition systems operate in *total variability space* based on a pre-trained *universal background model* (UBM) with factor analysis prior on mean supervectors [13]. The total variability model is given by,

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{T}\boldsymbol{w},$$

where $\boldsymbol{x}$ is mean supervector, $\boldsymbol{\mu}$ is a constant offset taken from the UBM, $\boldsymbol{T}$ is a rectangular matrix which defines the total variability subspace and $\boldsymbol{w}$ is a low-dimensional latent vector with standard normal distribution [13]. The maximum a posteriori estimate of $\boldsymbol{w}$ is known as the *i-vector* of the utterance.

2.2. Probabilistic Linear Discriminant Analysis (PLDA)

Given a collection of i-vectors $\{\boldsymbol{w_1}, ..., \boldsymbol{w}_H\}$, each corresponding to one recording of the same speaker, the Gaussian PLDA model [14] assumes that the i-vectors are distributed according to:

$$\boldsymbol{w}_h = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{\epsilon}_h,$$

where $\boldsymbol{m}$ is a constant speaker- and session-independent mean, $\boldsymbol{y}$ is the latent speaker identity variable which has a standard normal prior and $\boldsymbol{\epsilon}_h$ is a Gaussian residual with zero mean and full-covariance matrix. The PLDA model provides a closed-form expression for the likelihood ratio of two alternative hypotheses: (1) both vectors belong to the same speaker and (2) the vectors belong to different speakers. We use this likelihood ratio as a similarity measure for a pair of speech segments.

2.3. Adopting PLDA for Diarization

In speaker verification, i-vectors are extracted from full-length utterance, usually of several minutes in duration. Diarization, on the other hand, is a segmentation task where scores are computed from short segments. Therefore, we train the PLDA hyperparameters using i-vectors extracted from short segments. The PLDA training utterances contain only one session per speaker. We observed a slight improvement compared with training on full utterances. Another critical design consideration is the choice of the segment length. For very short segments, i-vector estimate is noisy. For long segments, in turn, separability of the target and impostor score distributions increases, but the number of mixed segments containing speech from more than one speaker increases, too. We found the optimal length to be in the range 0.5 – 1 sec. Following [15], we apply length normalization to i-vectors to compute PLDA scores.

A similar stage in speaker diarization system has been described in [16] but, instead of PLDA, the authors used support vector machine and additional set of utterances to construct similarity matrices fed as an input to agglomerative clustering algorithm.

3. PROPOSED MODEL

3.1. A Generative Model for Similarity Matrices

If we ignore any temporal relations between feature vectors, the diarization task coincides with the clustering task. A common approach to data clustering is to construct a model in which data are generated from a mixture of probability distributions. In a conventional diarization setting, such as [3, 8], each mixture component corresponds to a speaker-specific distribution of feature vectors, extracted from a segment of the input signal. But in our proposed model, a speech segment is represented not by the feature vector but by the vector of similarity scores with all the other segments of the signal. Let d_{ij} be a similarity score (PLDA likelihood ratio) between segments i and j. Now assume that the data consists of K clusters and is represented by the similarity matrix $\boldsymbol{D}$ of size $N \times N$ with entries $\{d_{ij}\}_{i,j=1}^{N}$. This matrix includes $K(K-1)/2$ groups of impostor (between-cluster) scores and K groups of target (within-cluster) scores. If we rearrange the matrix rows and columns according to the cluster (speaker) labels, it has a block structure, where the block D_{lm} contains the scores for all pairs of segments belonging to l-th and m-th clusters. Diagonal blocks contain the targets scores, and off-diagonal blocks the impostor scores. For well-separated clusters, scores in each block tend to have similar values but at the same time differ from the values of the other blocks. This is the main assumption in our model (see Fig. 1 (b)).

To formulate our model, we define the following distribution for the elements of the similarity matrix:

$$p(d_{ij}|\boldsymbol{z}, \boldsymbol{\theta}) = \prod_{\substack{l,m=1 \\ l<m}}^{K} p(d_{ij}|\theta_{lm})^{z_i^m z_j^l + z_i^l z_j^m} \prod_{k=1}^{K} p(d_{ij}|\theta_k)^{z_i^k z_j^k}. \quad (1)$$

Here, θ_{lm} are the parameters of the score distribution for block lm and z denotes cluster assignment for the i-th segment. Parameters θ_{kk} correspond to diagonal blocks (target scores) and θ_{lm} to the impostor scores. Since we assume symmetric similarity scores, notation $z_i^l z_j^m + z_i^m z_j^l$ means that there are two possibilities for score d_{ij} to get into lm-th block: i-th segment belongs to l-th cluster and j-th segment belongs to m-th cluster, or vice versa. In summary, we have

individual distributions for each pair of disjoint clusters and one distribution for each block of within-class scores. This model does not restrict the possible form of the likelihood function (1), but, since we work with real-valued PLDA scores, Gaussian distribution seems the most natural choice. Thus, the model can be viewed as a Gaussian mixture model (GMM) with $K(K+1)/2 = K(K-1)/2 + K$ components.

3.2. Automatic Detection of the Number of Speakers

We assumed above that the number of clusters (speakers) is given. Since this is usually *not* the case, we adopt Dirichlet processes [17] which enables building a nonparametric alternative to the finite mixture model, with unbounded (theoretically infinite) number of mixture components. Infinite mixtures are commonly used in Bayesian clustering [18] due to their appealing ability to automatically determine the number of clusters and their parameters from the data. Here, we view speaker diarization as a clustering problem. The nonparametric construction of Dirichlet processes allows us to get a posterior distribution on the number of speakers in a recording, therefore eliminating the need to specify it in advance.

A *Dirichlet process*, denoted by $\mathrm{DP}(\alpha, G_0)$, is a distribution of distributions, characterized by a *base measure* G_0 and a positive scalar α known as *concentration parameter*. In this study, we use a so-called *stick-breaking* construction of DP [19], based on representation of G as an infinite mixture of atoms drawn independently from G_0. Formally, $G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$ where $\theta_i \sim G_0$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ is an infinite vector of mixing weights summing up to 1. It is constructed as follows:

$$\pi_k(v) = v_k \prod_{i=1}^{k-1} (1 - v_i), v_i \sim Beta(1, \alpha) \tag{2}$$

Since exact Bayesian inference with an infinite number of parameters is intractable, a common approach to get the approximate posterior is via *truncated* stick-breaking representation [20], in which $\pi_k(v)$ is set to zero for $k > K$ for some K. The truncation level K is not part of the prior model specification and can be freely set. Based on the truncated stick-breaking representation of DP, [21] introduced a *mean-field inference* to approximate the posterior of latent variables using a factorized variational distribution. Here, we adopt their variational Bayesian (VB) approach, due to its better scalability compared to sampling based [22] inference in terms of computation time.

3.3. Model Specification and Parameter Inference

To simplify inference, we choose conjugate exponential priors for the model parameters. We place a Normal-Gamma prior over the means and the precisions of the similarity score distributions. Further, in order to take into account the effect α on the number of active mixture components, we impose a Gamma hyperprior over the concentration parameter. The complete generative process for the model is as follows:

$$\begin{aligned}
&\alpha \sim \mathrm{Gamma}(s_1, s_2) \\
&\pi|\alpha \sim \mathrm{Stick}(\alpha) \\
&Z|\pi \sim \mathrm{Multinomial}(\pi) \\
&\lambda_{lm} \sim \mathrm{Gamma}(a_0, b_0) \\
&\mu_{lm}|\lambda_{lm} \sim \mathrm{Normal}(m_0, (\beta_0 \lambda_{lm})^{-1}) \\
&d_{ij}|Z, \mu, \lambda \sim \mathrm{Normal}(\mu_{z_i z_j}, \lambda_{z_i z_j}^{-1}),
\end{aligned}$$

where $l, m = 1, 2, \dots, \infty$, $Z = \{z_i\}_{i=1}^N$ are the speaker labels, $\{\mu_{lm}, \lambda_{lm}\} = \theta_{lm}$ are the parameters for normal distribution that generate observations d_{ij} and α is the concentration parameter. Thus, $\{s_1, s_2, a_0, b_0, m_0, \beta_0\}$ are the hyperparameters for the proposed model. Despite the large number of them, most of them do not require careful fine-tuning. $\mathrm{Stick}(\alpha)$ denotes the stick-breaking process (Eq. 2). The graphical model representation of described generative process is shown on Fig. 1 (a).

An important part of the model are the hyperparameters m_0 which are the prior mean values of the target and impostor score distributions. They allow the model to distinguish between these two types of scores. Thus, we have two different mean hyperparameters: one for the within-class distributions (m_0^{tar}) and one for the between-class distributions (m_0^{imp}). As the target PLDA scores are generally larger than the impostor scores, m_0^{tar} must be greater than m_0^{imp}. Additionally, hyperparameter β_0 controls how much the estimated means can deviate from their prior values m_0.

We can now outline the variational Bayesian updates for the proposed model. Due to the conjugate exponential prior for all the parameters, the variational posterior is expected to be of the same form as the prior [21]. The probability of assigning the i-th segment to the k-th speaker is updated as follows:

$$\gamma_i^k \propto \exp\left(\sum_{j \neq i}^{N} \sum_{m=1}^{K} \gamma_j^m \langle \log \mathcal{N}(d_{ij}|\mu_{km}, \lambda_{km}^{-1}) \rangle + \langle \log \pi_k \rangle \right)$$

,where $\langle\rangle$ denotes expectations with respect to factorized posterior distribution.

Posterior updates for the means and precisions are slightly different from the standard updates for Bayesian Gaussian mixture [23] in how the sufficient statistics for Normal-Gamma posterior are computed. For the (l, m)-th mixture component, the zero- (N), the first- (F) and the second-order (S) statistics are calculated as,

$$\begin{aligned}
N_{lm} &= \sum_{i=1}^{N} \sum_{j<i}^{N} r_{ij}^{lm} \\
F_{lm} &= \frac{1}{N_{lm}} \sum_{i=1}^{N} \sum_{j<i}^{N} r_{ij}^{lm} d_{ij} \\
S_{lm} &= \frac{1}{N_{lm}} \sum_{i=1}^{N} \sum_{j<i}^{N} r_{ij}^{lm} (d_{ij} - F_{lm})^2,
\end{aligned}$$

where $r_{ij}^{lm} = \gamma_i^l \gamma_j^m + \gamma_i^m \gamma_j^l$ for $l \neq m$ and $r_{ij}^{lm} = \gamma_i^l \gamma_j^m$ for $l = m$. Other variational updates have the standard form and can be found in [21].

Since mean-field inference is sensitive to initialization, we used evidence lower bound as a score for model selection, as explained in the next Section.

4. EXPERIMENTS

In our experiments we compare the proposed model to eigenvoice based diarization system [3]. Unfortunately, the only data available to us were the telephone recordings from the NIST SRE 2008. Thus, we tested our method only on the dialogue conversations.

4.1. Front-end Processing

The front-end computes 14 Mel-frequency cepstral coefficients plus log-energy to yield a 15 dimensional vector per frame. For compara-

bility with prior work [3], we use only base cepstral cofficients without any delta or double-delta features. We use gender-independent UBM of 1024 Gaussians with diagonal covariance matrices trained on telephone part of the NISTs 1998-2006 SRE corpora.

The total variability matrix of rank 100 was then trained using this UBM from the same files. Using less (50) or more (200) dimensions, we observed degradation of performance similar to [6]. To extract i-vectors segments of the signal defined by speech activity boundaries was divided into portions, having a length not exceeding 100 feature vectors (one second).

When building the similarity matrix we use s-normalized [24] PLDA scores obtained using a subset of 600 additional utterances (also of one-second length). This normalization both makes the score distributions closer to Gaussian and compensates for small shifts of their values.

We implemented an eigenvoice (further termed as EV) diarization system according to the first implementation described in [3]. Hence, we evaluate the proposed EV algorithm just after the first pass without further Viterbi re-segmentation. This enables us to compare the core functionality of the proposed clustering model without mixing in the effects of re-segmentation procedure. The eigenvoice matrix of rank 100 was trained on telephone recordings from the NIST 1998-2006 SRE corpora.

4.2. Evaluation protocol

We evaluate the performance of our diarization system on the summed-channel telephone data from the NIST 2008 SRE corpus consist-ing of 2215 conversations. We use diarization error rate (DER), developed by NIST, as the performance measure [25]. To examine only speaker confusion error, which is the one that demonstrate the performance of speaker clustering, we used reference speech activity boundaries, so that errors caused by speech detectors mismatch were reduced to zero. We estimate DER using `md-eval-v21.pl` Perl script from the NIST website [25]. Following traditional conventions, a forgiveness collar of 0.25 sec was set around the speaker change points.

4.3. Implementation Details

We set the number of VB iterations to 20. Since the deterministic mean-field inference tends to converge to a local optimum and overestimate the number of speakers, we use to the variational lower bound as a score for selecting the resulting segmentation. Specifically, we repeat our algorithm for different upper bound values for the number of speakers, $S \in [S_{\min} = 1, S_{\max} = 5]$, and choose the segmentation with the largest lower bound. The diarization process for the baseline EV system is performed the same way.

As noted above, the hyperparameters of the Normal-Gamma prior have a considerable influence on the performance. We examine two implementation variants. In the first variant (**Proposed I**), the values of β_0 should not be too large to allow the means of target and impostor distributions to be chosen from a broad prior, so that their values could slightly differ from the apriori fixed m_0^{tar} and m_0^{imp} values. In the second variant (**Proposed II**), the hyperparameters m_0^{tar}, m_0^{imp} are pretrained on a development set consisting of 500 utterances from the 2006 NIST SRE corpus. Furthermore, β_0 should be large enough to prevent the deviations of mean values m_0^{tar}, m_0^{imp}.

It should be noted, that we do not perform any external iterations as [7, 8] which generally benefit in diarization performance helping to avoid premature convergence to a local optimum.

5. RESULTS

Table 2 shows the diarization results for both the baseline EV and the proposed systems. The first row of demonstarates that EV system achieves the best DER=1.29%, *if the number of speakers is known a priori* ($S_{\max}$ was set to actual numer of speakers). But if this is not the case (the second row), DER grows to 23%, an order of magnitude degradation. In contrast, our model yields a good practical result in both cases, demonstrating viability of Bayesian model selection. Table 1 below shows the results of determining the number of speakers. Note that all the recordings are dialogue conversations between two parties, thus 2 is the correct answer. Our approach found the correct number in almost 92 % of the cases.

Table 1. Results for determining the number of speakers. The percentage indicates in how many cases (files) the proposed method deemed the corresponding number of speakers.

2	3	4
91.8 %	7.3 %	0.9 %

The last two rows of the Table 2 reveal that hyperparameter learning results in improvement of performance for the proposed system. In case of pre-trained hyperparameters (Proposed II) our model yields DER=1.74%, which is quite close to the best result of EV system.

Table 2. Mean and standard deviation σ of diarization error rates (DER) for the baseline eigenvoice (EV) and the proposed systems on the NIST SRE 2008 telephone data. S_{max} is the upper bound on number of speakers. (For the proposed system in both cases it was set to 5).

System	mean DER, %	σ DER, %
EV (S_{max} = 2)	1.29	4.8
EV (S_{max} = 5)	23.41	11.2
Proposed I	2.35	6.6
Proposed II	1.74	5.3

6. CONCLUSION

We have developed a new generative model for clustering speech data represented by matrices of pairwise similarities. The proposed approach uses rich prior information encoded in speaker verification scores which is benefical in diarization as well. We tackled the problem of unknown number of speakers via Bayesian approach to model selection and parameter estimation. Our experiments demonstrate that, in most cases, the correct number of speakers is detected.

The present study has presented preliminary experimental evaluation on the NIST 2008 SRE corpus consisting of 2-speaker recordings. The evaluation should be extended to datasets containing more than than two speakers, for instance RT07 meeting data [26].

There are several directions for future work. As suggested in [8], one might incorporate temporal information to account for the dependencies across neighboring segments. Another one would be reducing the number of hyperparameters in the model, especially the expected means that now require careful initialization. This could be done by achieving the invariance with respect to any constant shift of scores caused by adding some real number to the similarity matrix.

7. REFERENCES

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

[2] X Anguera, Simon Bozonnet, Nicholas W D Evans, Corinne Fredouille, O Friedland, and O Vinyals, "Speaker diarization : A review of recent research," *"IEEE Transactions On Audio, Speech, and Language Processing" (TASLP), special issue on "New Frontiers in Rich Transcription", February 2012, Volume 20, N2, ISSN: 1558-7916*, 05 2011.

[3] Patrick Kenny, Douglas A. Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *J. Sel. Topics Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[4] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *ICASSP*, 2008, pp. 4133–4136.

[5] Carlos Vaquero, Alfonso Ortega, Jesús A. Villalba, Antonio Miguel, and Eduardo Lleida, "Confidence measures for speaker segmentation and their relation to speaker verification," in *INTERSPEECH*, 2010, pp. 2310–2313.

[6] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas A. Reynolds, and James R. Glass, "Exploiting intra-conversation variability for speaker diarization," in *INTERSPEECH*, 2011, pp. 945–948.

[7] Stephen Shum, Najim Dehak, and Jim Glass, "On the use of spectral and iterative methods for speaker diarization," in *INTERSPEECH*, 2012.

[8] Stephen Shum, Najim Dehak, Réda Dehak, and James R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[9] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, 2001, pp. 849–856.

[10] Rómer Rosales, Kannan Achan, and Brendan J. Frey, "Learning to cluster using local neighborhood structure," in *ICML*, 2004.

[11] S. J. D. Prince, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *Proc. International Conference on Computer Vision*, 2007.

[12] Morten Mørup and Mikkel N. Schmidt, "Bayesian community detection," *Neural Computation*, vol. 24, no. 9, pp. 2434–2456, 2012.

[13] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] N. Brummer, "EM for probabilistic LDA," Available: `https://sites.google.com/site/nikobrummer`, 2010.

[15] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011, pp. 249–252.

[16] Oleg Kudashev and Alexander Kozlov, "The diarization system for an unknown number of speakers," in *SPECOM*, 2013, pp. 340–344.

[17] Yee-Whye Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. Springer, 2010.

[18] Carl Edward Rasmussen, "The infinite gaussian mixture model," in *In Advances in Neural Information Processing Systems 12*. 2000, pp. 554–560, MIT Press.

[19] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[20] Hemant Ishwaran and Lancelot F. James, "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

[21] David M. Blei and Michael I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[22] Radford M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, pp. 249–265, 2000.

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

[24] Mohammed Senoussaoui, Patrick Kenny, Pierre Dumouchel, and Fabio Castaldo, "Well-calibrated heavy tailed bayesian speaker verification for microphone speech," in *ICASSP*, 2011, pp. 4824–4827.

[25] "Diarization error rate scoring code. NIST," Available: `http://www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl`, 2006.

[26] "Rich transcription evaluation project," Available: `http://www.itl.nist.gov/iad/mig/tests/rt/`, 2007.