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

Latent Class Model for Single Channel Speaker Diarization

Liang He, Xianhong Chen, Can Xu, and Jia Liu

Department of Electronic Engineering, Tsinghua University, Beijing, China.
heliang@mail.tsinghua.edu.cn

Abstract

Inspired by P. Kenny's variational Bayes (VB) method, we derive a latent class model (LCM) for single channel speaker diarization. Similar to the VB method, the LCM uses soft information and avoids premature hard decisions in its iterations. Different from the VB method, the LCM provides an iterative framework for multi-objective optimization and allows a more flexible way to compute the probability that given a speaker, a segment occurs. Based on this model, we propose a latent class model-i-vector-probabilistic linear discriminant analysis (LCM-Ivec-PLDA) system. Besides, as the divided segments are very short, their neighbors are taken into consideration. To overcome the initial sensitivity problem, we use an agglomerative hierarchical cluster (AHC) to do initialization and present hard and soft priors. Experiments on the NIST RT09 speaker diarization database and our collected database show that the proposed systems are superior to the traditional VB system.

1. Introduction

Speaker diarization aims to address the problem of "who spoke when" and is able to split an utterance into homogeneous regions by speaker identities [1].

Generally, there are three stages in a typical speaker diarization system: voice activity detection (VAD), in which non-speech (silence or noise) segments are removed; speaker segmentation, in which an audio recording is usually split into speaker homogeneous segments; and speaker clustering, in which the divided segments belonging to the same speaker are grouped into a cluster [2].

The variational Bayes (VB) system is proposed by P. Kenny [3–6]. This system has two characteristics. First, unlike the mainstream approach, it uses a uniform segmentation instead of speaker change point detection to do speaker segmentation. As the segmented segment is short enough, each segment can be regarded as containing only one speaker. Second, it utilizes a soft clustering approach that avoids premature hard decision and suppresses the propagation of errors. Despite its superior performance, there are some deficiencies. The VB method is a single-objective method. Its goal is to increase the overall likelihood, not to distinguish speakers. As the segmented segments are very short, the $p(x_m|y_s)$ which represents the probability that given a speaker s, the segment x_m occurs, may be inaccurate and degrades the system performance. In addition, some researchers have also noted that, the VB system is sensitive to its initialization [7]. When one speaker dominates the recording, a random prior tends to assign the segments to each speaker evenly, leading to a poor result.

We derive a latent class model which establishes an iterative framework for multi-objective optimization that allows us to compute $p(x_m|y_s)$ in a more flexible and discriminant way. We present a latent class model-i-vector-probabilistic linear discriminant analysis [8] (LCM-Ivec-PLDA) system subsequently. To address the problem caused by the shortness of each segment, we take x_m's neighbors into account to improve the accuracy of $p(x_m|y_s)$ by considering speaker temporal relevance. Based on the segmental i-vectors, we use an agglomerative hierarchical cluster (AHC) [9] to handle with the initial sensitivity problem. Experiments on RT09 and our own database show that the proposed system has a better performance compared with the VB system [10].

The remainder of this paper is organized as follows. Section 2 introduces the latent class model. Section 3 presents LCM-Ivec-PLDA speaker diarization system. Section 4 analyzes and discusses the experiment results. A conclusion is drawn in Section 5.

2. Latent Class Model

Suppose a sequence X is divided into fixed length short segments $\{x_m\}$, where the subscript m is the time index, $1 \leq m \leq M$. Let $Y = \{y_s\}$ be parameters of each latent class model, where the subscript s is the class index, $1 \leq s \leq S$. We define the latent class matrix $Q = \{q_{ms}\}$ as the probability of class s existing at the time m and have a constraint $\sum_{s=1}^{S} q_{ms} = 1, q_{ms} \geq 0$, see Fig. 1.

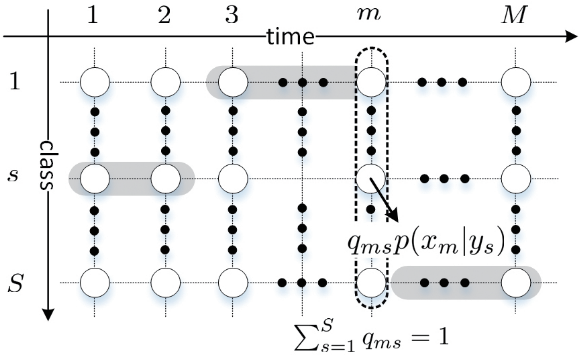

Figure 1: Latent class model

X is the observable data and Q, Y are hidden variables. Maximum likelihood estimation is used. The goal is to find proper Q and Y to maximize $\log p(X)$.

$$\log p(X) = \left[\sum_{m=1}^{M} \log \left(\sum_{s=1}^{S} q_{ms} p(x_m|y_s) \right) \right] \qquad (1)$$

where $p(x_m|y_s)$ is the conditional probability that given y_s, x_m occurs, $P = \{p(x_m|y_s)\}$.

Following the classical expectation−maximization (EM) algorithm [11], we obtain its lower bound by Jensen's inequali-

10.21437/Odyssey.2018-18

ty [12]

$$\begin{aligned}&\sum_{m=1}^{M} \log\left(\sum_{s=1}^{S} q_{ms} p(x_m|y_s)\right)\\ &\geq \sum_{m=1}^{M}\sum_{s=1}^{S}\Gamma_{ms}\log\left(\frac{q_{ms}p(x_m|y_s)}{\Gamma_{ms}}\right)\end{aligned} \tag{2}$$

where Γ_{ms} is an auxiliary function, $\sum_{s=1}^{S}\Gamma_{ms} = 1$ and $\Gamma_{ms} \geq 0$. Thus, the optimization of the objective function is decomposed into E step and M step.

In the E step, we calculate the Γ_{ms} which maximizes the log likelihood function given Q and P

$$\Gamma_{ms} = \frac{q_{ms}p(x_m|y_s)}{\sum_{s'=1}^{S} q_{ms'}p(x_m|y_{s'})} \tag{3}$$

In the M step, we compute Q, Y and P. To calculate Q, the Γ and P are assumed to be known. Maximizing $\sum_{m=1}^{M}\sum_{s=1}^{S}\Gamma_{ms}\log q_{ms}$ with constraint $\sum_{s=1}^{S} q_{ms} = 1, q_{ms} \geq 0$, we get its solution by the Lagrangian multiplier method.

$$q_{ms} = \Gamma_{ms} \tag{4}$$

Note that, we do not limit $p(x_m|y_s)$ to a specific model in the above derivation. That means different statistical models can be applied in (3) depending on the specified motivations.

3. LCM-Ivec-PLDA Speaker Diarization System

We use two methods alternately to compute P and build a hybrid iterative model, as shown in Fig. 2. In the inner loop, to find the optimal Y, we compute P according to the VB method proposed in the Kenny's paper [3]. In the outer loop, to discriminate speakers, we compute P by introducing PLDA [8,13,14].

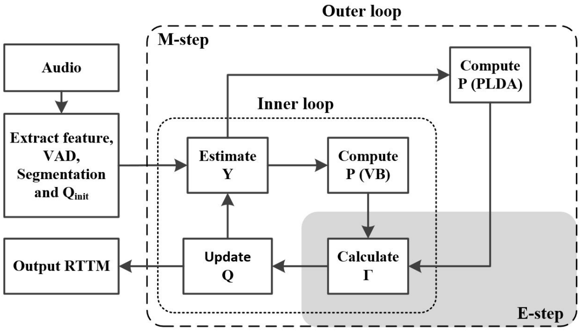

Figure 2: Flow chart of LCM-Ivec-PLDA speaker diariazation system

3.1. Estimate Y

To find the optimal Y, we calculate $p(x_m|y_s)$ according to VB method [3]. Let T denote the total variability space, the i-vector model [15] is as follows

$$p_{\text{GMM-Ivec}}(x_m|y_s) = \sum_{c=1}^{C} w_c \mathcal{N}(x_m|\mu_{\text{ubm},c} + T_c y_s, \Sigma_c) \tag{5}$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ is a Gaussian distribution with a mean vector μ and a covariance matrix Σ. w_c is weight. c is Gaussian mixture index, $1 \leq c \leq C$.

Different from speaker recognition in which all the X are assumed to be from one speaker, the x_m belongs to a speaker s with the probability q_{ms} in the case of speaker diariazation. It is necessary to consider this when estimating y_s. Submit (5) into (2), we use the Jensen's inequality again and obtain the lower bound of $\log p(X)$ as follows

$$\begin{aligned}&\log p(X) \geq\\ &\sum_{m=1}^{M}\sum_{s=1}^{S} q_{ms}\sum_{c=1}^{C}\gamma_{\text{ubm},mc}\log\mathcal{N}(x_m|\mu_{\text{ubm},c} + T_c y_s, \Sigma_c)\end{aligned} \tag{6}$$

where

$$\gamma_{\text{ubm},mc} = \frac{w_c\mathcal{N}(x_m|\mu_{\text{ubm},c}, \Sigma_{\text{ubm},c})}{\sum_{c'=1}^{C} w_{c'}\mathcal{N}(x_m|\mu_{\text{ubm},c'}, \Sigma_{\text{ubm},c'})} \tag{7}$$

The above objective function is a quadratic optimization problem with the optimal solution

$$y_s = (\varepsilon I + T^t \hat{N}_s \Sigma^{-1} T)^{-1} T^t \Sigma^{-1} \hat{F}_s \tag{8}$$

where $\hat{N}_s$, $\hat{F}_s$, Σ, and T are concatenations of $\hat{N}_{sc}$, $\hat{F}_{sc}$, Σ_c, and T_c, respectively. ε is a small positive constant. I is an identity matrix. The $\hat{N}_{sc}$, $\hat{F}_{sc}$ are defined as follows

$$\begin{aligned}\hat{N}_{sc} &= \sum_{m=1}^{M} q_{ms}\gamma_{\text{ubm},mc}\\ \hat{F}_{sc} &= \sum_{m=1}^{M} q_{ms}\gamma_{\text{ubm},mc}(x_m - \mu_{\text{ubm},c})\end{aligned} \tag{9}$$

In the above estimation, the T and Σ are assumed to be known. They can be estimated on a large auxiliary database in a traditional way.

3.2. Compute P (PLDA)

As y_s is obtained, $p(x_m|y_s)$ can be computed in Kenny's paper [3] as follows:

$$\text{ln}p(x_m|y_s) = G_m + H_{ms} \tag{10}$$

where

$$G_m = \sum_{c=1}^{C} N_{mc}\text{ln}\frac{1}{(2\pi)^{F/2}|\Sigma_c|^{1/2}} - \frac{1}{2}\text{tr}\left(\Sigma^{-1}S_m\right)$$

$$H_{ms} = y_s^{\mathrm{T}}T^{\mathrm{T}}\Sigma^{-1}F_m - \frac{1}{2}y_s^{\mathrm{T}}T^{\mathrm{T}}N_m\Sigma^{-1}Ty_s$$

N_m, F_m, and S_m are the zero, first, and second Baum-Welch statistics of segment x_m, respectively.

However, it does not take into account the exclusion of channel interference. To discriminate different speakers, we introduce a PLDA to improve its performance.

For each segment x_m, we extract an i-vector y_m. Based on the simplified PLDA model [16], the $p(x_m|y_s)$ can be calculated by the log likelihood ratio.

$$p(x_m|y_s) = \log\frac{P(y_m, y_s|\theta_{tar})}{P(y_m, y_s|\theta_{non})} \tag{11}$$

where θ_{tar} (θ_{non}) represents that y_m and y_s are from the same (different) speaker.

As we mentioned above, each segment x_m is usually very short to ensure its speaker homogeneity. However, this shortness will lead to inaccuracy when calculating the $p(x_m|y_s)$. Intuitively, if a speaker s appears at time m, the speaker will appear at a great probability in the vicinity of time m. We can take advantages of x_m's neighbors to improve the accuracy of $p(x_m|y_s)$. Let $X_m = \{x_{m-\Delta M}, \cdots, x_m, \cdots, x_{m+\Delta M}\}$, where $\Delta M > 0$. We use X_m instead of only x_m to extract i-vector y_m to represent x_m and then use (11) to get $p(x_m|y_s)$, as shown in Fig. 3, the bottom part. Although X_m may contain more than one speaker, it does not matter. Because this $p(x_m|y_s)$ reflects the probability that the speaker s appears at the time m, not at the time range $(-\Delta M, \cdots, \Delta M)$. Besides, X_m is long enough to ensure more robust estimates, thereby improving the system performance.

3.3. Update Q

After the computation of $p(x_m|y_s)$, the Q is updated by (3). Here, we again use x_m's neighbors to boost the performance. Suppose $x_{m+\Delta m}$ is the neighbor of x_m. Given the condition that x_m belongs to the speaker s, we consider the probability that $x_{m+\Delta m}$ does not belong to the speaker s. If we define the appearance of speaker change point as an event, the above process can be approximated as a homogeneous Poisson point process [17]. Under this assumption, the probability that a speech segment from time m to time $m + \Delta m$ belongs to the same speaker is equivalent to the probability that the speaker change point does not appear from time m to time $m + \Delta m$.

$$p(\Delta m) = e^{-\lambda \Delta m}, \Delta m \geq 0 \tag{12}$$

where λ is the rate parameter. We consider the contribution of $p(x_{m+\Delta m}|y_s)$ to $p(x_m|y_s)$ by updating $p(x_m|y_s)$ as follows, see Fig. 3, the top part.

$$p(x_m|y_s) \leftarrow \sum_{\Delta m=-\Delta M}^{\Delta M} [p(\Delta m) p(x_{m+\Delta m}|y_s)] \tag{13}$$

After the PLDA scoring (11), we use (13) to update the $p(x_m|y_s)$. The q_{ms} is subsequently updated according to (3).

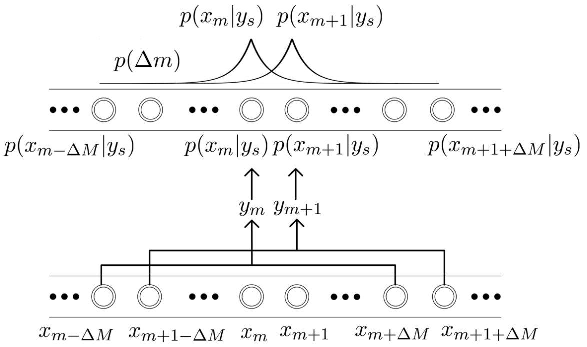
$p(x_m|y_s)$ $p(x_{m+1}|y_s)$
$p(\Delta m)$
$p(x_{m-\Delta M}|y_s)$
$p(x_m|y_s)$ $p(x_{m+1}|y_s)$
$p(x_{m+1+\Delta M}|y_s)$
y_m y_{m+1}
$x_{m-\Delta M}$ $x_{m+1-\Delta M}$ x_m x_{m+1} $x_{m+\Delta M}$ $x_{m+1+\Delta M}$

Figure 3: Taking neighbors into account to update $p(x_m|y_s)$

3.4. AHC Initialization

Although random initialization works well in most cases, it tends to assign the segments to each speaker evenly, leading to bad results when a speaker dominates the whole conversation. To address it, we recommend a more robust and informative AHC initialization method. After using PLDA to compute the log likelihood ratio between two segment i-vectors [18, 19], AHC is applied to get the clustering results. Based on the AHC results, two prior calculation methods, hard prior and soft prior, are proposed.

3.4.1. Hard Prior

According to the AHC clustering results, if segment x_m is classified to speaker s, we will assign q_{ms} with a relatively larger value q. The hard prior is as follows:

$$q_{ms} = I(x_m \in s)q + I(x_m \notin s)\frac{1-q}{S-1} \tag{14}$$

where $I(\cdot)$ is the indicator function. $I(x_m \in s)$ means x_m is spoken by speaker s.

3.4.2. Soft Prior

For soft prior, we first calculate the center of each estimated cluster s:

$$s_{\text{center}} = \frac{\sum_m I(x_m \in s) y_m}{\sum_m I(x_m \in s)} \tag{15}$$

If segment x_m belongs to cluster s, the distance between segment x_m and s_{center} is

$$d_{ms} = \|y_m - s_{\text{center}}\|_2 \tag{16}$$

The prior that segment x_m is spoken by speaker s is

$$q_{ms} = \frac{1}{2}[\frac{e^{-(\frac{d_{ms}}{d_{\max,s}})^k} - e^{-1}}{1-e^{-1}} + 1] \tag{17}$$

where $d_{\max,s} = \max_{x_m \in s}(d_{ms})$, k is a constant value. This soft prior varies from 0.5 to 1 and makes sure that if segment x_m is closer to s_{center}, q_{ms} will be larger. For other speakers, the prior is

$$q_{mj|j\neq s} = \frac{1-q_{ms}}{S-1} \tag{18}$$

4. Experiments

4.1. Database

We do experiments on the NIST Rich Transcription Meeting Recognition Evaluation 2009 (RT09) and our collected database.

The NIST RT09 SPKD evaluation database has 7 meeting audio recordings and around 3 hours in length. Only one channel recording from a single distant microphone (SDM) audio recording is used to demonstrate our single channel speaker diarization algorithm. All of them are English.

The training part of our collected database contains 57 speakers (30 female and 27 male). The total duration is about 94 hours. All of them are natural conversations (Mandarin) recorded in a quiet office condition. The evaluation part has 3 audio recordings (*TL 7-9*). They are also recorded in a quiet office, but there is one speaker who dominates the whole conversation ($>$ 80% in length). Each recording has two speakers and is about 20 minutes.

All the above audio recordings are converted to 8kHz, 16bits PCM format.

4.2. Configuration and Parameters

Perceptual linear predictive (PLP) features with 19 dimensions are extracted from audio recordings with a 25 ms Hamming window and a 10 ms stride. PLP and log-energy constitute a 20 dimensional basic feature. This basic feature and its first derivatives are concatenated as our acoustic feature. Speech/silence segmentation (VAD) is executed by the frame log-energy and subband spectral entropy.

The UBM is composed of 512 diagonal Gaussian components. The rank of total variability matrix T is 300. ε in (8) is 0.1. For the PLDA, the rank of subspace matrix is 150. For x_m's neighbors, ΔM is 40 and λ is 0.005. We use a forward-backward algorithm to smooth Q. The loop probability is 0.998 and the non-loop probabilities are equal. In the AHC initialization, q is set to be 0.7 in the hard prior setting and k is 10 in the soft prior setting. Speaker number is assumed to be known in advance.

For the RT09, we use Switchboard-P1, RT05 and RT06 to train UBM, T and PLDA parameters. For the *TL 7-9*, we use the training part to train the above parameters.

We adopt speaker error rate (SPKR) and diarization error rate (DER) to measure the system performance according to the RT09 evaluation plan [10].

4.3. Experiment Results and Discussion

TABLE 1 shows the experiment results of different systems on the RT09. All these systems are randomly initialized. The baseline system is VB1, which is well described in paper [3] and partly realized by the python code downloaded from: http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoiceand-hmm-priors.

Compared with VB1, VB2 takes each segment neighbors into consideration. When calculating $p(x_m|y_s)$, equation (10) is firstly used and then followed by (13). The other three systems are the proposed LCM-Ivec-PLDA (LIP) systems with their inner loop in Fig. 2 being the same as VB1. As for the outer loop, LIP1 use only (11) to calculate $p(x_m|y_s)$. Compared with LIP1, LIP2 uses X_m instead of x_m to extract i-vector y_m. And LIP3 uses both X_m and (13) to improve its performance.

From TABLE 1, we can see that VB2 is better than VB1; LIP2 is better than LIP1; LIP3 is better than LIP2 in both DER and SPKR. This is because taking x_m's neighbors into account can improve the robustness and accuracy of $p(x_m, |y_s)$ both in VB and LIP systems.

Moreover, LIP2 and LIP3 is better than VB2. This demonstrates that the introduction of PLDA is very effective. VB is a single-objective method with the objective function of maximizing $\log p(X)$ in (1). Whereas, LIP is a multi-objective method with inner loop being the same as VB1 and outer loop aiming at distinguishing different speakers, which is in line with the basic requirements of speaker diarization task and contributes to its performance improvement. In most audio recordings, the SPKR is significantly small compared with the DER, which means that the VAD is the main source of errors.

We also do the AHC initialization experiment on our collected audio recordings, in which there is a speaker dominating the conversation. It can be seen in TABLE 2 that in this case random initialization method has poor results both in VB2 and LIP3 system. The proposed AHC hard and soft prior can improve the system performance significantly. And the soft prior is more robust than the hard prior, because soft prior gives each segment its prior according to its distance to the estimated speaker centers. With the AHC initialization, the LIP3 system also surpasses the VB2 system.

It should be noted that, the experiment results in Table 1 are

Table 1: Experiment Results of Different Systems on RT09

DER	VB1	VB2
EDI_20071128-1000_ci01_d03	11.2	10.0
EDI_20071128-1500_ci01_d07	20.2	19.6
IDI_20090128-1600_ci01_d08	7.0	6.4
IDI_20090129-1000_ci01_d07	36.0	35.2
NIST_20080201-1405_d05	49.3	47.6
NIST_20080227-1501_d07	23.4	23.8
NIST_20080307-0955_d05	29.8	27.1

SPKR	VB1	VB2
EDI_20071128-1000_ci01_d03	2.9	1.9
EDI_20071128-1500_ci01_d07	5.0	4.3
IDI_20090128-1600_ci01_d08	1.5	0.9
IDI_20090129-1000_ci01_d07	16.0	15.3
NIST_20080201-1405_d05	29.4	27.8
NIST_20080227-1501_d07	12.2	12.6
NIST_20080307-0955_d05	21.5	18.8

DER	LIP1	LIP2	LIP3
EDI_20071128-1000_ci01_d03	10.3	9.9	9.3
EDI_20071128-1500_ci01_d07	38.2	20.3	19.5
IDI_20090128-1600_ci01_d08	7.3	7.2	6.4
IDI_20090129-1000_ci01_d07	44.9	31.9	32.1
NIST_20080201-1405_d05	63.8	44.8	39.7
NIST_20080227-1501_d07	64.4	14.6	14.3
NIST_20080307-0955_d05	51.6	18.2	16.1

SPKR	LIP1	LIP2	LIP3
EDI_20071128-1000_ci01_d03	2.0	1.6	1.1
EDI_20071128-1500_ci01_d07	22.7	4.8	4.2
IDI_20090128-1600_ci01_d08	1.7	1.7	0.9
IDI_20090129-1000_ci01_d07	25.1	11.9	12.1
NIST_20080201-1405_d05	43.9	25.0	20.2
NIST_20080227-1501_d07	53.3	3.4	3.1
NIST_20080307-0955_d05	43.4	9.9	7.9

[1] The VB1 is the system described in the P. Kenny's paper [3]. The VB2 is the VB1 + (13).

[2] The LIP1 only uses (11) in the outer loop. LIP2 is the same as the LIP1, except that it uses X_m instead of x_m to extract i-vector y_m. LIP3 is the LIP2 + (13).

[3] Because we only focus on single-channel speaker diarization task, there is no use of microphone array processing or speech enhancement. Our methods runs on all the single-channel audio recordings in the SDM condition and we select the best result to report here.

Table 2: Experiment Results with Random Initialization and AHC Initialization

	Random Prior		AHC Hard Prior		AHC Soft Prior	
VB2	SPKR	DER	SPKR	DER	SPKR	DER
TL 7	36.9	40.1	1.7	4.9	1.9	5.2
TL 8	24.1	28.7	6.1	10.8	1.3	6.1
TL 9	30.6	32.4	6.6	8.4	1.1	2.9
LIP3	SPKR	DER	SPKR	DER	SPKR	DER
TL 7	38.8	42.0	1.5	4.7	0.3	3.5
TL 8	32.2	36.9	2.3	7.1	0.8	5.5
TL 9	44.7	46.5	6.2	8.0	1.1	2.9

all without AHC initialization. It shows that when the recordings are not dominated by one speaker, the system performances are not so depend on the AHC initialization.

5. Conclusion

In this paper, we present a latent class model and then realize a LCM-Ivec-PLDA system for speaker diarization. The LCM provides an iterative framework for multi-objective optimization and allows more flexible and discriminant $p(x_m|y_s)$ models. The proposed LCM-Ivec-PLDA system which significantly outperforms the VB system on the RT09 and our collected database. There are three main reasons. First, the introduced PLDA in the computation of $p(x_m|y_s)$ enhances the system's ability at distinguishing speakers. Second, the proper usage of x_m's neighbors is important. One way is taking x_m and its neighbors to constitute X_m to calculate y_m. The other way is considering the contribution of neighbors by multiplying a weight. At last, the AHC initialization is crucial in the case when one speaker dominates the whole conversation. Both the hard and soft prior work well and the soft prior is more informative. Overall, compared with the classical VB system, the SPKR and DER of LCM-Ivec-PLDA are significantly reduced on the NIST RT09 SPKD and our collected database respectively.

6. Appendix

Proof of equation (3) and (4).

The objective function is:

$$\max_{q_{ms}} \sum_{m=1}^{M} \log\left(\sum_{s=1}^{S} q_{ms} p(x_m|y_s)\right),$$

$$s.t. \sum_{s=1}^{S} q_{ms} = 1 (1 \le m \le M)$$

The above formula is similar to GMM. Following the EM algorithm, we introduce an auxiliary function Γ_{ms}, $\sum_{s=1}^{S} \Gamma_{ms} = 1$, $\Gamma_{ms} \ge 0$. Then have

$$\begin{aligned} f &= \sum_{m=1}^{M} \log\left(\sum_{s=1}^{S} \Gamma_{ms} \frac{q_{ms} p(x_m|y_s)}{\Gamma_{ms}}\right) \\ &\ge \sum_{m=1}^{M} \sum_{s=1}^{S} \Gamma_{ms} \log \frac{q_{ms} p(x_m|y_s)}{\Gamma_{ms}} \end{aligned}$$

In the E step, we calculate Γ_{ms} which maximizes the log likelihood function given Q and P. The equality holds if and only if

$$\frac{q_{ms} p(x_m|y_s)}{\Gamma_{ms}} = \text{constant}$$

Because $\sum_{s=1}^{S} \Gamma_{ms} = 1$, we can get $\sum_{s=1}^{S} q_{ms} p(x_m|y_s) =$ constant. Then

$$\Gamma_{ms} = \frac{q_{ms} p(x_m|y_s)}{\sum_{s'=1}^{S} q_{ms'} p(x_m|y_{s'})}$$

In the M step, we estimate Q given Γ_{ms}. When estimate Q, we need to maximize the following function

$$\begin{aligned} f &= \sum_{m=1}^{M} \sum_{s=1}^{S} \Gamma_{ms} \log \frac{q_{ms} p(x_m|y_s)}{\Gamma_{ms}} \\ &= \sum_{m=1}^{M} \sum_{s=1}^{S} \Gamma_{ms} \log q_{ms} + \sum_{m=1}^{M} \sum_{s=1}^{S} \Gamma_{ms} \log \frac{p(x_m|y_s)}{\Gamma_{ms}} \end{aligned}$$

$$s.t. \sum_{s=1}^{S} q_{ms} = 1 (1 \le m \le M)$$

The second item is not related to q_{ms}, and can be considered as a constant. We introduce new variable λ_m ($\lambda_m \ne 0$) called a Lagrange multiplier and study the Lagrange function

$$L = \sum_{m=1}^{M} \sum_{s=1}^{S} \Gamma_{ms} \log q_{ms} + \sum_{m=1}^{M} \lambda_m \left(\sum_{s=1}^{S} q_{ms} - 1\right)$$

Take its partial derivative for q_{ms}, and let it be zero.

$$\begin{aligned} &\frac{\partial L}{\partial q_{ms}} \\ =& \frac{\sum_{m=1}^{M} \sum_{s=1}^{S} \Gamma_{ms} \log q_{ms} + \sum_{m=1}^{M} \lambda_m (\sum_{s=1}^{S} q_{ms} - 1)}{q_{ms}} \\ =& \frac{\Gamma_{ms}}{q_{ms}} + \lambda_m = 0 \end{aligned}$$

So $q_{ms} = -\frac{\Gamma_{ms}}{\lambda_m}$.

Because $\sum_{s=1}^{S} q_{ms} = -\sum_{s=1}^{S} \frac{\Gamma_{ms}}{\lambda_m} = 1$, thus $\lambda_m = -\sum_{s=1}^{S} \Gamma_{ms} = -1$. Then we can get

$$q_{ms} = \Gamma_{ms}.$$

7. References

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

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

[3] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, Dec 2010.

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

[5] D. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Proceedings of Interspeech 2009*, sep 2009, pp. 1047–1050.

[6] F. Valente, *Variational Bayesian methods for audio indexing*, Ph.D. thesis, Eurecom, Sophia-Antipolis, France, 2005.

[7] A. E. Bulut, H. Demir, Y. Z. Isik, and H. Erdogan, "Plda-based diarization of telephone conversations," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4809–4813.

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

[9] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomera-tive hierarchical clustering in a speaker diarization system," in *Proceedings of Interspeech 2007*, Aug 2007, pp. 1853–1856.

[10] *The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan*, 2009.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society, series B*, vol. 39, no. 1, pp. 1–38, 1977.

[12] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, San Diego, CA: Academic Press, 7th ed. edition, 2000.

[13] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, “Exploiting intra-conversation variability for speaker diarization,” in *Proceedings of Interspeech 2011*, Aug 2011, pp. 945–948.

[14] W. Zhu and J. Pelecanos, “Online speaker diarization using adapted i-vector transforms,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5045–5049.

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

[16] Carol Espy-Wilson Daniel Garcia-Romero, “Analysis of i-vector length normalization in speaker recognition systems,” in *Proceedings of Interspeech 2011*, Aug 2011, pp. 249–252.

[17] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*, Springer-Verlag, 1991.

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

[19] G. L. Lan, D. Charlet, A. Larcher, and S. Meignier, “Iterative plda adaptation for speaker diarization,” in *Proceedings of Interspeech 2016*, Sep 2016, pp. 2175–2179.