# PLDA-BASED DIARIZATION OF TELEPHONE CONVERSATIONS

*Ahmet Emin Bulut*[1,2]*, Hakan Demir*[1]*, Yusuf Ziya Işık*[1,2]*, Hakan Erdogan*[2]

[1]TUBITAK BILGEM , Gebze, Turkey
[2]Faculty of Engineering and Natural Sciences, Sabanci University, Turkey
{ahmet.bulut,hakan.demir,yusuf.ziya}@tubitak.gov.tr, haerdogan@sabanciuniv.edu

## ABSTRACT

This paper investigates the application of the probabilistic linear discriminant analysis (PLDA) to speaker diarization of telephone conversations. We introduce using a variational Bayes (VB) approach for inference under a PLDA model for modelling segmental i-vectors in speaker diarization. Deterministic annealing (DA) algorithm is imposed in order to avoid local optimal solutions in VB iterations. We compare our proposed system with a well-known system that applies k-means clustering on principal component analysis (PCA) coefficients of segmental i-vectors. We used summed channel telephone data from the National Institute of Standards and Technology (NIST) 2008 Speaker Recognition Evaluation (SRE) as the test set in order to evaluate the performance of the proposed system. We achieve about 20% relative improvement in Diarization Error Rate (DER) compared to the baseline system.

***Index Terms***— speaker diarization, i-vector, PLDA, deterministic annealing, variational Bayes

## 1. INTRODUCTION

Nowadays, with the explosive growth of audio documents, there is an increasing interest towards applying speech technologies to automatic searching, indexing, and retrieval of audio information. Speaker diarization, which gives the "who spoke when" information without any prior knowledge about speakers, is an important sub-task to address mentioned problems. To illustrate, for an automatic speech recognition system such information allows us to determine the occurrences of specific speaker for a given utterance, which in turn improves transcription performance by speaker adaptation. Moreover, successful diarization of conversations would also increase the performance of speaker verification systems. Speaker diarization of audio data has been studied for different domains, such as meeting, broadcast and telephone recordings [1, 2, 3].

Basically speaker diarization consists of three stages. In the first step, speech activity detection is employed in order to extract speech containing parts from a given utterance. As the second step, the extracted speech parts are further divided into segments according to the speaker changes in such a way that each segment contains the speech of a single speaker. This stage is called speaker segmentation in the literature. Finally, in the clustering stage, all the segments are passed over and the ones spoken by the same speaker are labeled identically. Speaker-based clustering can also be followed by cluster recombination, which refines the speaker clusters for more purity. Among all the components of a speaker diarization system, performance of clustering stage is crucial for the success of the overall system. Many systems have been designed and tuned based on Bayesian Information Criterion (BIC). One such system [4], developed by MIT Lincoln Laboratory, serves as a baseline for a number of studies.

Upon the recent successes of factor analysis based methods, this study explores a new set of such approaches to speaker diarization. We adapt the methods from speaker recognition in order to make use of the concept of inter-speaker variability for the diarization of telephone conversations. Factor analysis based speaker diarization was first introduced in [5] using a stream-based approach. In the study of Kenny et al. [3], they modify Valente's [6] speaker diarization system based on the VB method and they incorporate the factor analysis priors defined by eigenvoices and eigenchannels [7]. Also, in a recent study [8], PLDA is introduced to the problem of speaker diarization. They use factor analysis to extract low-dimensional representation of a sequence of acoustic feature vectors, namely i-vectors [9] which are modelled by PLDA. As the metric for clustering, they use log-likelihood ratio of the probability of hypothesis that two clusters represented by corresponding i-vectors share the same identity and have distinct identities, rather than BIC-based clustering as used in [4]. The authors in [10] proposed k-means clustering for i-vector based diarization approach which constitutes our baseline system. We also extract i-vectors for each segment in a similar way, however we represent i-vectors with a PLDA model and use a VB approach for inference under the model [11].

The rest of the paper is organized as follows. Section 2 provides the overview of our speaker diarization system. The experimental setup and results are then described in section 3. Section 4 is devoted to conclusion and future work, and relation to prior works is explained in section 5.

## 2. PLDA-BASED SPEAKER DIARIZATION SYSTEM

PLDA is originally used for the face recognition task [12]. Later, it is successfully applied to the speaker detection task as well [13, 14]. In our study, PLDA is adapted to the speaker diarization problem by proposing a special generative story for segment i-vectors. This is the first study, to the best of our knowledge, PLDA is used for modelling the extracted segment i-vectors and inference under the model is realized by VB for speaker diarization.

Our speaker diarization system is composed of mainly three parts. Speaker change point detection, alignment of segments over speakers, and re-segmentation. The implementation details of the first and the last parts are similar with the earlier study in [4]. For the second part, where we assign segments to speakers, we follow a VB approach with different initialization methods and a DA variant of VB [15].

### 2.1. Two Covariance PLDA Model

The i-vector features, contain information relevant to factors like channel, microphone, speaking style, language in addition to speaker identity. In speaker verification, PLDA model is used to extract speaker identity related factors from i-vectors. A variant of PLDA, known as two covariance PLDA [16], assumes that the i-vectors are generated by addition of two terms; a speaker vector $y$ unique to a speaker and a residual vector $\epsilon$ unique to the utterance. The speaker vector $y$ is assumed to be sampled from a Gaussian distribution with mean $\mu$ and covariance $\Lambda^{-1}$, and the residual vector is assumed to be sampled from a Gaussian with zero mean and covariance $\mathcal{L}^{-1}$.

### 2.2. Modelling Assumptions

We assume that we have a two covariance PLDA model trained on a separate training set at hand. We assume that we are given a conversation involving $S$ speakers and the speaker change points are specified. Let us denote the set of segment i-vectors by $\mathbf{\Phi} = \{\phi_1, ..., \phi_M\}$. For each segment $m = 1, ..., M$, we define an $S \times 1$ indicator vector $i_m$ whose components are defined as $i_{ms} = 1$ if speaker $s$ is talking in the segment $m$ and $i_{ms} = 0$ otherwise. Let $\mathbf{I} = \{i_1, ..., i_M\}$ be the set of all indicator vectors belonging to the given utterance. We also assign a prior probability to the event that a speaker $s$ is talking in a given segment; we denote and set by $\pi_s = \frac{1}{S}$. The generative story for our PLDA based diarization model is as follows:

- For each speaker $s$ sample $y_s$, from $\mathcal{N}(y; \mu, \Lambda^{-1})$.
- For each segment:
  - Sample $i_m$ from the multinomial distribution $Mult(\Pi)$ where $\Pi = (\pi_1, ..., \pi_S)$. Let $k$ be the index for which $i_{mk} = 1$, with all the other entries of $i_m$ being 0.
  - Sample $\epsilon_m$ from $\mathcal{N}(\epsilon; \bar{0}, \mathcal{L}^{-1})$.
  - The observed segment i-vector is obtained as $\phi_m = y_k + \epsilon_m$.

Let $\mathbf{Y} = \{y_1, ..., y_S\}$ be the set of speaker vectors of the speakers talking in the given utterance. Using this model, we can summarize the diarization problem as of calculating the posterior probability of the speaker talking in a given segment. With these assumptions, obtaining the posterior probability, $P(\mathbf{Y}, \mathbf{I}|\mathbf{\Phi})$ produce intractable integrals. Therefore we resort to the approximate inference methods, namely mean-field VB, in order to approximate $P(\mathbf{Y}|\mathbf{\Phi})$ and $P(\mathbf{I}|\mathbf{\Phi})$.

### 2.3. Variational Bayes for PLDA based i-vectors

The basic assumption for mean-field variational methods is that the approximate posterior factorizes as:

$$Q(\mathbf{Y}, \mathbf{I}) = Q(\mathbf{Y})Q(\mathbf{I}) \tag{1}$$

Approximate segment and speaker posteriors, $Q(\mathbf{I})$ and $Q(\mathbf{Y})$, are defined as:

$$Q(\mathbf{I}) = \prod_{m=1}^{M} \prod_{s=1}^{S} q_{ms}^{i_{ms}} \tag{2}$$

$$Q(\mathbf{Y}) = \prod_{s=1}^{S} \mathcal{N}(y_s \mid \mu_s, C_s^{-1}) \tag{3}$$

In equation (2), we define $q_{ms}$ as the posterior probability of speaker $s$ talking in segment $m$ and in equation (3), it turns out that approximate speaker posterior distributions are Gaussian with mean $\mu_s$ and precision $C_s$.

Adapting the formulation in [17], we formulate segment and speaker posteriors for the VB approach as follows:

1. Update rule for segment posteriors:

$$q_{ms} = \frac{\tilde{q}_{ms}}{\sum_{s'=1}^{S} \tilde{q}_{ms'}} \tag{4}$$

where

$$\log \tilde{q}_{ms} = \mu_s^T \mathcal{L} \phi_m - \frac{1}{2} \mathrm{tr}(\mathcal{L}(C_s^{-1} + \mu_s \mu_s^T)) + \text{const} \tag{5}$$

where const stands for speaker independent terms.

2. Update rule for speaker posteriors:

$$C_s = \Lambda + \sum_{m=1}^{M} q_{ms} \mathcal{L} \tag{6}$$

$$\mu_s = C_s^{-1}(\Lambda\mu + \sum_{m=1}^{M} q_{ms}\mathcal{L}\phi_m) \quad (7)$$

The speaker and segment posteriors are updated alternately throughout the variational e-step. On convergence, diarization is performed by assigning each segment $m$ to the speaker given by $\underset{s}{\operatorname{argmax}}(q_{ms})$ [3].

Initializing the VB algorithm by just assigning random values to the segment posteriors $q_{ms}$ is proved to be ineffective especially for the recordings that one speaker dominates the conversation [3]. For that recordings, two speaker posteriors found by the VB algorithm only model the dominant speaker, and the diarization error rate may be very high corresponding to the average. In order to overcome this problem we try various initialization heuristics for a better start up for the VB iterations and also use a DA variant of the variational algorithm to avoid local optimal results for speaker posteriors.

### 2.4. Initialization of VB Iterations

Firstly, we adopt a heuristic approach in order to initialize segment posteriors similar to the study in [3]. In this setup, instead of starting with two speakers, we randomly initialize the segment posteriors with three speakers. After running the VB algorithm, we compute the pairwise distances among the speakers using their corresponding mean vectors and take the most distant two speakers. Moreover, we iterate this procedure ten times and choose the final speaker pair among the most distant speakers of each iteration. Speaker pair which yields the furthest distant is chosen to be our starting point. We continue to the VB e-step iterations with these two speaker posteriors. As a distance metric we use cosine similarity and likelihood ratio scoring with the PLDA model [16, 12].

### 2.5. Deterministic Annealing variant of Variational Bayes

DA is introduced to the VB method in order to avoid trapping in poor local optimal solutions. This process simply consists of introducing a temperature parameter, $\beta$, to the free energy for controlling the annealing process deterministically [15]. The DA variant of update formulation in section 2.3 can be adapted as follows:

$$\begin{aligned}\log \tilde{q}_{ms} =& \beta(\mu_s^T \mathcal{L}\phi_m - \frac{1}{2}\mathrm{tr}(\mathcal{L}(C_s^{-1} + \mu_s\mu_s^T))) \\ &+ \text{const}\end{aligned} \quad (8)$$

$$C_s = \beta(\Lambda + \sum_{m=1}^{M} q_{ms}\mathcal{L}) \quad (9)$$

$$\mu_s = C_s^{-1}\beta(\Lambda\mu + \sum_{m=1}^{M} q_{ms}\mathcal{L}\phi_m) \quad (10)$$

By introducing the temperature parameter $\beta$ to the formulation, we attain a control on the convergence of the VB algorithm by lowering the precision $C_s$ of the speaker posterior distribution as seen in equation (9).

## 3. EXPERIMENTAL SETUP AND RESULTS

We use 20 dimensional static MFCC features. We use telephone part of the NIST 2004/2005/2006 SRE corpora in order to train gender-independent universal background model (UBM) of 1024 Gaussians. We train gender-independent i-vector model of rank 600 on the same dataset. We extract 600-dimensional i-vectors by using the sufficient statistics collected from the UBM in each segment.

### 3.1. Segmentation

After extraction of MFCC features, we use BIC based penalized likelihood ratio test to detect speaker change points. We check whether the data in the two sides of a candidate change point is better modeled with a single distribution or two. We use full covariance Gaussian distribution for modelling. This is the most widely used approach to speaker diarization for segmentation. Readers may refer to [4] for detailed formulation and configurations.

### 3.2. K-means clustering i-vector System

This system is based on the work described in [10]. After extracting an i-vector for each speech segment in a given utterance, we apply principle component analysis (PCA) based projection. We choose the dimension of PCA-projected vectors for each utterance separately, so that 50% of the energy is preserved. Then, we apply k-means ($K = 2$) clustering to the projected i-vectors based on the cosine distance.

### 3.3. i-vector PLDA System

In our proposed system, we apply linear discriminant analysis (LDA) to the segment i-vectors. After LDA, we apply whitening and unit length normalization before training the PLDA model. We use the same dataset with UBM training for training LDA and PLDA models. In speaker verification, a major source of intra-speaker variability is microphone and channel variations between utterances. For speaker diarization, we have a single session, and phonetic content variabilities are one of the major sources of variation between segment i-vectors of a given speaker. Hence, to obtain a better PLDA model for our task, we take a single utterance from every speaker in the training set. We use the i-vectors extracted from this full utterance, as well as from random cuts between 2 and 20 seconds extracted from it, in LDA and PLDA training. We observe a minor improvement compared to training on multi-session full utterances.

### 3.4. Viterbi re-segmentation

After we complete the initial clustering step by using the VB algorithm, we conduct a frame-based Viterbi re-segmentation to improve the diarization result. We use the labels obtained from the initial clustering step to train 32 mixture GMMs for each speaker. We run the Viterbi algorithm, by fixed self-transition probability, over all speech frames with the two GMMs to obtain final alignments.

### 3.5. Evaluation Protocol

The performance measurement of speaker diarization system is evaluated using diarization error rate (DER). This performance metric is calculated as alignment of reference diarization output with a system diarization output by summing up time weighted combination of: *Miss* - classifying speech as non-speech, *False Alarm* - classifying non-speech as speech and *Speaker Error* - confusing one speakers speech as from another [18]. The evaluation code ignores errors of less than 250ms in the locations of segment boundaries. We take the reference speech activity boundaries as given by using time marks from the speech recognition transcripts produced on each channel separately. Clearly, miss and false alarm errors are mainly caused by a mismatch of the reference speech activity detector and the diarization system output. For a more efficient metric in order to evaluate the effectiveness of our speaker diarization system based on the use of reference speech/non-speech boundaries, we set both miss and false alarm error rates to zero [3, 10].

### 3.6. Results

We use NIST SRE 2008 summed channel telephone data as test set. The dataset consists of 2215 conversations. Each conversation is approximately five minutes in duration ($\approx 200$ hours in total) and involving just two speakers. In the experiments, we use 600-dimensional i-vectors to which we apply a dimensionality reduction procedure. For our system 150-dimensional LDA projection is employed and for the baseline system, we use utterance specific PCA projection keeping $50\%$ of the eigenvalue mass.

Table 1 shows the results of the baseline system (KM-PCA) as well as the results of our proposed system (VB-PLDA) which is initialized with two speakers and randomly generated segment posteriors.

**Table 1**. Comparative results of baseline and proposed system. We randomly initialize $q_{ms}$ with two speakers for VB iterations.

| | KM-PCA | VB-PLDA |
|---|---|---|
| mean DER (%) | 2.72 | 4.14 |
| $\sigma$ (%) | 5.83 | 9.16 |

Table 2 shows the results obtained from our proposed system with two different heuristic initializations and a DA variant of VB. We use two metrics for initialization with cosine similarity (VB-COS) and PLDA log-likelihood ratio (VB-LLR) described in section 2.4. We apply four VB iterations in order to determine best two speaker models out of three for each ten attempts. For obtaining the results of DA variant of VB (DA-VB) system detailed in section 2.5, we set initial value of temperature parameter as, $\beta_{init} = 0.2$ and update as, $\beta_{new} = \beta_{current} \times 1.05$ and continue to the VB iterations as long as $\beta_{new} < 1$. By using DA, we obtain comparable performance to the cumbersome heuristic initialization methods.

**Table 2**. Comparative results of proposed systems with two different VB initializations and the DA variant of VB.

| | VB-COS | VB-LLR | DA-VB |
|---|---|---|---|
| mean DER (%) | 2.18 | 2.19 | 2.28 |
| $\sigma$ (%) | 5.55 | 5.42 | 5.73 |

## 4. CONCLUSION

Motivated by a previous study which utilizes factor analysis with a VB method [3], we develop a system that uses PLDA modelling with a VB method for inference in the speaker diarization problem. We successfully apply DA method to avoid the suboptimal heuristic initialization in VB. We obtain competitive performance as far as the study in [10] is concerned in our experiments.

Our future efforts will continue to apply proposed system to meeting and broadcast data involving an unknown number of speakers.

## 5. RELATION TO PRIOR WORK

In our proposed study we are inspired from a previous study [3], which exploits eigenvoice priors for VB. By our proposed system, we try to obtain a better modeling for the underlying distribution of the speaker factors of the i-vector in a probabilistic framework with the PLDA model which proved to be very successful in speaker recognition. In an another study [8], PLDA is introduced in speaker diarization to compute the log-likelihood ratio as a substitute to BIC scores in the clustering stage. However, we use the PLDA model to represent segmental i-vectors and apply a VB approach for inference in this framework. Moreover, we introduce a formulation based on the DA variant of VB by which we overcome the initialization problem handled by a heuristic method in [3].

## 6. ACKNOWLEDGEMENTS

We would like to thank Patrick Kenny for giving the advice about using DA approach in VB.

## 7. REFERENCES


[1] Hanwu Sun, Bin Ma, Swe Zin Kalayar Khine, and Haizhou Li, "Speaker diarization system for RT07 and RT09 meeting room audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2010, pp. 4982–4985.

[2] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[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] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to broadcast audio and telephone conversations," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[5] 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, IEEE.

[6] Fabio Valente, *Variational Bayesian methods for audio indexing*, Ph.D. thesis, University of Nice Sophia-Antipolis, 2005.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[8] Jan Prazak and Jan Silovský, "Speaker diarization using PLDA-based speaker clustering," in *IEEE 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS, Volume 1*, 2011, pp. 347–350.

[9] 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.

[10] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas A. Reynolds, and James R. Glass, "Exploiting intra-conversation variability for speaker diarization," in *12th Annual Conference of the International Speech Communication Association*, 2011, pp. 945–948.

[11] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 1 edition, 2007.

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

[13] N. Brummer, L. Burget, P. Kenny, P. Matejka, Edward Villiers de, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, Doris Baum, and Mohammed Senoussauoi, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*, 2010, pp. 1–20.

[14] Douglas E. Sturim, William M. Campbell, Najim Dehak, Zahi Karam, Alan McCree, Douglas A. Reynolds, Fred Richardson, Pedro A. Torres-Carrasquillo, and Stephen Shum, "The MITLL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition.," in *ICASSP*. 2011, pp. 5272–5275, IEEE.

[15] K Katahira, K Watanabe, and M Okada, "Deterministic annealing variant of variational Bayes method," *Journal of Physics: Conference Series*, vol. 95-012015, no. 1, 2008.

[16] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010, p. 34.

[17] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," Tech. Rep., 2010.

[18] *Diarization Error Rate (DER) scoring code*, Avaliable: www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl.