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

# VB-HMM Speaker Diarization with Enhanced and Refined Segment Representation

*Xianhong Chen, Liang He, Can Xu, Yi Liu, Tianyu Liang, Jia Liu*

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
{chenxianhong,heliang,xucan,ty_liang, liuj}@mail.tsinghua.edu.cn

## Abstract

Variational Bayes hidden Markov model (VB-HMM) is a soft speaker diarization system. It is often combined with fixed length segmentation (FLS) instead of speaker change detection (SCD) to avoid SCD error propagation. However, as each segment is too short to provide enough speaker information, the emission probability (given a speaker, a segment occurs) will be noisy and inaccuracy. Therefore, we propose a VB-HMM speaker diarization system with enhanced and refined segment representation. First, it enhances the segment representation with stream neighbors to extract more information of the same speaker to improve the accuracy of emission probability, and then it further refines the segment representation with speaker change points in the iteration to dislodge the information of other different speakers. The experiment results on RT09 demonstrate that, VB-HMM with enhanced and refined segment representation has a relative improvement of 22.9% compared with VB-HMM with only FLS.

## 1. Introduction

The task of speaker diarization aims at determining "who spoke when". It has many useful applications such as automatic audio retrieving, and speech recognition [1, 2].

A classical speaker diarization usually has three parts: voice activity detection (VAD), where non-speech (silence or noise) segments are removed; speaker segmentation, where an audio is split into speaker homogeneous segments; and speaker clustering, where multiple segments belonging to the same speaker are grouped into a cluster.

In the speaker segmentation stage, there are two dominant approaches. The first approach is speaker change detection (SCD) based on a speaker homogeneous evaluation criteria. One of the most popular criteria is Bayesian information criterion (BIC) [3]. It is based on an unsupervised method balancing between likelihood function and free model parameters. One will use two adjacent sliding windows on the audio, compute the distance between them, then decide whether the two windows are generated from the same speaker. Recently, there are some attempts based on supervised method trying to improve the performance of SCD, such as factor analysis [4], deep neural networks (DNN) [5, 6], convolutional neural networks (CNN) [7, 8], and recurrent neural network (RNN) [9, 10]. Paper [7] firstly introduces CNN to SCD. It makes a decision directly on the signal spectrogram. Following this work, the output of CNN is applied in [8] to refine the statistics for a segment i-vector extraction. Paper [10] regards SCD as a sequence labelling task, and addresses it with a bidirectional long short-term memory. All these supervised methods need large amount of labeled data and might suffer from lack of robustness when working in different acoustic environment. Therefore, Jati and Georgiou [11] train a DNN on unlabeled data to learn speaker manifold for SCD. Even with all these effort, SCD performance is still not sufficient, and the resulted speaker change points are fixed, so SCD error will propagate to the subsequent procedures.

The second approach is fixed length segmentation (FLS) which divides the audio into fixed length short segments [12, 13, 14, 15, 16]. As each segment is short enough, it can be seen as a speaker homogeneous one. The critical design consideration of this approach is the choice of the segment length. For very short segment, its speaker information is poor and the estimation is noisy. For long segment, the total length of mixed segments containing more than one speaker increases. The limitation of FLS is the minimal length of the segment from which the identity of the speaker can be extracted. In this case, segments should be long enough to allow the extraction of speaker information while limiting the risk of a speaker change point existing in it. Sholokhov et al. [14] found that the optimal segment length should be in the range of 0.5 - 1 second. Though this approach has no need to find speaker change points, it leaves the tough challenge to speaker clustering. For short segments with little information, many clustering methods are helpless.

In the speaker clustering stage, agglomerative hierarchical clustering (AHC) is the most popular algorithm [17]. It treats the divided segments as individual cluster and merges a pair of nearest clusters into a new one. This merging process is repeated until a stopping criterion, which might be a threshold method or pre-estimated speaker numbers [18, 16, 19], is satisfied. However, the drawback of AHC is that the clustering error may be propagated and the premature hard decision can't be remedied in the later iterations. Besides, the sequential information is not fully taken into account during the clustering stage, thus it may result in frequent speaker changes in a short period and needs a resegmentation [20].

To overcome the problem induced by hard decision, a soft speaker clustering method inspired by variational Bayes (VB) framework is introduced in [21]. VB treats speaker i-vectors and speaker labels as latent variables that are iteratively updated by maximizing a lower bound until convergence. VB speaker clustering method combined with a hidden Markov model (VB-HMM) makes soft decision in its iterations and gains a better performance [22, 23, 12]. VB-HMM is often combined with FLS speaker segmentation method to avoid SCD error propagation. However, as each segment is too short to provide enough speaker information, the emission probability (given a speaker, a segment occurs) will be very noisy and inaccuracy. A more robust and informative emission probability is desired.

In this paper, we propose VB-HMM with enhanced and refined segment representation. The number of speakers is assumed to be known in advance. First, we enhance the segmen-


10.21437/Odyssey.2018-19

t representation with its neighbor segments. It aims to extract more information of the same speaker from the neighbors to improve the emission probability. Then we further make use of the speaker change points in the iteration to refine segment representation. It aims to dislodge the information of other speakers from the neighbors. The experiment results on RT09 demonstrate that VB-HMM with segment representation enhanced and refined has better performance compared with VB-HMM with only FLS.

The remainder of this paper is organized as follows. Section 2 introduces the VB-HMM speaker diarization system, and the segment representation is enhanced with neighbors and refined with speaker change points in Section 3. Section 4 compares the method with the most relevant work. Section 5 describes the experimental setup and discusses the results. Section 6 presents the conclusion.

## 2. VB-HMM Speaker Diarization System

### 2.1. Variational Bayes Speaker Diarizaiton Theory

VB is a soft speaker clustering method introduced in [12, 21]. Suppose the recording is uniformly segmented into fixed length segments $X = (x_1, \cdots, x_m, \cdots, x_M)$, where the subscript $m$ is the time index, $1 \leq m \leq M$. Each of which has its own representation. Let $Y = (y_1, \cdots, y_s, \cdots, y_S)$ be the speaker factors, each of which obeys a normal distribution, and $q_{ms}$ represents the probability that segment $x_m$ is spoken by speaker $s$, $Q = \{q_{ms}\}$. It meets the requirement $\sum_{s=1}^{S} q_{ms} = 1$.

In speaker diarization, $X$ is the observable data; $Y$ and $Q$ are the hidden variables. Our goal is to find proper $Y$ and $Q$ to maximize $P(X)$. According to the Kullback-Leibler divergence, the lower bound of the log likelihood $\ln P(X)$ can be expressed as

$$\ln P(X) \geq \int P(Y,Q) \ln \frac{P(X,Y,Q)}{P(Y,Q)} d(Y,Q) \qquad (1)$$

The equality holds if and only if $P(Y,Q) = P(Y,Q|X)$. VB uses a factorizable $P(Y,Q) = P(Y)P(Q)$ to approximate the true posterior $P(Y,Q|X)$. Then $P(Y)$ and $P(Q)$ are iteratively refined to increase the lower bound of $P(X)$.

### 2.2. VB-HMM System Realization

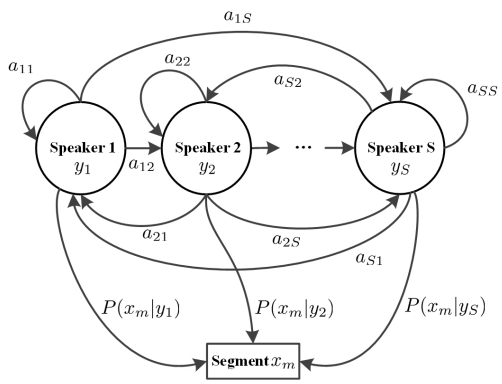



Figure 1: The probabilistic graphical model of the HMM

VB is usually applied with hidden Markov model (VB-HMM). Fig. 1 is the probabilistic graphical model of the HMM adopted in our speaker diarization system, each state represents a speaker. The emission probability $P(x_m|y_s)$ represents that given a speaker $s$, the probability of segment $x_m$ occurrence, and the transition probability $A = \{a_{ij}\}$ represents the probability that a speaker jumps to itself and other speakers. It should be noted that, the HMM in this paper is not frame based. The $x_m$ in emission probability $P(x_m|y_s)$ is not a frame but a segment.

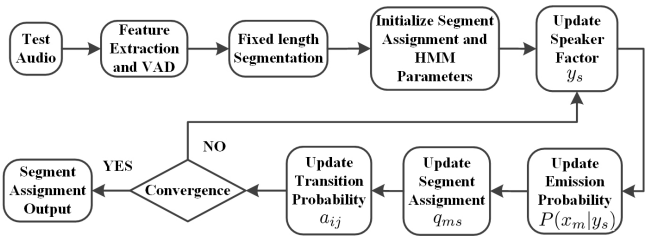



Figure 2: Flow chart of a VB-HMM speaker diarization system

Fig. 2 is the flow chart of a VB-HMM speaker diarization system. After feature extraction and VAD, the recording is uniformly segmented into fixed length segments $X = \{x_1, \cdots, x_m, \cdots, x_M\}$. Each segment is represented by its statistics, and then the speaker factor $Y$ and HMM parameters are iteratively updated until converge.

- Segment Representation

The zero-, first-, and second-order Baum-Welch statistics extracted from segment $x_m$ are as follows,

$$\begin{aligned} N_{mc} &= \sum_t \gamma_{mt}(c) \\ F_{mc} &= \sum_t \gamma_{mt}(c)(x_{mt} - \mu_{ubm,c}) \\ S_{mc} &= \text{diag}\left(\sum_t \gamma_{mt}(c)(x_{mt} - \mu_{ubm,c})(x_{mt} - \mu_{ubm,c})^{\mathrm{T}}\right) \end{aligned} \qquad (2)$$

where $x_{mt}$ is the $t$th frame of $x_m$, $\mu_{ubm,c}(c = 1, \cdots, C)$ is the subvector of UBM mean supervector $\mu_{ubm}$, and $\gamma_{mt}(c)$ is the posterior that $x_{mt}$ is generated by mixture component $c$. Let $N_m$ be the matrix whose diagonal blocks are $N_{mc}I$, $F_m$ be the supervector obtained by concatenating $F_{mc}$, and $S_m$ be the diagonal matrix whose diagonal blocks are $S_{mc}$. Each segment can be represented by its statistics: $N_m$, $F_m$, and $S_m$.

- Update speaker factor $y_s$

The mean vector $\omega_s$ and the precision matrix $\Lambda_s$ of speaker factor $y_s$ are updated as follows

$$\begin{aligned} \Lambda_s &= I + T^{\mathrm{T}}\Sigma^{-1}N(s)T \\ \omega_s &= \Lambda_s^{-1}T^{\mathrm{T}}\Sigma^{-1}F(s) \end{aligned} \qquad (3)$$

where $\Sigma$ is a covariance matrix with diagonal blocks $\{\Sigma_1, \cdots, \Sigma_C\}$. $\Sigma_c$ is the covariance matrix associated with mixture component $c$ of the universal background model (UBM). $T$ is the total variability matrix. $N(s)$ and $F(s)$ are the speaker dependent Baum-Welch statistics, which are obtained by taking the segment assignment probability $q_{ms}$ into consid-

eration:

$$N(s) = \sum_{m=1}^{M} q_{ms} N_m$$
$$F(s) = \sum_{m=1}^{M} q_{ms} F_m \quad (4)$$

It should be noted that, the speaker Baum-Welch statistic $N(s)$ and $F(s)$ are weighted by $q_{ms}$. So it avoids hard decisions and belongs to a soft clustering algorithm.

- Update emission probability $P(x_m|y_s)$

The update of the emission probability is:

$$\ln P(x_m|y_s) = G_m + H_{ms} \quad (5)$$

where

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

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

- Update $q_{ms}$

With the emission probability and other HMM parameters, the forward-backward algorithm is used to update $q_{ms}$.

- Update transition probability $a_{ij}$

For the transition probability $a_{ij}$, as all the segments are short, the probability of a speaker jump to itself is larger than to others. We define:

$$a_{ii} = loop$$
$$a_{ij} = (1 - loop)\frac{sp_j}{\sum_{s=1, s\neq i}^{S} sp_s}; (j \neq i) \quad (6)$$

where $sp_s = \sum_{m=1}^{M} q_{ms}$ represents the total number of segments spoken by speaker $s$. The larger the $sp_s$ is, the more likely that other speakers will jump to speaker $s$.

Once the system converges, the speaker diarization is performed by assigning each segment to the speaker given by $\arg\max_s q_{ms}$.

# 3. Segment Representation Enhancement and Refinement

In VB-HMM, each segment is short enough to ensure its homogeneous. As each segment is too short to provide enough speaker information for the subsequent clustering, we further enhance segment representation with its stream neighbors and refine segment representation with speaker change points obtained in the iteration.

Fig. 3 is the sketch of the segment representation enhancement and refinement method. Each segment is represented by a black dot. In (b), segment representation is enhanced with neighbor segments. The triangle represents the weight added to the neighbors. It aims to extract more information of the same speaker from the neighbor segments. In (c), the segment representation is further refined by speaker change points. The weight added to the neighbors outer a speaker point is set to be zero. It aims to dislodge the information of different speakers. Compared (c) and (d), we know that in different iterations, the speaker change points might be different. So the refinement will be different too.

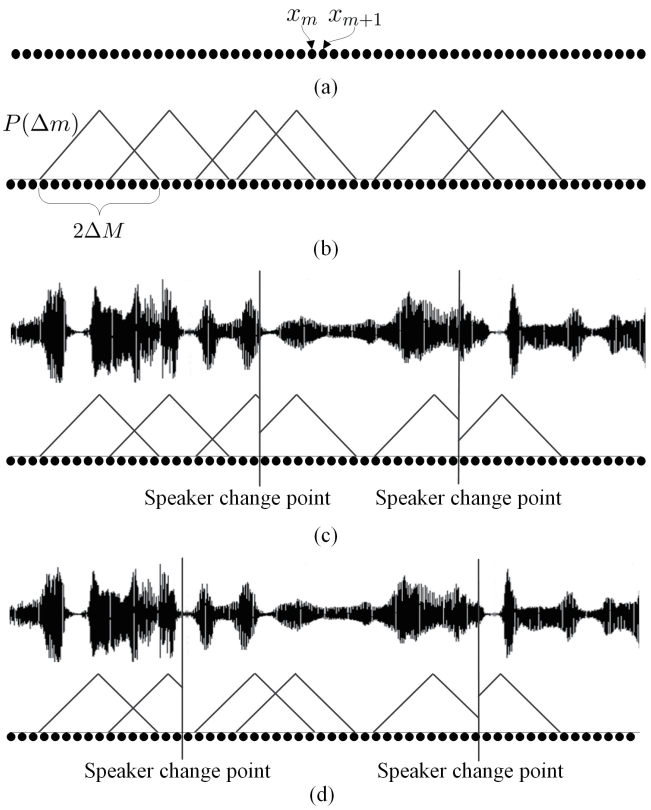



Figure 3: (a) Fixed length segmentation, each segment is represented by a black dot. (b) Representation enhanced with stream neighbors (Enhance). (c) Representation enhanced and refined with neighbors and speaker change points (Enhance-Refine). (d) The speaker change points obtained in different iterations might be different, resulting in different $p(\Delta m)$.

### 3.1. Representation Enhancement

Given an audio containing multiple speakers, it is very unlikely that speaker will change frequently, such as once per second. Therefore, there is a strong possibility that each segment and its time stream neighbors belong to a single speaker, and the more close two segments are from each other, the more larger this probability will be. Based on this perception, we enhance the segment representation with its stream neighbors to extract more speaker information as follows:

$$\hat{N}_m = \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) N_{m+t}$$
$$\hat{F}_m = \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) F_{m+t} \quad (7)$$
$$\hat{S}_m = \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) S_{m+t}$$

where $2\Delta M$ is the number of neighbor segments considered in the enhancement. $P(\Delta m)$ is a weight added to each neighbor segment. It can take many forms, as shown in Fig. 3(b), it is exhibited as a triangle function,

$$P(\Delta m) = \begin{cases} \frac{\Delta M - \Delta m}{\Delta M} & \text{for } \Delta m \leq \Delta M \\ 0 & \text{for } \Delta m > \Delta M. \end{cases} \quad (8)$$

$P(\Delta m)$ can also be model as a Poisson distribution [24]. Suppose $x_{m+\Delta m}$ is the neighbor of $x_m$. The distribution of the number of speaker change points between $x_m$ and $x_{m+\Delta m}$ can be modeled as a Poisson distribution:

$$P(k, \Delta m) = \frac{e^{-\lambda|\Delta m|}(\lambda|\Delta m|)^k}{k!} \quad (9)$$

where $k$ is the number of speaker change points, $\lambda$ is the average number of speaker change points in unit segment. Under this assumption, the probability that audio segments from time $m$ to time $m + \Delta m$ belong to a single speaker is equivalent to the probability that the speaker change point does not appear from time $m$ to time $m + \Delta m$. That means $k = 0$. Therefore, $P(\Delta m)$ can be represented as:

$$P(\Delta m) = e^{-\lambda|\Delta m|} \quad (10)$$

### 3.2. Representation Refinement

Segment representation can be further refined by taking the speaker change points into consideration. In each iteration, there will be a clustering result and a set of speaker change points. If these exist speaker change points between the segment and its neighbor in the last iteration, then in the next iteration, this neighbor would be discarded and the weight $P(\Delta m)$ added to it would be 0, as shown in Fig. 3(c). This idea tries to avoid the segment representation being mixed with other speaker's information. Different from SCD, the speaker change points and $P(\Delta m)$ for each segment will change in different iterations, as shown in Fig. 3(d), thus avoiding the error of speaker change points propagating.

### 3.3. VB-HMM with Enhanced and Refined Representation

In the VB-HMM iterations, each speaker factor is updated according to all the audio segments multiplied with $q_{ms}$. $q_{ms}$ is updated by forward-backward algorithm. Only the emission probability $p(x_m|y_s)$ is updated according to only one segment $x_m$. If segment $x_m$ is too short to provide enough speaker information, this emission probability will be very noisy and inaccuracy. Therefore, emission probability is the key point to be ameliorated in the VB-HMM speaker diarization system. So, the enhanced and refined segment representation are applied in the emission probability update.

$$\ln P(x_m|y_s) = G_m + H_{ms}$$

where

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

$$H_{ms} = \omega_s^{\mathrm{T}} T^{\mathrm{T}} \Sigma^{-1} \hat{F}_m - \frac{1}{2}\omega_s^{\mathrm{T}} T^{\mathrm{T}} \hat{N}_m \Sigma^{-1} T \omega_s$$

Here, we use the enhanced and refined $\hat{N}_m$, $\hat{F}_m$, $\hat{S}_m$ instead of $N_m$, $F_m$, $S_m$.

## 4. Compared to Prior Work

The work most similar to ours is paper [8] proposed by Zajic. It refines the segment representation with a weight, which is produced by a CNN, to found more accurate representation. The difference between it and ours is as follows:

- In reference [8], SCD is used to do speaker segmentation. So each segment, in most cases, is long enough for speaker information extraction. Whereas in this paper, fixed length segmentation is applied. Thus, each segment is too short to provide enough information.
- In [8], the weight is applied to the frames in a single segment. It aims to dislodge the information of other speakers in the same segment. Whereas in this paper, the weight is applied to neighbor segments. It mainly aims to extract more information of the same speaker from the neighbor segments. The speaker change points dislodge information of different speakers from neighbor segments.

## 5. Experiments

### 5.1. Database

Experiments are implemented on the National Institute of Standards and Technology (NIST) Rich Transcription 2009 (RT09) database. This database has 7 meeting audio and each audio has 4-11 speakers and multi microphone channels. But only one channel audio is used to demonstrate our method under the condition of single distant microphone (SDM). Switchboard-P1, RT05, and RT06 database are used as training data. All the above audio are converted to 8kHz 16bits Pulse Code Modulation (PCM) format.

### 5.2. Configuration and Parameters

Perceptual linear predictive (PLP) features with 19 dimensions are extracted from the audio using a 25 ms Hamming window and a 10 ms stride. PLP and log-energy constitute a 20 dimensional basic feature. This base feature along with its first derivatives are concatenated as our acoustic feature vector. VAD is implemented using the frame log-energy and subband spectral entropy.

Audio is divided into 0.3 second segments. The length is not in the range of 0.5 - 1 second provided in [14]. Thanks to the proposed segment representation enhancement and refinement method, we can extract more speaker information for each segment. So the limitation of minimal length of each segment is relaxed.

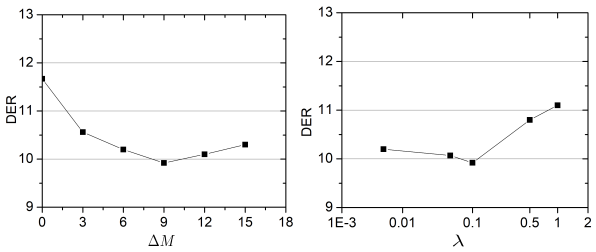
The number of speakers is assumed to be known in advance. For segment representation refinement, we adopt Poisson weight (10) for $P(\Delta m)$. Its $\Delta M$ is set to be 10 and $\lambda$ is 0.1. UBM consists of 512 diagonal Gaussian components, and the rank of total variability matrix $T$ is 300. In the AHC system, PLDA with 150 dimensions is used as the distance metric. In the first iteration, only neighbor information is used, because there is still no speaker change point. Switchboard-P1, RT05, and RT06 database are used to train UBM, $T$, and PLDA.

Diarization error rate (DER) and speaker error rate (SPKR) estimated using md-eval-v21.pl [25] is adopted to measure the system performance according to the RT09 evaluation plan [26].

### 5.3. Result Discussion

Firstly, we study the influence of $\Delta M$ and $\lambda$ in equ. (10) to the VB system performance. Fig. 4(a) shows the DER varied with $\Delta M$ of audio RT09 'edi_20071128-1000_ci01_d03'. It can be seen that when $\Delta M = 0$, that means the proposed speaker segmentation is not used, the performance of the speaker diarization is poor. As $\Delta M$ becomes larger, DER firstly decreases and then increases slightly. This demonstrates that our method

works and extracts more speaker information from the segment and its neighbors. But if $\Delta M$ grows too large, it begins to mix with other speaker's information.




(a) (b)

Figure 4: DER varies with $\Delta M$ and $\lambda$ in equ. (10).

The relationship between DER and $\lambda$ is shown in Fig. 4(b). When $\lambda$ approaches zero, the value of $P(\Delta m)$ approaches to 1, and the Poisson weight degrades to a rectangular weight, resulting in lower performance. As $\lambda$ gets larger, the weight becomes sharper. It firstly excludes the information of different speaker, and then excludes the information of the same speaker. So the DER of the diarization will decrease and then increase.

The VB-HMM system experiment results on RT09 7 audio are list in Table. 1 and Table. 2. It can be seen that, for some audio, the application of our method obviously improves the performance of VB-HMM system, especially when VB-HMM with FLS has a poor performance. For some audio, the improvement is not so obvious.

Table 1: The SPKR of VB-HMM systems on RT09 under SDM condition. 'FLS' means the original VB-HMM without any segment representation improvement. 'Enhance' means VB-HMM with enhanced segment representation. 'E-R' means VB-HMM with enhanced and refined segment representation.

| SPKR[%] | FLS | Enhance | E-R |
|---|---|---|---|
| edi_20071128-1000_ci01_d03 | 3.3 | 1.6 | 1.6 |
| edi_20071128-1500_ci01_d08 | 7.8 | 4.9 | 4.5 |
| idi_20090128-1600_ci01_d07 | 8.1 | 1.7 | 1.7 |
| idi_20090129-1000_ci01_d04 | 16.2 | 13.9 | 12.1 |
| nist_20080201-1405_d05 | 27.6 | 25.2 | 24.1 |
| nist_20080227-1501_d04 | 11.3 | 3.5 | 3.6 |
| nist_20080307-0955_d06 | 19.6 | 10.6 | 9.9 |
| average | 13.41 | 8.77 | 8.21 |

Table 2: The DER of VB-HMM systems on RT09 under SDM condition.

| DER[%] | FLS | Enhance | E-R |
|---|---|---|---|
| edi_20071128-1000_ci01_d03 | 11.67 | 9.91 | 9.89 |
| edi_20071128-1500_ci01_d08 | 25.24 | 23.32 | 19.68 |
| idi_20090128-1600_ci01_d07 | 17.52 | 7.27 | 7.27 |
| idi_20090129-1000_ci01_d04 | 36.21 | 33.35 | 32.37 |
| nist_20080201-1405_d05 | 48.17 | 44.87 | 43.05 |
| nist_20080227-1501_d04 | 22.41 | 14.65 | 14.66 |
| nist_20080307-0955_d06 | 26.14 | 18.24 | 17.41 |
| average | 26.77 | 21.66 | 20.62 |

From the average DER of VB-HMM system on RT09 database, we can see that, VB-HMM without segment representation improvement performs worst. Adding the stream neighbors information (Enhance), the system performance is better. This is because the enhancement extracts more information of the same speaker from the neighbor segments to improve the accuracy of emission probability. The system performs best when the segment representation is enhanced with neighbors and refined with speaker change points (E-R). Because the speaker change points further dislodge the information of different speakers from the neighbor segments. Compared with FLS, the relative improvements of 'Enhance' and 'E-R' are 19.1% and 22.9%, respectively.

We also compare our system performance with other research works in the literatures, which do not know the number of speakers in advance. Table 3 lists the average performance of different methods on the RT09 database. All of these systems except [27] are under a SDM condition. It can be seen that our method has best DER except for [28].

Table 3: Compared with other work performance on RT09. Scoring overlapped speech is accounted in the error rates.

| works | SPKR[%] | DER[%] |
|---|---|---|
| [27] | 14.3 | 27.0 |
| [28] | - | 19.54 |
| [29] | - | 31.3 |
| [30] | - | 21.1 |
| ours | 8.21 | 20.62 |

## 6. Conclusion

In this paper, we propose a VB-HMM speaker diarization system with enhanced and refined segment representation. First, it enhances the segment representation with stream neighbors, and then it refines the segment representation with speaker change points in the iteration. The experiment results on RT09 demonstrate that, VB-HMM with enhanced segment representation has a relative improvement of 19.1% compared with VB-HMM with only FLS. This is because the enhancement extracts more information of the same speaker from neighbors to improve the accuracy of emission probability. VB-HMM with enhanced and refined segment representation has best performance, and the relative improvement reaching to 22.9%. This is because the refinement further dislodges the information of different speakers.

## 7. References


[1] 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.

[2] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[3] Chuck Wooters and Marijn Huijbregts, *The ICSI RT07s Speaker Diarization System*, pp. 509–519, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[4] Brecht Desplanques, Kris Demuynck, and Jean-Pierre Martens, "Factor analysis for speaker segmentation and improved speaker diarization," in *Proceedings of Interspeech 2015*, 2015, pp. 3081–3085.

[5] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using auto associative neural networks," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 4, pp. 667 – 675, 2009.

[6] V. Gupta, "Speaker change point detection using deep neural nets," in *Proceedings of ICASSP*, April 2015, pp. 4420–4424.

[7] M. Hruz and Z. Zajic, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *Proceedings of ICASSP*, March 2017, pp. 4945–4949.

[8] Zbynek Zajic, Marek Hruz, and Ludek Muller, "Speaker diarization using convolutional neural network for statistics accumulation refinement," in *Proceedings of Interspeech 2017*, 2017, pp. 3562–3566.

[9] Herve Bredin, "Tristounet: Triplet loss for speaker turn embedding," in *Proceedings of ICASSP*, March 2017, pp. 5430–5434.

[10] Ruiqing Yin, Herve Bredin, and Claude Barras, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in *Proceedings of Interspeech 2017*, 2017, pp. 3827–3831.

[11] Arindam Jati and Panayiotis Georgiou, "Speaker2vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation," in *Proceedings of Interspeech 2017*, 2017, pp. 3567–3571.

[12] 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.

[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] Alexey Sholokhov, Timur Pekhovsky, Oleg Kudashev, Andrei Shulipa, and Tomi Kinnunen, "Bayesian analysis of similarity matrices for speaker diarization," in *Proceedings of ICASSP*, May 2014, pp. 106–110.

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

[16] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, Jan 2014.

[17] 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.

[18] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[19] Gregory Sell, Alan McCree, and Daniel Garcia-Romero, "Priors for speaker counting and diarization with ahc," in *Proceedings of Interspeech 2016*, Sep 2016, pp. 2194–2198.

[20] Gregory Sell and Daniel Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *Proceedings of ICASSP*, 2015, pp. 4794 – 4798.

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

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

[23] 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.

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

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

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

[27] Sree Harsha Yella and Herve Bourlard, "Information bottleneck based speaker diarization of meetings using non-speech as side information," in *Proceedings of ICASSP*, May 2014, pp. 96–100.

[28] Hanwu Sun, Bin Ma, Swe Zin Kalayar Khine, and Haizhou Li, "Speaker diarization system for rt07 and rt09 meeting room audio," in *Proceedings of ICASSP*, March 2010, pp. 4982–4985.

[29] Gerald Friedland, Adam Janin, David Imseng, Xavier Anguera, Luke Gottlieb, Marijn Huijbregts, Mary Tai Knox, and Oriol Vinyals, "The icsi rt-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 371–381, Feb 2012.

[30] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The liaeurecom rt'09 speaker diarization system: Enhancements in speaker modelling and cluster purification," in *Proceedings of ICASSP*, March 2010, pp. 4958–4961.