SABANCI UNIVERSITY

Diarization of Telephone Conversations using Probabilistic Linear Discriminant Analysis

by

Ahmet Emin Bulut

Submitted to
the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

SABANCI UNIVERSITY

January 2015

DIARIZATION OF TELEPHONE CONVERSATIONS USING PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

APPROVED BY

Assoc. Prof. Dr. Hakan ERDOĞAN (Thesis Supervisor) .....

Assoc. Prof. Dr. Berrin YANIKOĞLU .....

Assoc. Prof. Dr. Özgür ERÇETİN .....

DATE OF APPROVAL: .....

To my family...

Acknowledgements

I would like to express my gratitude to my thesis supervisor Hakan Erdoğan for his invaluable guidance, tolerance, positiveness, support and encouragement throughout my thesis.

I also would like to thank my thesis jury members Berrin Yanıkoğlu and Özgür Erçetin for their valuable ideas.

My parents, Aysel and Bahattin, receive my deepest gratitude and love for their dedication and many years of support during my undergraduate and graduate studies that provided the foundation for this work.

Moreover, I would like to thank my wife, Tuba for her unflagging love and support throughout my thesis. I cannot put into words how much her support means to me and I am beyond fortunate to have her.

I am thankful to all members of TÜBİTAK Speech and Natural Language Processing Laboratory especially Yusuf Ziya Işık and Hakan Demir for sharing their valuable ideas and experiences.

DIARIZATION OF TELEPHONE CONVERSATIONS USING PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

AHMET EMİN BULUT

CS, M.Sc. Thesis, 2015

Thesis Supervisor: Hakan Erdoğan

Keywords: speaker diarization, i-vector, PLDA, deterministic annealing, variational Bayes

Abstract

Speaker diarization can be summarized as the process of partitioning an audio data into homogeneous segments according to speaker identity. This thesis investigates the application of the probabilistic linear discriminant analysis (PLDA) to speaker diarization of telephone conversations. We introduce a variational Bayes (VB) approach for inference under a PLDA model for modeling segmental i-vectors in speaker diarization. Deterministic annealing (DA) algorithm is employed in order to avoid locally optimal solutions in VB iterations. We compare our proposed system with a well-known system that applies k-means clustering on principal component analysis coefficients of segmental i-vectors. We used summed channel telephone data from the National Institute of Standards and Technology 2008 Speaker Recognition Evaluation as the test set in order to evaluate the performance of the proposed system. We achieve about 20% relative improvement in diarization error rate as compared to the baseline system.

TELEFON KONUŞMALARININ OLASILIKSAL DOĞRUSAL AYIRTAÇ ANALİZİ KULLANILARAK BÖLÜTLENMESİ

AHMET EMİN BULUT

CS, Yüksek Lisans Tezi, 2015

Tez Danışmanı: Hakan Erdoğan

Anahtar Kelimeler: konuşmacı bölütleme, i-vektör, ODAA, belirleyici tavlama, değişkenli Bayes

Özet

Konuşmacı bölütleme, ses verisinin konuşmacı kimliğine göre homojen bölütlere ayrılması süreci olarak özetlenebilir. Bu tezde olasılıksal doğrusal ayırtaç analizi (ODAA) metodunun telefon konuşmaları üzerinde konuşmacı bölütleme alanına uygulanması incelenmiştir. Konuşmacı bölütlemede kullanılan bölütsel i-vektörlerin ODAA modeli altında değişkenli Bayes (DB) yöntemi ile çıkarsaması ilk olarak bu çalışmada denenmiştir. Değişkenli Bayes iterasyonlarında yerel en uygun sonuçlardan kaçınmak için belirleyici tavlama (BT) algoritması kullanılmıştır. Önerilen sistem, bu alanda bilinen bir sistem olan, bölütsel i-vektörlerin temel bileşenler analizi katsayıları üzerinde k-ortalama topaklama yönteminin uygulandığı sistem ile karşılaştırılmıştır. Performans değerlendirmesi Ulusal Standartlar ve Teknoloji Enstitüsü tarafından 2008 Konuşmacı Tanıma Değerlendirmesi için belirlenen test veri kümesi üzerinde yapılmıştır. Önerilen sistem, baz alınan sistemin Bölütleme Hata Oranı'na göre %20 daha iyi performans göstermiştir.

Contents

List of Figures

List of Tables

Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
DER	**D**iarization **E**rror **R**ate
DA	**D**eterministic **A**nnealing
EM	**E**xpectation **M**aximization
GMM	**G**aussian **M**ixture **M**odel
JFA	**J**oint **F**actor **A**nalysis
KL	**K**ullback **L**iebler (Divergence)
LDA	**L**inear **D**iscriminant **A**nalysis
MFCC	**M**el **F**requency **C**epstral **C**oefficients
NIST	**N**ational **I**nstitute of **S**tandards and **T**echnology
PCA	**P**rincipal **C**omponent **A**nalysis
PLDA	**P**robabilistic **L**inear **D**iscriminant **A**nalysis
RTF	**R**eal **T**ime **F**actor
SRE	**S**peaker **R**ecognition **E**valuation
SV	**S**peaker **V**erification
TVS	**T**otal **V**ariability **S**pace
UBM	**U**niversal **B**ackground **M**odel
VAD	**V**oice **A**ctivity **D**etector
VB	**V**ariational **B**ayes

Chapter 1

Introduction

Along with the recent explosive growth of audio documents, there has been 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 [3–5].

Although speaker diarization goes hand-in-hands with speaker recognition in terms of the methods for distinguishing between speakers they differs in many ways. One of the major differences is that in the speaker recognition setting prior models are trained on the data sets of target speakers, whereas for speaker diarization, there is no prior information regarding any of the speakers in the recording. And the second main difference is the average length of the test and train data when distinguishing between speakers which is generally very short in the case of speaker diarization while for speaker recognition data scarcity occures exceptionally.

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. The general schema is illustrated in Figure 1.1. Speaker-based clustering can also be followed by cluster re-combination, 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 [1], developed by MIT Lincoln Laboratory, serves as a baseline for a number of studies.

1.1 Motivation

In many systems speaker diarization is used as a preprocessing stage for example like in an automatic speech recognition (ASR) system or speaker verification (SV) system. The motivation of this study is to enable us with a successful speaker based diarization in order to obtain a meaningful and efficient result for the system that utilizes that information. Imagine a call center of a company where all conversations are recorded for the sake of customer satisfaction. Transcription of these recordings with an ASR system is crucial for reporting and archiving. By preprocessing with a speaker diarization system one can improve the accuracy of the ASR system output by tuning the system with acoustic characteristics of each customer or representative. Also determining the start and the end time of the speech for each speaker make the transcripts more readable for human readers. Another problem is making an audio search in that bulk collection of telephone data. For a given model of representative for example one can want to make a SV test by sliding a window throughout each recording in order to determine the ones

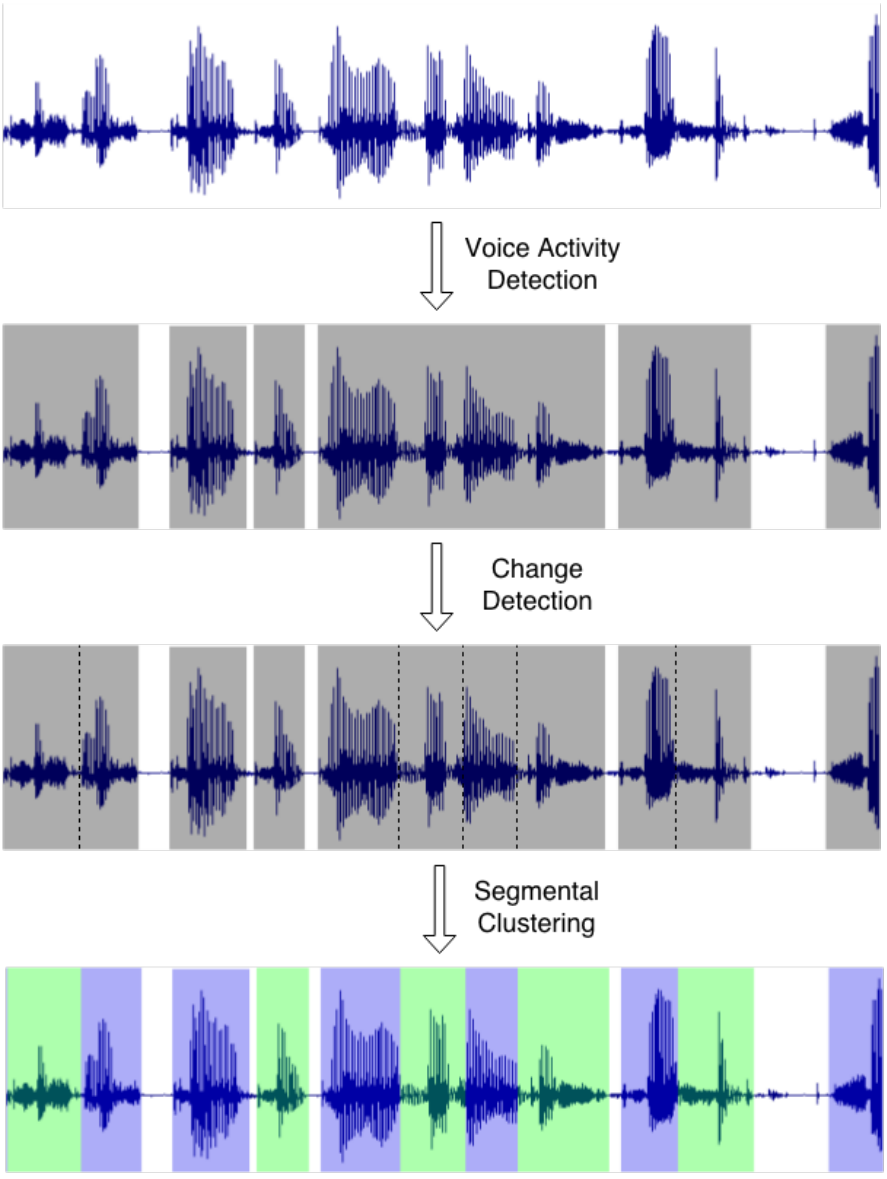

FIGURE 1.1: General schematic overview of a Speaker Diarization process that can be divided into three main subtasks: Voice activity detection, change detection, and segmental clustering.

in which specific representative is speaking. However, clearly that sort of testing approach may decrease the SV performance because of short and impure test data. By employing a speaker diarization preprocessing over the test data can definitely provide sufficient data for a better SV test.

1.2 Contributions

In our proposed study we are inspired from a previous study [5], which exploits eigenvoice priors for variational Bayes (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 probabilistic linear discriminant analysis (PLDA) model which proved to be very successful in speaker recognition. In another study [6], PLDA is introduced in speaker diarization to compute the log-likelihood ratio as a substitute to Bayesian information criterion (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 deterministic annealing (DA) variant of VB by which we overcome the initialization problem handled by a heuristic method in [5].

1.3 Outline

The rest of the thesis is organized as follows. Chapter 1 provides the overview of speaker diarization, contribution and outline of the this. Related work is detailed in Chapter 2. Chapter 3 is devoted to our proposed system. The experimental setup and results are then described in Chapter 4. Chapter 5 is devoted to conclusion and future work.

Chapter 2

Related Work

2.1 Agglomerative Clustering based Speaker Diarization

The problem of "who spoke when" in an audio document has been explored by various researchers using a variety of methods. The agglomerative clustering method which is one of the popular bottom-up clustering methods is initially proposed by Reynolds et al. [1]. Main components of this type of speaker diarization system is provided in Figure 2.1.

2.1.1 Bayesian Information Criterion

The Bayesian information criterion (BIC) is initially used in speaker diarization in [7]. It is a model selection criterion in order to describe a given data set penalized by the number of parameters in the model. To illustrate, let $\mathcal{X} = \{x_i : i = 1, ..., N\}$ be the data set we are modelling and $\Lambda = \{\lambda_i : i = 1, ..., M\}$ be the probable models. We aim at maximizing likelihood function, say $\mathcal{L}(\mathcal{X}, \lambda_i)$ for each model λ_i. The BIC score is defined as follows:

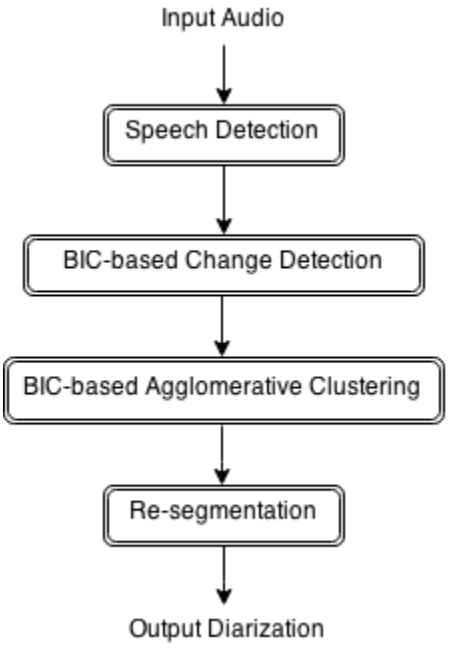

FIGURE 2.1: System diagram of the basic speaker diarization system using BIC-based agglomerative clustering [1].

$$BIC(\lambda_i) = \log \mathcal{L}(\mathcal{X}, \lambda_i) - \alpha \frac{1}{2} \#(\lambda_i) \times \log(N) \tag{2.1}$$

where α is called the BIC weight and $\#(\lambda_i)$ is the number of parameters that are need to be estimated in model λ_i. We can clearly observe from the formula that we search for a model has a better fit to the data but has fewer free parameters.

2.1.2 Stages of Agglomerative Clustering

Speech detection is the first step for a basic speaker diarization system. Generally, the input data consists of many non-speech parts such as silence, music, and background noise. In order to separate speech parts from the other sources generally Gaussian mixture model (GMM) based voice activity detection is used [1]. However, for an easier problem where the segments consists mostly of just speech and silence, energy based voice activity detection can be employed [8].

The next step after the speech detection is change detection which aims at finding the possible speaker change point. Initially in [7] and later in [1] BIC based change detection is used. As described in Figure 2.2, the method searches for a change point within a window using a penalized likelihood ratio (S_{BIC}) between modeling the probability density function of the window as a single full-covariance Gaussian and two full-covariance Gaussians [1]. The BIC score for segment z supposed to consists of two speaker segments namely x and y is defined as follow:

$$S_{BIC} = \log \frac{p(x|\lambda_x)p(y|\lambda_y)}{p(z|\lambda_z)} - \alpha P, \tag{2.2}$$

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N \tag{2.3}$$

where λ_x, λ_y, and λ_z are corresponding full-covariance Gaussian segment models, α is the BIC weight (typically set to 1.0 [1]) and P is the BIC penalty for d-dimensional full-covariance Gaussian model in a window of size N.

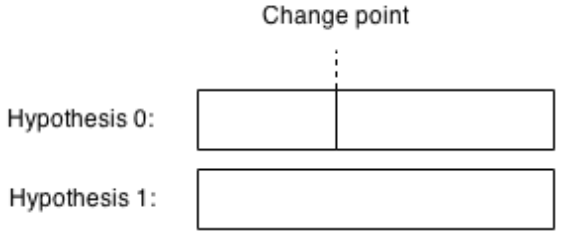

FIGURE 2.2: Illustration of BIC-based speaker change point detection within a given window. Hypothesis 0: window is modeled by two distributions, that is, there exists a change point, Hypothesis 1: window is modeled by a single distribution, that is, no change point found [1].

We can calculate probability density of a segment x, that consists of N_x frames, for a given full-covariance Gaussian segment model, $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ as follows [1]:

$$\log p(x|\lambda_x) = -\frac{1}{2}N_x \log |\mathbf{\Sigma}_x| - \frac{1}{2}N_x(x - \boldsymbol{\mu}_x)^T \mathbf{\Sigma}_x^{-1}(x - \boldsymbol{\mu}_x). \tag{2.4}$$

We decide the existence of possible change point when S_{BIC} is greater then zero [1]. If a change point is found, the starting point of the window is set to the change point and the search is restarted. Otherwise, the window is enlarged to the next break point and the procedure is repeated again.

After change point detection agglomerative clustering part is initialized with the segments in which presumably one speaker speaks. In clustering stage we aim at getting the segments together with respect to the same speaker. The outline of the agglomerative clustering stage can be summarized as follows [1]:

1. Initialize bottom clusters of the segment tree obtained from the change detection stage.
2. Compute pair-wise distances between each cluster.
3. Merge closest clusters and set distances to infinity between these clusters and the remaining ones.
4. Compute distances between new clusters and the remaining ones.
5. Iterate steps 3-4 until the stopping criterion is met.

Generally BIC based metric is used for distance calculations and stopping criterion, as in [1]. However for some studies that number of speakers are known a priori, algorithm is terminated when number of clusters reach to the number of speakers, as in [5].

In order to reduce computation time for comparing the clusters and each calculation of cluster mean and covariance, one can use Hotelling's T^2 statistics [9]. The method suggests a new likelihood ratio test statistic for the clusters to be compared by using the mean and the covariance of the clusters. At each instance of

merging two clusters the method gives a way to calculate new mean and covariance out of previously calculated mean and covariance of the clusters.

After completing the clustering stage, as a final step frame-based Viterbi re-segmentation is conducted by speaker and non-speech models. This part will be detailed in Section 4.4. The main purpose of this stage is improving the diarization result by refining segment boundaries.

2.2 Factor Analysis Based Systems

Upon the recent successes of factor analysis based methods, a new set of such approaches are applied to speaker diarization. The methods are adapted 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 [10] using a stream-based approach. In the study of Kenny et al. [5], they modify Valente's [11] speaker diarization system based on the VB method and they incorporate the factor analysis priors defined by eigenvoices and eigenchannels [12]. Theoretical background will be described in Section 2.2.1 about these priors. Also, in a recent study [6], 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 [13] which are modelled by PLDA. The approaches i-vector and PLDA will be described in detail in Section 2.2.2 and Section 3, respectively. 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 [1]. The authors in [14] proposed k-means clustering for i-vector based diarization approach which constitutes our baseline system detailed in Section 2.2.4. In this thesis 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 [15]. The study on VB approach will be described generally in Section 2.2.3.

2.2.1 Joint Factor Analysis

The Joint Factor Analysis (JFA) method is initially used in [16] for speaker verification in order to compensate inter-session and inter-speaker variabilities. Different from the classical UBM-MAP adaptation method [17], JFA method achieves a better model for speaker and channel dependent factors. The theory of JFA can be outlined as follows:

First we define the Gaussian Mixture Model (GMM) which is also called the universal background model (UBM) in the context of speaker verification problems. This is a widely used generative model for speech data. Given a GMM model θ, consisting C components and a dimension of F feature vectors, the likelihood of observing a given feature vector $\boldsymbol{y}$ is computed as follows:

$$p(\boldsymbol{y}|\theta) = \sum_{c} w_c \mathcal{N}_c(\boldsymbol{y}|\boldsymbol{m}_c, \boldsymbol{\Sigma}_c) \tag{2.5}$$

where w_c is the mixture components which all sum up to one and $\mathcal{N}_c(\boldsymbol{y}|\boldsymbol{m}_c, \boldsymbol{\Sigma}_c)$ is a multivariate Gaussian distribution with F dimensional mean vector, $\boldsymbol{m}_c$ and $F \times F$ dimensional covariance matrix, $\boldsymbol{\Sigma}_c$ defined for each mixture component as follows:

$$\mathcal{N}_c(\boldsymbol{y}|\boldsymbol{m}_c, \boldsymbol{\Sigma}_c) = \frac{1}{(2\pi)^{\frac{F}{2}}|\boldsymbol{\Sigma}_c|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{m}_c)^T \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{y}-\boldsymbol{m}_c)\right\}, \tag{2.6}$$

note that $(.)^T$ stands for transpose for above formulation and throughout the thesis.

Assume we are given a UBM defined as a large GMM trained to represent the speaker-independent distribution of features. Then a speaker supervector of dimension $C \cdot F$ is adapted from the UBM with a mean supervector of dimension

$CF \times 1$ and a block diagonal covariance matrix of dimension $CF \times CF$. This mean supervector, and covariance matrix is obtained by concatenating all the Gaussian component means and covariances.

The main idea behind the JFA is that the GMM supervector, $\boldsymbol{M}$ can be obtained by the sum of speaker and channel independent, speaker-dependent, channel-dependent, and a residual component, as:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{D}\boldsymbol{z} \tag{2.7}$$

where $\boldsymbol{m}$ is speaker and channel independent $CF \times 1$ supervector obtained from the UBM. $\boldsymbol{V}$ and $\boldsymbol{U}$ are lower dimensional speaker and channel subspaces, so called eigenvoices and eigenchannels respectively. Moreover, $\boldsymbol{D}$ is diagonal $CF \times CF$ matrix which models the residual variabilities that are not captured by $\boldsymbol{V}$ and $\boldsymbol{U}$. Lastly, $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are the normally distributed hidden vectors that need to be estimated.

By calculating some sufficient statistics across the speakers we can train $\boldsymbol{V}$, $\boldsymbol{U}$, and $\boldsymbol{D}$ matrices in given order and in turn estimation of speaker and channel factor vectors $\boldsymbol{y}$ and $\boldsymbol{x}$ and residual factor vector $\boldsymbol{z}$ can be achieved by using those trained speaker and channel subspace matrices. Formulation details and proofs can be found in [18].

2.2.2 Total Variability Approach

Total variability approach to speaker recognition is initially used in [19] and extended in [13] which propose a simplified solution to the problem at hand in terms of both theory and implementation. Different from the JFA, this approach represents all the speaker and channel factors in a single total variability space. The GMM supervector, $\boldsymbol{M}$ with this approach is defined as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{\phi} \tag{2.8}$$

where $\boldsymbol{m}$ is again speaker and channel independent $CF \times 1$ supervector obtained by concatenating each mixture mean of the UBM, assume we are given a UBM with C mixture component and for each mixture c we denote w_c, $\boldsymbol{m}_c$, and $\boldsymbol{\Sigma}_c$ as the corresponding mixture weight, mean vector, and covariance matrix. $\boldsymbol{T}$ is lower dimensional $CF \times R$ total variability space, where $R \ll CF$ and $\boldsymbol{\phi}$ is the normally distributed hidden vector, so called i-vector, stands for the intermediate vector for its intermediate representation between an acoustic feature vector and a supervector. The process of training the total variability matrix, $\boldsymbol{T}$ and extracting the i-vector, $\boldsymbol{\phi}$ can be summarized as follows:

In order to continue to the calculations component wise, we can see the model defined in equation (2.8) as follows:

FIGURE 2.3: Illustration of the model with supervector $\boldsymbol{M} = \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_C$, UBM mean supervector $\boldsymbol{m} = \boldsymbol{m}_1, ..., \boldsymbol{m}_C$, and total variability matrix $\boldsymbol{T} = \boldsymbol{T}_1, ..., \boldsymbol{T}_C$ stands for all mixture components of the UBM with i-vector $\boldsymbol{\phi}$.

$$\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_C \end{bmatrix}_{CF\times 1} = \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \\ \vdots \\ \boldsymbol{m}_C \end{bmatrix}_{CF\times 1} + \left[\begin{array}{ccc} \dots & \boldsymbol{T}_1 & \dots \\ \hline \dots & \boldsymbol{T}_2 & \dots \\ \hline & \vdots & \\ \hline \dots & \boldsymbol{T}_C & \dots \end{array}\right]_{CF\times R} \begin{bmatrix} | \\ \boldsymbol{\phi} \\ | \end{bmatrix}_{R\times 1}$$

then we can write each equation for the corresponding UBM mixture component c as follows:

FIGURE 2.4: Illustration of the model for a mixture c of the UBM.

$$\begin{bmatrix} | \\ \boldsymbol{\mu}_c \\ | \end{bmatrix}_{F\times 1} = \begin{bmatrix} | \\ \boldsymbol{m}_c \\ | \end{bmatrix}_{F\times 1} + \begin{bmatrix} & | & \\ \dots & \boldsymbol{T}_c & \dots \\ & | & \end{bmatrix}_{F\times R} \begin{bmatrix} | \\ \boldsymbol{\phi} \\ | \end{bmatrix}_{R\times 1}$$

$$\boldsymbol{\mu}_c = \boldsymbol{m}_c + \boldsymbol{T}_c\boldsymbol{\phi}. \tag{2.9}$$

Let we are given an utterance s represented as a sequence of L acoustic feature vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_L$. The i-vector $\boldsymbol{\phi}$ is a latent variable whose posterior distribution can be calculated using Baum-Welch statistics extracted from the UBM [13]. These sufficient statistics are defined, as follows:

$$N_c = \sum_t \gamma_t(c), \tag{2.10}$$

$$\boldsymbol{F}_c = \sum_t \gamma_t(c)\boldsymbol{x}_t, \tag{2.11}$$

$$\boldsymbol{S}_c = \sum_t \gamma_t(c)\boldsymbol{x}_t\boldsymbol{x}_t^T \tag{2.12}$$

where $c = 1, ..., C$ is the index of corresponding Gaussian component of the UBM and $\gamma_t(c)$ is the posterior probability that $\boldsymbol{x}_t$ is generated by the mixture component c, defined as follows:

$$\gamma_t(c) = \frac{w_c\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}_c, \boldsymbol{\Sigma}_c)}{\sum_c w_c\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}_c, \boldsymbol{\Sigma}_c)}. \tag{2.13}$$

Then for a given utterance $\boldsymbol{x}_1, ..., \boldsymbol{x}_L$, we can calculate posterior mean and covariance of the i-vector, $\boldsymbol{\phi}$ as follows [20]:

$$\langle\boldsymbol{\phi}\rangle = \text{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi}) \sum_c \boldsymbol{T}_c^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{F}_c - N_c\boldsymbol{m}_c), \tag{2.14}$$

$$\mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi}) = (\boldsymbol{I} + \sum_{c} N_c \boldsymbol{T}_c^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{T}_c)^{-1} \tag{2.15}$$

where $\boldsymbol{I}$ stands for identity matrix.

For estimating the total variability matrix, $\boldsymbol{T}_c$ we can use the first and second order moments of the posterior distribution of $\boldsymbol{\phi}$, $\langle \boldsymbol{\phi}(s) \rangle$ and $\langle \boldsymbol{\phi}(s) \boldsymbol{\phi}(s)^T \rangle$ for each utterance s [20] as follows:

$$\boldsymbol{T}_c = \left(\sum_{s} \boldsymbol{F}_c(s) \langle \boldsymbol{\phi}^T(s) \rangle \right) \left(\sum_{s} N_c(s) \langle \boldsymbol{\phi}(s) \boldsymbol{\phi}^T(s) \rangle \right)^{-1} \tag{2.16}$$

where the sum over s is for all utterances in the training set and for each utterance s zeroth and first order statistics $N_c(s)$ and $\boldsymbol{F}_c(s)$ are defined as in equation (2.10) and (2.11). And also second order moment can be easily calculated as follows:

$$\langle \boldsymbol{\phi}(s) \boldsymbol{\phi}^T(s) \rangle = \mathrm{Cov}(\boldsymbol{\phi}(s), \boldsymbol{\phi}(s)) + \langle \boldsymbol{\phi}(s) \rangle \langle \boldsymbol{\phi}^T(s) \rangle. \tag{2.17}$$

Actually, the training procedure for the $\boldsymbol{T}$ matrix is same as the training of $\boldsymbol{V}$ matrix in equation (2.7) detailed in [18], but differently treat all conversation sides of all training speakers as belonging to different speakers. Theoretical background, formulation detail and proofs can be found in [21].

2.2.3 Variational Bayes System

The variational Bayes method of speaker diarization developed by Kenny et al. [5] is one of the first systems where factor analysis is used. In a previous study [11] Valente used a fully Bayesian treatment for the problem of estimating speaker

GMM models, however in this study they use eigenvoice and eigenchannel priors on GMMs by employing a variational Bayesian framework.

Two speaker telephone conversations are used as a test set and an initial segmentation of uniform one second intervals is applied on segments after removing silences. The diarization problem is formulated as one of calculating, for each speaker segment, the posterior probabilities of the events that speaker s is talking in the segment m, denoted as q_{ms}. As well as, two speaker posteriors are calculated which are multivariate Gaussian distributions on speaker factors with mean $\boldsymbol{a}_s$ and precision $\boldsymbol{\Lambda}_s$. The alignment of speech frames with Gaussians in speaker GMMs is carried out with a UBM such that the sufficient Baum-Welch statistics is extracted from each of the speaker segments by means of the UBM. Modelling assumptions on distribution of speaker supervector is as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} \tag{2.18}$$

where $\boldsymbol{m}$ denotes $CF \times 1$ mean supervector and $\boldsymbol{\Sigma}$ denotes $CF \times CF$ supervector sized diagonal covariance matrix for an UBM with C mixtures in an F dimensional space. We assume $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ are obtained by concatenating mean vectors $\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_C$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_C$, respectively, for each mixture of the UBM. And $\boldsymbol{V}$ denotes $CF \times R$ block diagonal eigenvoice matrix and $\boldsymbol{y}$ denotes $R \times 1$ speaker factor vector with standard normal distribution.

The sufficient statistics N_c, $\boldsymbol{F}_c$, and $\boldsymbol{S}_c$ for a given segment feature vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_L$ are defined as equations (2.10), (2.11), and (2.12). For formulation convenience centralized first and second order Baum-Welch statistics $\tilde{\boldsymbol{F}}_c$ and $\tilde{\boldsymbol{S}}_c$ are defined as follows:

$$\tilde{\boldsymbol{F}}_c = \boldsymbol{F}_c - N_c \boldsymbol{m}_c, \tag{2.19}$$

$$\tilde{\boldsymbol{S}}_c = \boldsymbol{S}_c - \text{diag}(\boldsymbol{F}_c \boldsymbol{m}_c^T + \boldsymbol{m}_c \boldsymbol{F}_c^T - N_c \boldsymbol{m}_c \boldsymbol{m}_c^T) \tag{2.20}$$

where $\boldsymbol{m}_c$ is the subvector of $\boldsymbol{m}$ which corresponds to the mixture component c of the UBM. Also, let $\boldsymbol{N}$ be the $CF \times CF$ diagonal matrix with diagonal blocks of $N_c \boldsymbol{I}$ and $\tilde{\boldsymbol{F}}$ be the $CF \times 1$ supervector obtained by concatenating $\tilde{\boldsymbol{F}}_c$ for $c = 1, 2, ..., C$.

The details of training procedure of the variational Bayes system is as follows:

1. Updating the segment posterior q_{ms} for segment m and speaker s

$$q_{ms} = \frac{\tilde{q}_{ms}}{\sum_{s'=1}^{S} \tilde{q}_{ms'}} \tag{2.21}$$

where

$$\begin{aligned} \log \tilde{q}_{ms} = \boldsymbol{a}_s^T \boldsymbol{V}^T \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{F}}_m - \frac{1}{2} \boldsymbol{a}_s^T \boldsymbol{V}^T N_m \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{a}_s \\ - \frac{1}{2} \text{tr} \left(\boldsymbol{V}^T N_m \boldsymbol{\Sigma}^{-1} \boldsymbol{V} \boldsymbol{\Lambda}_s^{-1} \right) + \text{const} \end{aligned} \tag{2.22}$$

where const stands for speaker independent terms and N_m and $\tilde{\boldsymbol{F}}_m$ are centralized Baum-Welch statistics extracted from the segment m.

2. Updating speaker posteriors for speaker s

$$\boldsymbol{\Lambda}_s = \boldsymbol{I} + \boldsymbol{V}^T \boldsymbol{\Sigma}^{-1} \left(\sum_m q_{ms} N_m \right) \boldsymbol{V}, \tag{2.23}$$

$$\boldsymbol{a}_s = \boldsymbol{\Lambda}_s^{-1} \boldsymbol{V}^T \boldsymbol{\Sigma}^{-1} \left(\sum_m q_{ms} \tilde{\boldsymbol{F}}_m \right). \tag{2.24}$$

3. The speaker and segment posteriors are updated alternately until the variational lower bound, $\mathcal{L}$ converges

$$\begin{aligned} \mathcal{L} = & \sum_m \sum_s q_{ms} \log \tilde{q}_{ms} + \frac{1}{2} \left\{ RS - \sum_s \left(\log |\boldsymbol{\Lambda}_s| + \operatorname{tr} \left(\boldsymbol{\Lambda}_s^{-1} + \boldsymbol{a}_s \boldsymbol{a}_s^T \right) \right) \right\} \\ & - \sum_m \sum_s q_{ms} \log q_{ms} \end{aligned} \tag{2.25}$$

where q_{ms} and $\tilde{q}_{ms}$ are given as equation (2.21) and equation (2.22) and Λ_s and a_s are given as equation (2.23) and equation (2.24). On convergence, diarization is performed by assigning each segment m to the speaker given by $\underset{s}{\operatorname{argmax}}\, q_{ms}$.

Theoretical background, formulation detail and proofs can be found in technical report written by P. Kenny [22].

Initializing the Variational Bayes algorithm by assigning random values to the segment posteriors is found to be ineffective for recordings where one speaker dominates the conversation. In order to prevent the variational Bayes algorithm from modelling the dominant speaker, some sort of heuristic initialization method is applied [5].

After variational iterations completed, in order to refine initial segmentation boundaries Viterbi re-segmentatiton and Baum-Welch soft speaker clustering is applied at the accoustic feature level. For each speaker a Gaussian mixture model is trained by the data determined by the previous steps. Then, by using the Baum-Welch statistics the posterior probabilities of each speaker are calculated given each feature frame.

Speaker change points and speaker assignments found by Viterbi re-segmentation are used to initialize a second run of Variational Bayes. This procedure is called second-pass which aims to provide further refinement in speaker boundaries.

2.2.4 K-means Clustering System

This system is one of the factor analysis based systems proposed by S. Shum [14]. In this study, the total variability space is used in order to represent the speaker segments. The method is detailed in Section 2.2.2 which has achieved considerable performance in the task of speaker verification [13]. Starting with an initial segmentation, i-vector extraction is obtained from each speech segment. The main focus of the work is to concentrate on intra-session variability rather than inter-session variability. Because the work is performed on summed-channel telephone data and the diarization process includes to detect speaker changes in a conversation different from the speaker verification task which assumes that a given utterance contains speech from only one speaker.

Two types of initial segmentation is applied in order to detect speech parts, one is a harmonicity and modulation frequency based voice activity detector (VAD) described in [23] and the other is the use of reference boundaries as the initial speech/non-speech segmentation. After voice activity detection, in order to have a better identification of speaker i-vector directions the principle component analysis (PCA) is applied to the segment i-vectors within the total variability space. PCA-based projection is applied in a way that p proportion of eigenvalue mass determines the number of principle components. That is, minimum number of n largest eigenvalues are chosen as follows:

$$\min_{n} \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \geq p \tag{2.26}$$

where D is i-vector dimension and λ's are eigenvalues indexed with decreasing order.

Then, in order to align segments with speakers, PCA-projected i-vectors are subjected to k-means clustering algorithm based on the cosine distance. At the end, a re-segmentation stage is applied like the end of the variational Bayes system detailed in Section 2.2.3 in order to refine speaker boundaries.

Chapter 3

PLDA-based Speaker Diarization System

PLDA is originally used for the face recognition task [24]. Later, it is successfully applied to the speaker detection task as well [25, 26]. 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, where 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 [1] which are detailed in Section 2.1.2. 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 [27].

3.1 LDA-based Dimensionality Reduction

Assuming that we are given an initial segmentation and thus can extract an i-vector for each segment as detailed in Section 2.2.2. In order to improve our discrimination performance between speakers of i-vectors we apply Linear Discriminant Analysis (LDA) technique by projecting i-vectors into an orthogonal basis. In the below formal definition of PLDA, the classes and feature vectors for each class can be considered as speakers and corresponding i-vectors, respectively.

Given a set of features belonging to different classes, this dimensionality reduction method tries to find a mapping in order to represent the features that enables better discrimination between different classes by maximizing between-class variance and minimizing within class variance. We can find that sort of orthogonal basis by defining a mapping (projection matrix) $\boldsymbol{A}$ composed of the eigenvectors associated with the greatest eigenvalues of the generalized eigenvalue equation:

$$\boldsymbol{S}_b\boldsymbol{v} = \lambda\boldsymbol{S}_w\boldsymbol{v} \tag{3.1}$$

where λ's are the eigenvalues and v's are corresponding eigenvectors. The matrices $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are between-class and within-class covariance matrices, respectively. For a given training set containing S speakers with n_s utterances per speaker, these covariance matrices are estimated as follows:

$$\boldsymbol{S}_b = \sum_{s=1}^{S}(\boldsymbol{\phi}_s - \bar{\boldsymbol{\phi}})(\boldsymbol{\phi}_s - \bar{\boldsymbol{\phi}})^T, \tag{3.2}$$

$$\boldsymbol{S}_w = \sum_{s=1}^{S}\frac{1}{n_s}\sum_{n=1}^{n_s}(\boldsymbol{\phi}_s^{(n)} - \boldsymbol{\phi}_s)(\boldsymbol{\phi}_s^{(n)} - \boldsymbol{\phi}_s)^T \tag{3.3}$$

where $\boldsymbol{\phi}_s^{(n)}$ is the n^{th} element of class s and class mean $\boldsymbol{\phi}_s$ and global mean $\bar{\boldsymbol{\phi}}$ can be calculated as follows:

$$\boldsymbol{\phi}_s = \frac{1}{n_s} \sum_{n=1}^{n_s} \boldsymbol{\phi}_s^{(n)}, \tag{3.4}$$

$$\bar{\boldsymbol{\phi}} = \frac{1}{S} \sum_{n=1}^{S} \boldsymbol{\phi}_s. \tag{3.5}$$

The LDA-projected i-vectors $\boldsymbol{\phi}_{LDA}$ can be calculated by using the projection matrix $\boldsymbol{A}$, whose columns are the eigenvectors corresponding to the largest eigenvalues of generalized eigenvalue equation (3.1), as follows:

$$\boldsymbol{\phi}_{LDA} = \boldsymbol{A}^T \boldsymbol{\phi}. \tag{3.6}$$

In speaker recognition literature LDA is mainly used for compensating channel and session differences. Channel and session independent representation of i-vectors can be obtained by reducing the intra-class variability. As declared in the study [2], in the problem of summed channel telephone diarization compensation of inter-session variability may be unnecessary because there is exactly one session in nature of the problem. However, the meaning of the LDA-projection depends mainly on the choice of the LDA model training set. For a speaker recognition problem training set consists of various recordings of different speakers which are recorded by different microphones for each session. However, creating a training set

by letting each class including one recording of a specific speaker and various non-overlapping random cuts extracted from that recording may convert the meaning of reducing intra-class variability into the compensating intra-speaker variations. Classes of LDA and PLDA models training set can be illustrated as in Figure 3.1(a) and 3.1(b) for a general speaker recognition problem and our proposed speaker diarization system.

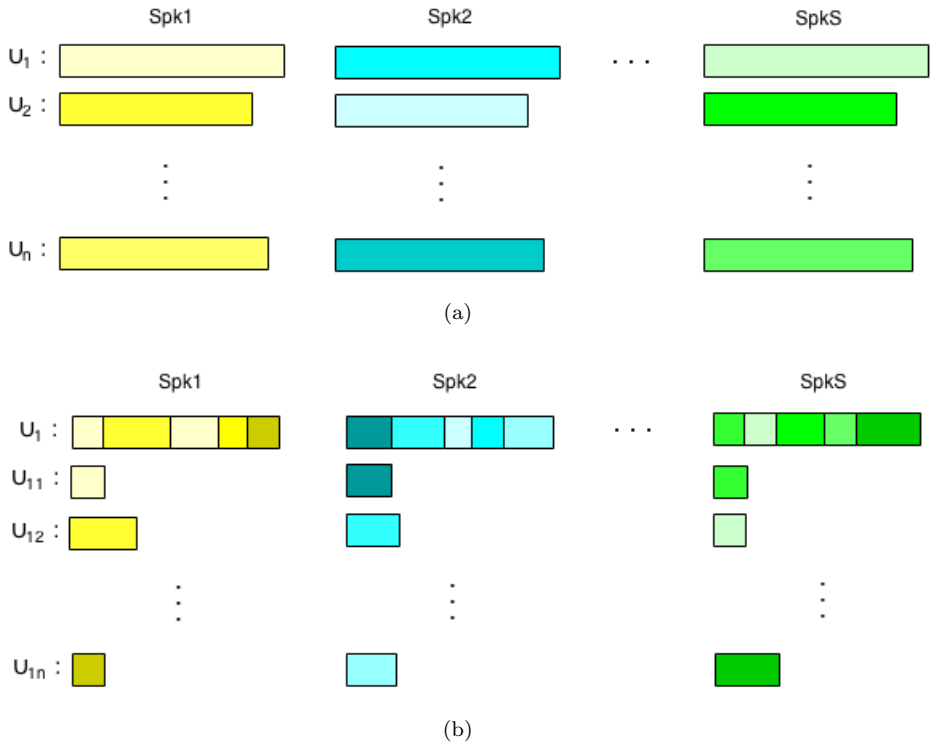

FIGURE 3.1: (a) Classes of LDA and PLDA models training set for a speaker recognition task. (b) Classes of LDA and PLDA model training set for our proposed speaker diarization system.

All i-vectors used in modelling and formulation of the system are considered as LDA-projected throughout the chapter.

3.2 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 [28], assumes that the i-vectors are generated by addition of two terms; a speaker vector $\boldsymbol{y}$ unique to a speaker and a residual vector $\boldsymbol{\epsilon}$ unique to the utterance. The speaker vector $\boldsymbol{y}$ is assumed to be sampled from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Lambda}^{-1}$, and the residual vector is assumed to be sampled from a Gaussian with zero mean and covariance $\boldsymbol{\mathcal{P}}^{-1}$. These model parameters namely mean $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Lambda}^{-1}$ and $\boldsymbol{\mathcal{P}}^{-1}$ are estimated during the PLDA model training stage.

3.3 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 $\Phi = \{\phi_1, ..., \phi_M\}$. For each segment $m = 1, ..., M$, we define an $S \times 1$ indicator vector $\boldsymbol{i}_m$ whose components are defined as $i_{ms} = 1$ if speaker s is talking in the segment m and $i_{ms} = 0$ otherwise. Let $\mathcal{I} = \{\boldsymbol{i}_1, ..., \boldsymbol{i}_M\}$ 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 $\boldsymbol{y}_s$, from $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$.
- For each segment:
  - Sample $\boldsymbol{i}_m$ from the multinomial distribution $Mult(\boldsymbol{i}; \Pi)$ where $\Pi = \{\pi_1, ..., \pi_S\}$. Let s be the index for which $i_{ms} = 1$, with all the other entries of $\boldsymbol{i}_m$ being 0.
  - Sample $\boldsymbol{\epsilon}_m$ from $\mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{\mathcal{P}}^{-1})$.
  - The observed segment i-vector is obtained as $\boldsymbol{\phi}_m = \boldsymbol{y}_s + \boldsymbol{\epsilon}_m$.

Graphical model for the proposed generative story of our PLDA based diarization model is illustrated in Figure 3.2. With those asserted generative story, we assume that speaker identity and intra-speaker accoustic subspaces span the entire i-vector space.

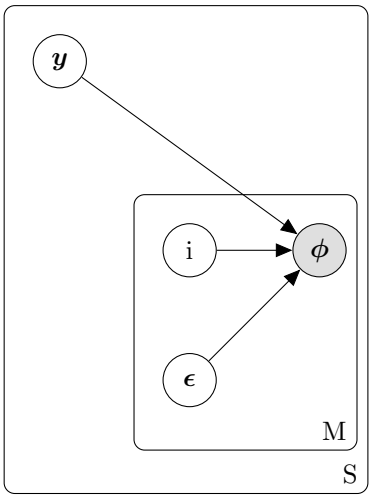

FIGURE 3.2: Graphical model for the proposed generative story of our PLDA based diarization system. Shaded node represents the observed variable while other nodes represent the latent variables. Outer plate representations denotes a set of M and S repeated occurences.

Let $\mathcal{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_S\}$ be the set of all 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, $P(\mathcal{Y}, \mathcal{I}|\Phi)$ produce intractable integrals. Therefore we resort to the approximate inference methods, namely mean-field VB, in order to approximate $P(\mathcal{Y}|\Phi)$ and $P(\mathcal{I}|\Phi)$.

3.4 Variational Bayes for PLDA based i-vectors

Supposing above probabilistic model we are given that observed variable Φ is generated with two hidden variables $\mathcal{I}$ and $\mathcal{Y}$. Lets denote hidden variables as $\theta = (\mathcal{I}, \mathcal{Y})$. Our generative story specifies the joint distribution $P(\Phi, \theta)$, and our goal is to find an approximation for the posterior distribution $P(\theta|\Phi)$ and for the model evidence $P(\Phi)$. We can define log marginal probability as,

$$\log P(\Phi) = \mathcal{L}(Q) + KL(Q\|P) \tag{3.7}$$

where we have defined variational lower bound of approximate posterior distribution and Kullback-Liebler (KL) divergence between true and approximate posterior distributions as follows:

$$\mathcal{L}(Q) = \int Q(\theta) \log \left\{ \frac{P(\Phi, \theta)}{Q(\theta)} \right\} d\theta, \tag{3.8}$$

$$KL(Q\|P) = -\int Q(\theta) \log \left\{ \frac{P(\theta|\Phi)}{Q(\theta)} \right\} d\theta. \tag{3.9}$$

In order to minimize KL divergence we can maximize the lower bound $\mathcal{L}(Q)$ by optimization with respect to $Q(\theta)$. The maximum of the lower bound occurs when the KL divergence vanishes, which occurs when $Q(\theta)$ equals the posterior distribution $P(\theta|\Phi)$. However, we propose a probabilistic model such that the true posterior distribution is intractable. In order to obtain both tractability and better approximation to the true posterior, we consider a restricted family of distributions for $Q(\theta)$ remember that $\theta = (\mathcal{I}, \mathcal{Y})$.

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

$$Q(\mathcal{Y}, \mathcal{I}) = Q(\mathcal{Y})Q(\mathcal{I}). \tag{3.10}$$

This factorized form of variational inference refers to an approximation framework

developed in physics called mean field theory [15].

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

$$Q(\mathcal{I}) = \prod_{m=1}^{M} \prod_{s=1}^{S} q_{ms}^{i_{ms}}, \tag{3.11}$$

$$Q(\mathcal{Y}) = \prod_{s=1}^{S} \mathcal{N}(\boldsymbol{y}_s | \boldsymbol{\mu}_s, \boldsymbol{C}_s^{-1}). \tag{3.12}$$

In equation (3.11), we define q_{ms} as the posterior probability of speaker s talking in segment m and in equation (3.12), it turns out that approximate segment and speaker posteriors factorize in the same way as marginal distribution of segment and speaker distributions. Notice that approximate segment posterior distributions are multinomial with probability q_{ms} and approximate speaker posterior distributions are Gaussian with mean $\boldsymbol{\mu}_s$ and precision $\boldsymbol{C}_s$. Actually, we do not assume approximate segment and speaker posteriors to be in this form, but rather we derive this result by assuming a factorization for approximate posterior as in equation (3.10), details can be found in Appendix A.

We can summarize update formulas of approximate segment and speaker posteriors for the VB approach as follows:

1. Update rule for segment posteriors:

$$q_{ms} = \frac{\tilde{q}_{ms}}{\sum_{s'=1}^{S} \tilde{q}_{ms'}} \tag{3.13}$$

where

$$\log \tilde{q}_{ms} = \boldsymbol{\mu}_s^T \boldsymbol{\mathcal{P}} \boldsymbol{\phi}_m - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\mathcal{P}}(\boldsymbol{C}_s^{-1} + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T)\right) + \text{const} \tag{3.14}$$

where const stands for speaker independent terms.

2. Update rule for speaker posteriors:

$$\boldsymbol{C}_s = \boldsymbol{\Lambda} + \sum_{m=1}^{M} q_{ms} \boldsymbol{\mathcal{P}}, \tag{3.15}$$

$$\boldsymbol{\mu}_s = \boldsymbol{C}_s^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{m=1}^{M} q_{ms} \boldsymbol{\mathcal{P}} \boldsymbol{\phi}_m). \tag{3.16}$$

The speaker and segment posteriors are updated alternately throughout the variational e-step. Formulation details of the update formulas can be found in Appendix A. On convergence, diarization is performed by assigning each segment m to the speaker given by $\underset{s}{\operatorname{argmax}}\ q_{ms}$ [5].

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 [5]. 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.

3.5 Initialization of VB Iterations

Firstly, we adopt a heuristic approach in order to initialize segment posteriors similar to the study in [5]. 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 [24, 28].

3.6 Deterministic Annealing variant of Variational Bayes

DA is introduced to the VB method in order to avoid trapping in poor local optimal solutions without using any heuristic approach as in Section 3.5. This process simply consists of introducing a temperature parameter, β to the free energy for controlling the annealing process deterministically [27]. The DA variant of update formulation in Section 3.4 can be adapted as follows:

$$\begin{aligned}\log \tilde{q}_{ms} = \beta(\boldsymbol{\mu}_s^T \boldsymbol{\mathcal{P}} \boldsymbol{\phi}_m - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\mathcal{P}}(\boldsymbol{C}_s^{-1} + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T)\right)) \\ + \text{const},\end{aligned} \tag{3.17}$$

$$\boldsymbol{C}_s = \beta(\boldsymbol{\Lambda} + \sum_{m=1}^{M} q_{ms} \boldsymbol{\mathcal{P}}), \tag{3.18}$$

$$\boldsymbol{\mu}_s = \boldsymbol{C}_s^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{m=1}^{M} q_{ms}\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m) \tag{3.19}$$

where the temperature parameter, β is initialized to be much smaller than one and increased during iterations until to be equal one.

By introducing the temperature parameter, β to the formulation, we attain a control on the convergence of the VB algorithm by decreasing the precision $\boldsymbol{C}_s$ (increasing the covariance $\boldsymbol{C}_s^{-1}$) of the speaker posterior distribution as seen in equation (3.18), notice that $\boldsymbol{\mu}_s$ does not changed, details of the formulation can be found in Appendix B.

Chapter 4

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. Details about the model training sets can be found in Table 4.1.

4.1 Segmentation

After extraction of MFCC features, we use BIC based penalized likelihood ratio test, following the recipe in Section 2.1.1, in order 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 [1] for detailed formulation and configurations.

4.2 K-means clustering i-vector System

As a baseline system we choose a system which applies k-means clustering on principal component analysis coefficients of segmental i-vectors[14]. This system is chosen because of the fact that it was reported to have superior results with respect to the earlier studies [1, 5]. According to the comparative DER results in [14], the benchmark study has a DER of 0.9% as opposed to the study [5] and [1] with DER of 3.5% and 1.0%, respectively on NIST SRE 2008 summed channel telephone data. The benchmark system has also other advantages such that it has a similar system architecture with our proposed system and it is easy to implement.

This system is based on the work described in [14]. 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.

4.3 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 LDA and 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 ($\approx$ 12 in total per utterance), in LDA and PLDA

training. We observe a minor improvement compared to training on multi-session full utterances. Details about the model training sets can be found in Table 4.1.

TABLE 4.1: Number of speakers and utterances used for training UBM/i-vector models and LDA/PLDA models.

	#speakers	#utterances
UBM/i-vector	1628	22419
LDA/PLDA	1218	15744

4.4 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. By applying this process in frame level we obtain an opportunity to recover the speaker error present in individual segments. 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 emission probabilities of frame likelihoods given two GMMs, to obtain final alignments.

Overall system diagram of our proposed speaker diarization system can be seen in Figure 4.1.

4.5 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 (M)* - classifying speech as non-speech, *False Alarm (FA)* - classifying non-speech as speech and *Speaker Error (E)* - confusing one speakers speech as from another [29]. Each type of error is illustrated in Figure 4.2. 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

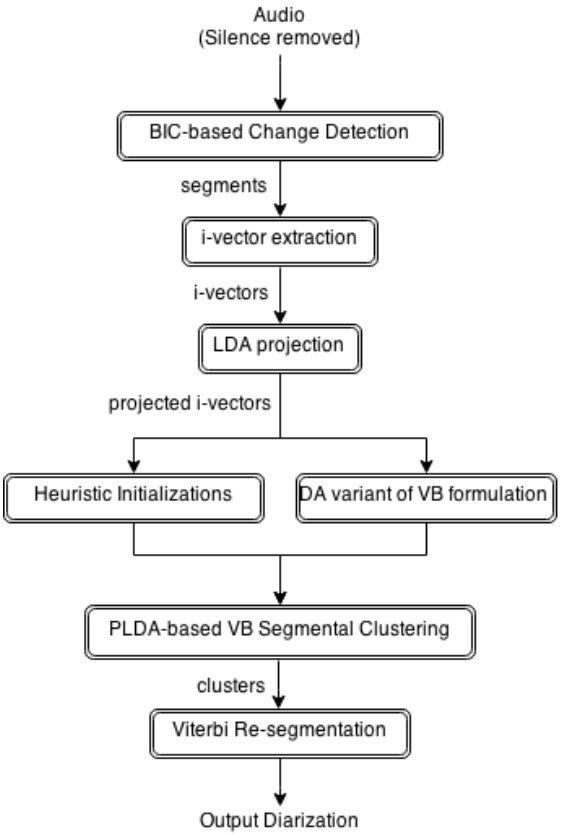

FIGURE 4.1: Main components of the proposed speaker diarization system.

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 as in the studies [5, 14].

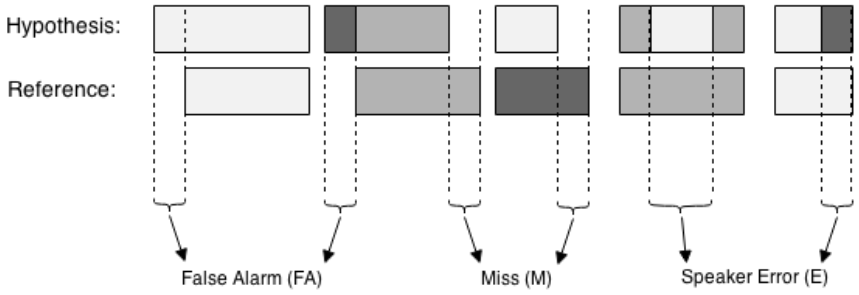

FIGURE 4.2: Illustration of each type of diarization error with reference and hypothesized system output [2].

4.6 Results

4.6.1 DER 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 as described in Section 3.1. Various experiments are conducted on a development set with different LDA projection dimensions. Results can be seen in Table 4.2.

TABLE 4.2: Results of our proposed system with various LDA projection dimensions on a development set.

	100	150	200	300
mean DER (%)	2.69	2.61	2.67	2.89
σ (%)	5.54	5.21	5.57	6.10

Analyzing the effect of applying LDA projection to the i-vectors, diarization performance increase constantly to some point by decreased dimension. As described in Section 3.1, LDA training set is structured in a way that intra-speaker acoustic variations are compensated and inter-speaker variations are become much more distinctive. Therefore, 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 which is considered ideal for k-means clustering [2].

Table 4.3 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. Our results for (KM-PCA) has higher DER than reported in [14] since we implemented a simplified version of the algorithm without multiple iteration of Viterbi resegmentation part and second pass of whole algorithm. In Table 4.3, we can see the effect of random initialization of speaker segments which clearly decreased the performance of proposed system as oppose to the baseline system.

TABLE 4.3: 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
σ (%)	5.83	9.16

Table 4.6 shows the results obtained from our proposed system with two different heuristic initializations, a DA variant of VB, and lastly again a VB system initialized with k-means algorithm. We use two metrics for initialization with cosine similarity (VB-COS) and PLDA log-likelihood ratio (VB-LLR) described in Section 3.5 for heuristic methods. 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 3.6, 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 $\beta_{new} < 1$. In order to obtain optimal values for the initial value of temperature parameter and the rate of increment for each iteration a variety of experiments are conducted on a development set of about, in total, one hour of telephone speech results can be seen in Table 4.4 and 4.5.

TABLE 4.4: Comparative results of DA-VB system with various initial value for temperature parameter with a fixed increment of 1.05 for each iteration on an $\approx 1h$ devset.

	0.001	0.01	0.2	0.8
mean DER (%)	3.58	3.55	1.34	1.36
σ (%)	11.37	11.37	1.96	1.97

TABLE 4.5: Comparative results of DA-VB system with various increment rate for each iteration with a fixed initial value of 0.2 for temperature parameter on an $\approx 1h$ devset.

	1.01	1.05	1.1	1.2
mean DER (%)	1.35	1.34	1.35	3.26
σ (%)	1.96	1.96	1.96	7.31

Analyzing the determination of optimal parameters of the DA algorithm, it is clear that DA algorithm improve performance according more to the initial value of the

temperature parameter than the increment rate. As described in Section 3.6, a relatively low value of initial temperature parameter decrease the precision which in turn increase the variance of the speaker posterior. This control of uncertainty on speaker space prevent VB algorithm from modelling the wrong speaker for the recordings that one speaker dominates the conversation.

TABLE 4.6: Comparative results of proposed systems with two different VB initializations, the DA variant of VB, and k-means initialized VB.

	VB-COS	VB-LLR	DA-VB	KM-VB
mean DER (%)	2.18	2.19	2.28	2.17
σ (%)	5.55	5.42	5.73	5.32

By using DA, we obtain comparable performance to the cumbersome heuristic initialization methods. And as a last attempt we try a further initialization with k-means clustering as described in Section 2.2.4 and then apply VB iterations (KM-VB). This last result can be considered as the fusion of baseline and proposed systems.

4.6.2 Error Analysis

Throughout the experiments we realized that about 50 test conversations have very high diarization error rate as opposed to remaining test conversations. When analyzing the properties of these utterances we observe that about 70% proportion is female-to-female conversations and 20% proportion is male-to-male conversations and the rest are female-to-male conversations. About 60% of the conversations are from far east countries, particularly tonal languages and the rest have generally very short speaker turns, background noise, and overlapping speaker parts in common. Actually, asserted properties of conversations make diarization problem very difficult to figure out such that even human ear can hardly determine which speaker is speaking at a particular time.

4.6.3 Run Time Performance

As for run time performance of the systems, we can divide overall diarization process into two main parts. The first part is segmental feature extraction part which is the part where segmental i-vectors are extracted. The first part is common to the benchmark study and our study. The second part is the segmental clustering part which includes the rest of the operations for obtaining the segment alignments over speakers. In Table 4.7, we present average run time performance of the segmental clustering part of the systems as a real time factor (RTF) for an approximately 5-minute conversation. RTF is a well-known metric in order to evaluate run time performance of a speech processing procedure, calculated as processing duration divided by input audio duration. Segmental feature extraction part has RTF of 0.1407 for all systems, however segmental clustering part differs in run time performance for benchmark and proposed systems.

TABLE 4.7: Comparative run time performance results of proposed systems and benchmark system in real time factor (RTF), with two different VB initializations and the DA variant of proposed VB system, baseline system and k-means initialized VB system.

	VB-COS	VB-LLR	DA-VB	KM-PCA	KM-VB
RTF	0.0042	0.0070	0.0032	0.0053	0.0095

It is clear that the feature extraction part dominates the speaker diarization process of a given utterance. It is approximately 30 times computationally more complex than the clustering part. Analyzing the clustering part of the systems, we can say that heuristic initializations detrimentally effect the computational complexity of the clustering part and also k-means algorithm takes longer time as opposed to the DA variant of VB which has the most effective performance of all. And clearly, the last attempt of k-means initialized VB system has the worst run time performance, yet it has the best diarization performance in terms of DER.

Chapter 5

Conclusion And Future Work

5.1 Conclusion

In this thesis, speaker based diarization of the telephone conversations which is one of the basic problems of the area of speech processing is discussed. Throughout the thesis, many methods were described and a VB diarization system is proposed with a novel initialization scheme using the DA method. As a starting point, in order to determine speaker change points a method [1] based on BIC is utilized. Then, general story behind the JFA and TVS are presented. Moreover, motivated by a previous study which utilizes factor analysis with a VB method [5], we develop a system that uses PLDA modelling with a VB method for inference in the speaker diarization problem. Also, a special preparation of training set proposed for the LDA and PLDA model in order to obtain best representation of speakers for the diarization of telephone conversations. At the final step, we successfully apply DA method to avoid the suboptimal heuristic initialization in VB. We obtain competitive performance as far as the study in [14] is concerned in our experiments.

5.2 Future Work

Actually, besides striving for a better system for improving the diarization performance, one can work on the detecting and excluding the overlapped speech segments in which two or more speakers speak at the same time [30, 31]. Since in this thesis we did not consider such analysis the overlapped segments assigned to the dominant speaker by contributing diarization error rate as speaker error.

In spite of the fact that the proposed system is tested on a database consisting two speaker telephone conversations, the formulation gives the system a chance to work for the databases consisting greater than two speakers. As a future work we assess the performance of our system on meeting and broadcast data involving, constantly, n number of speakers where $n > 2$.

Also, by going one step further our future efforts will continue to apply proposed system to meeting and broadcast data involving an unknown number of speakers. We are planning to apply a Bayesian nonparametric approach in order to estimate the number of speakers present in a given utterance like in the study [32] Fox et al. suggested.

Appendix A

PLDA-based Diarization System Variational Formulation

By assuming a factorized distribution for approximate posterior distributions as in equation (3.10) and by using the lower bound equation (3.8), which is a functional of the approximate posterior $Q(\theta)$, we can calculate the general variational mean-field update formulas for the approximate posteriors $Q(\mathcal{I})$ and $Q(\mathcal{Y})$. We would like to remind the reader that Φ denotes the observed variables (extracted i-vectors for segments) and $\mathcal{I}$ and $\mathcal{Y}$ are the hidden variables to be estimated, corresponding to the indicator variables for speakers at segments and the set of speaker vectors respectively. First we write $\mathcal{L}(Q)$ in terms of $Q(\mathcal{I})$ by keeping $Q(\mathcal{Y})$ constant, as follows [15]:

$$\begin{aligned}
\mathcal{L}(Q) &= \int Q(\theta) \left\{ \log P(\Phi, \theta) - \log Q(\theta) \right\} d\theta \\
&= \int Q(\mathcal{I}) \left\{ \int \log P(\Phi, \mathcal{I}, \mathcal{Y}) Q(\mathcal{Y}) d\mathcal{Y} \right\} d\mathcal{I} \\
&\quad - \int Q(\mathcal{I}) \log Q(\mathcal{I}) d\mathcal{I} + \text{const} \\
&= \int Q(\mathcal{I}) \log \tilde{P}(\Phi, \mathcal{I}) d\mathcal{I} - \int Q(\mathcal{I}) \log Q(\mathcal{I}) d\mathcal{I} + \text{const}
\end{aligned} \tag{A.1}$$

where we define a new distribution by setting

$$\log \tilde{P}(\Phi, \mathcal{I}) = \mathbb{E}_{\mathcal{Y}}[\log P(\Phi, \mathcal{I}, \mathcal{Y})] + \text{const.} \tag{A.2}$$

Here $\mathbb{E}_{\mathcal{Y}}[\,.\,]$ denotes the expectation with respect to the $Q(\mathcal{Y})$ so that

$$\mathbb{E}_{\mathcal{Y}}[\log P(\Phi, \mathcal{I}, \mathcal{Y})] = \int \log P(\Phi, \mathcal{I}, \mathcal{Y}) Q(\mathcal{Y}) d\mathcal{Y}. \tag{A.3}$$

As described in Section 3.4 our goal is to maximize $\mathcal{L}(Q)$ in order to find a better approximate posterior distribution. Note that we first suppose $Q(\mathcal{Y})$ is fixed and we would like to maximize $\mathcal{L}(Q)$ as calculated in equation (A.1) with respect to $Q(\mathcal{I})$. This can be achieved by recognizing that $\mathcal{L}(Q)$ is equal to the negative KL divergence between $Q(\mathcal{I})$ and $P(\Phi, \mathcal{I}, \mathcal{Y})$ as can be seen at the last step of the equation (A.1). Thus maximizing $\mathcal{L}(Q)$ is equivalent to minimizing KL divergence, and the minimum occurs when $Q(\mathcal{I}) = \tilde{P}(\Phi, \mathcal{I})$. Hence we obtain a general expression for the optimal solution as follows:

$$\log Q(\mathcal{I}) = \mathbb{E}_{\mathcal{Y}}[\log P(\Phi, \mathcal{I}, \mathcal{Y})] + \text{const} \tag{A.4}$$

we can calculate $Q(\mathcal{Y})$ in a similar manner by fixing $Q(\mathcal{I})$ and maximizing $\mathcal{L}(Q)$ with respect to $Q(\mathcal{Y})$,

$$\log Q(\mathcal{Y}) = \mathbb{E}_{\mathcal{I}}[\log P(\Phi, \mathcal{I}, \mathcal{Y})] + \text{const} \tag{A.5}$$

where constants are chosen so as to ensure that the total probability of distributions sum up to one. Now, in order to calculate approximate posteriors in particular we have to calculate joint probability as a first step.

Since we assume that hidden variables $\mathcal{I}$ and $\mathcal{Y}$ are independent from each other we can write joint probility distribution as follows:

$$P(\Phi, \mathcal{I}, \mathcal{Y}) = P(\Phi|\mathcal{I}, \mathcal{Y})P(\mathcal{Y})P(\mathcal{I}). \tag{A.6}$$

We define the marginal probabilities of hidden variables as in the generative story in Section 3.3 as follows:

$$P(\mathcal{I}) = \prod_m \prod_s \pi_s^{i_{ms}} \tag{A.7}$$

$$P(\mathcal{Y}) = \prod_s P(\boldsymbol{y}_s) \tag{A.8}$$

where π_s be the prior probability of the speaker and i_{ms} is the indicator vector for a given segment as defined in Section 3.3 and also we can define conditional probability of $P(\Phi|\mathcal{I}, \mathcal{Y})$ as follows:

$$P(\Phi|\mathcal{I}, \mathcal{Y}) = \prod_m \prod_s P(\boldsymbol{\phi}_m|\boldsymbol{y}_s)^{i_{ms}}. \tag{A.9}$$

Since $\boldsymbol{y}_s$ and $\boldsymbol{\epsilon}_m$ Gaussian with $\boldsymbol{y}_s \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ and $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\mathcal{P}}^{-1})$ we can conclude by generative model in Section 3.3 that $P(\boldsymbol{\phi}_m|\boldsymbol{y}_s)$ is Gaussian with:

$$P(\boldsymbol{\phi}_m|\boldsymbol{y}_s) \sim \mathcal{N}(\boldsymbol{y}_s, \boldsymbol{\mathcal{P}}^{-1}). \tag{A.10}$$

Then by using the definition of multivariate Gaussian distribution [15] and equations (A.6), (A.7), (A.8), (A.9), and (A.10) we can calculate log-joint probability, $\log P(\Phi, \mathcal{I}, \mathcal{Y})$ as:

$$\begin{aligned}\log P(\Phi, \mathcal{I}, \mathcal{Y}) = &-\frac{1}{2}\sum_m\sum_s i_{ms}\Big\{-\log|\boldsymbol{\mathcal{P}}| + \mathrm{tr}(\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^T) - 2y_s^T\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m + \mathrm{tr}(\boldsymbol{\mathcal{P}}\boldsymbol{y}_s\boldsymbol{y}_s^T)\Big\} \\ &-\frac{1}{2}\sum_s\Big\{-\log|\boldsymbol{\Lambda}| + \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{y}_s\boldsymbol{y}_s^T) - 2\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{y}_s + \boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\Big\} \\ &+\sum_m\sum_s i_{ms}\log\pi_s + \text{const.}\end{aligned} \tag{A.11}$$

Now we can calculate the update formula for $Q(\mathcal{I})$ by using the equation (A.4) as follows, note that we ignore the terms independent of $\mathcal{I}$ throughout the calculations:

$$\begin{aligned}\log Q(\mathcal{I}) &= \mathbb{E}_{\mathcal{Y}}[\log P(\Phi|\mathcal{I}, \mathcal{Y})] + \log P(\mathcal{I}) \\ &= \sum_m\sum_s i_{ms}\Big\{\frac{1}{2}\log|\boldsymbol{\mathcal{P}}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^T) + \mathbb{E}_{\boldsymbol{y}_s}[\boldsymbol{y}_s^T]\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m \\ &\quad -\frac{1}{2}\mathrm{tr}(\boldsymbol{\mathcal{P}}\mathbb{E}_{\boldsymbol{y}_s}[\boldsymbol{y}_s\boldsymbol{y}_s^T]) + \log\pi_s + \text{const}\Big\}\end{aligned} \tag{A.12}$$

We obtain that posterior distribution has the same form as the prior distribution. By this observation, we define unnormalized segment posterior, $\tilde{q}_{ms}$ as:

$$\begin{aligned}\log \tilde{q}_{ms} = \frac{1}{2}\log|\boldsymbol{\mathcal{P}}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^T) + \mathbb{E}_{\boldsymbol{y}_s}[\boldsymbol{y}_s^T]\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m \\ - \frac{1}{2}\mathrm{tr}(\boldsymbol{\mathcal{P}}\mathbb{E}_{\boldsymbol{y}_s}[\boldsymbol{y}_s\boldsymbol{y}_s^T]) + \log\pi_s + \text{const.}\end{aligned} \tag{A.13}$$

Normalizing so that segment posteriors sum to one, gives:

$$Q(\mathcal{I}) = \prod_m \prod_s q_{ms}^{i_{ms}} \tag{A.14}$$

where q_{ms} is defined as equation (3.13).

We can also calculate the update formula for $Q(\mathcal{Y})$ by using the equation (A.5) as follows, note that we ignore the terms independent of $\mathcal{Y}$ throughout the calculations:

$$\begin{aligned}\log Q(\mathcal{Y}) &= \mathbb{E}_{\mathcal{I}}[\log P(\Phi|\mathcal{I},\mathcal{Y})] + \log P(\mathcal{I}) \\ &= -\frac{1}{2}\sum_m\sum_s \mathbb{E}_{\mathcal{I}}[i_{ms}]\Big\{-\log|\boldsymbol{\mathcal{P}}| + \mathrm{tr}(\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^T) - 2\boldsymbol{y}_s^T\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m + \mathrm{tr}(\boldsymbol{\mathcal{P}}\boldsymbol{y}_s\boldsymbol{y}_s^T)\Big\} \\ &\quad -\frac{1}{2}\sum_s\Big\{-\log|\boldsymbol{\mathcal{P}}| + \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{y}_s\boldsymbol{y}_s^T) - 2\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{y}_s + \boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\Big\} \\ &\quad + \text{const}\end{aligned} \tag{A.15}$$

Now we will write approximate log-posterior of y_s for each speaker by knowing that $\mathbb{E}_{\mathcal{I}}[i_{ms}] = q_{ms}$ as in A.16, note that we ignore the terms independent of $\boldsymbol{y}_s$ throughout the calculations:

$$\begin{aligned}\log Q(\boldsymbol{y}_s) = & \sum_m \left\{ q_{ms} \left(\boldsymbol{y}_s^T \boldsymbol{\mathcal{P}} \boldsymbol{\phi}_m - \frac{1}{2} \boldsymbol{y}_s^T \boldsymbol{\mathcal{P}} \boldsymbol{y}_s \right) \right\} \\ & - \frac{1}{2} \boldsymbol{y}_s^T \boldsymbol{\Lambda} \boldsymbol{y}_s + \boldsymbol{y}_s \boldsymbol{\Lambda} \boldsymbol{\mu} + \text{const.}\end{aligned} \tag{A.16}$$

We observe that the right-hand side of this expression is a quadratic function of $\boldsymbol{y}_s$, so we can identify the distribution as Gaussian. Hence we obtain that posterior distribution has the same form as the prior distribution. Now by using the completing the square trick [15], we can equate the right-hand side of the equation to the form

$$-\frac{1}{2} \boldsymbol{y}_s^T \boldsymbol{C}_s \boldsymbol{y}_s + \boldsymbol{y}_s \boldsymbol{C}_s \boldsymbol{\mu}_s. \tag{A.17}$$

By following necessary calculations the approximate posterior mean, $\boldsymbol{\mu}_s$ and precision, $\boldsymbol{C}_s$ of $\boldsymbol{y}_s$ can be calculated as follows:

$$\boldsymbol{C}_s = \boldsymbol{\Lambda} + \sum_{m=1}^{M} q_{ms} \boldsymbol{\mathcal{P}}, \tag{A.18}$$

$$\boldsymbol{\mu}_s = \boldsymbol{C}_s^{-1} (\boldsymbol{\Lambda} \boldsymbol{\mu} + \sum_{m=1}^{M} q_{ms} \boldsymbol{\mathcal{P}} \boldsymbol{\phi}_m). \tag{A.19}$$

After obtaining approximate posterior mean and covariance we can calculate the posterior expectations in equation (A.12) as:

$$\mathbb{E}_{\boldsymbol{y}_s}[\boldsymbol{y}_s] = \boldsymbol{\mu}_s, \tag{A.20}$$

$$\mathbb{E}_{\boldsymbol{y}_s}[\boldsymbol{y}_s \boldsymbol{y}_s^T] = \boldsymbol{C}_s^{-1} + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T. \tag{A.21}$$

Appendix B

Deterministic Annealing Variant of Variational Bayes Formulation

In deterministic annealing (DA) variant of VB formulation, by introducing a temperature parameter, β to the free energy, approximate segment and speaker posterior distribution equations (A.4) and (A.5) take the form as follows [27]:

$$\log Q(\mathcal{I}) = \mathbb{E}_{\mathcal{Y}}[\beta \log P(\Phi, \mathcal{I}, \mathcal{Y})] + \text{const}, \tag{B.1}$$

$$\log Q(\mathcal{Y}) = \mathbb{E}_{\mathcal{I}}[\beta \log P(\Phi, \mathcal{I}, \mathcal{Y})] + \text{const}. \tag{B.2}$$

After following same steps as in Appendix A one can deduce the update formulas for segment and speaker posteriors as follows:

$$\begin{aligned} \log \tilde{q}_{ms} = \beta(\boldsymbol{\mu}_s^T \boldsymbol{\mathcal{P}} \boldsymbol{\phi}_m - \frac{1}{2}\text{tr}\left(\boldsymbol{\mathcal{P}}(\boldsymbol{C}_s^{-1} + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T)\right)) \\ + \text{const}, \end{aligned} \tag{B.3}$$

$$\boldsymbol{C}_s = \beta(\boldsymbol{\Lambda} + \sum_{m=1}^{M} q_{ms}\boldsymbol{\mathcal{P}}), \tag{B.4}$$

$$\boldsymbol{\mu}_s = \boldsymbol{C}_s^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{m=1}^{M} q_{ms}\boldsymbol{\mathcal{P}}\boldsymbol{\phi}_m) \tag{B.5}$$

notice that in equation (B.5) β is canceled out with the one in the inverse of speaker precision equation (B.4).

Bibliography

[1] 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.

[2] Stephen Shum. Unsupervised methods for speaker diarization. Master's thesis, Massachusetts Institute of Technology, 2011.

[3] 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*, pages 4982–4985, 2010. doi: 10.1109/ICASSP.2010.5495077. URL http://dx.doi.org/10.1109/ICASSP.2010.5495077.

[4] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1505–1512, 2006. doi: 10.1109/TASL.2006.878261. URL http://doi.ieeecomputersociety.org/10.1109/TASL.2006.878261.

[5] Patrick Kenny, Douglas A. Reynolds, and Fabio Castaldo. Diarization of telephone conversations using factor analysis. *J. Sel. Topics Signal Processing*, 4(6):1059–1070, 2010. doi: 10.1109/JSTSP.2010.2081790. URL http://dx.doi.org/10.1109/JSTSP.2010.2081790.

[6] 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*, pages 347–350, 2011. doi: 10.1109/IDAACS.2011.6072771. URL http://dx.doi.org/10.1109/IDAACS.2011.6072771.

[7] Scott Chen and Ponani Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, page 8. Virginia, USA, 1998.

[8] Kirill Sakhnov, Ekaterina Verteletskaya, and Boris Simak. Approach for energy-based voice detector with adaptive scaling factor. *IAENG International Journal of Computer Science*, 36(4):394, 2009.

[9] B. Zhou and Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *In Int. Conf. on Spoken Language Processing*. ICSLP, 2000.

[10] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair. Stream-based speaker segmentation using speaker factors and eigenvoices. In *ICASSP*, pages 4133–4136. IEEE, 2008. ISBN 1-4244-1484-9. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2008.html#CastaldoCDLV08.

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

[12] 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*, 16(5):980–988, 2008. ISSN 1558-7916. doi: 10.1109/TASL.2008.925147.

[13] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[14] 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*, pages 945–948, 2011. URL http://www.isca-speech.org/archive/interspeech_2011/i11_0945.html.

[15] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738.

[16] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. Speaker and session variability in GMM-based speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1448–1460, 2007.

[17] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[18] Patrick Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[19] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Interspeech*, volume 9, pages 1559–1562, 2009.

[20] Patrick Kenny. A small foot-print i-vector extractor. In *Proc. Odyssey*, 2012.

[21] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel. Eigenvoice modeling with sparse training data. *Speech and Audio Processing, IEEE Transactions on*, 13(3):345–354, 2005.

[22] Patrick Kenny. Bayesian analysis of speaker diarization with eigenvoice priors. *CRIM, Montreal, Technical Report*, 2008.

[23] Ekapol Chuangsuwanich, Scott Cyphers, James Glass, and Seth Teller. Spoken command of large mobile robots in outdoor environments. In *Spoken*

Language Technology Workshop (SLT), 2010 IEEE, pages 306–311. IEEE, 2010.

[24] 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*, pages 1–8, 2007. doi: 10.1109/ICCV.2007.4409052. URL http://dx.doi.org/10.1109/ICCV.2007.4409052.

[25] 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*, pages 1–20, 2010. URL http://www.fit.vutbr.cz/research/view_pub.php?id=9346.

[26] 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*, pages 5272–5275. IEEE, 2011. ISBN 978-1-4577-0539-7. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2011.html#SturimCDKMRRTS11.

[27] K Katahira, K Watanabe, and M Okada. Deterministic annealing variant of variational Bayes method. *Journal of Physics: Conference Series*, 95-012015 (1), 2008. URL http://stacks.iop.org/1742-6596/95/i=1/a=012015.

[28] Niko Brümmer and Edward de Villiers. The speaker partitioning problem. In *Odyssey 2010: The Speaker and Language Recognition Workshop*, page 34, 2010. URL http://www.isca-speech.org/archive_open/odyssey_2010/od10_034.html.

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

[30] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. Overlapped speech detection for improved speaker diarization in multiparty

meetings. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4353–4356. IEEE, 2008.

[31] Jürgen T Geiger, Ravichander Vipperla, Nicholas Evans, Björn Schuller, and Gerhard Rigoll. Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 340–344. IEEE, 2012.

[32] Emily B Fox, Erik B Sudderth, Michael I Jordan, Alan S Willsky, et al. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A):1020–1056, 2011.