The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

DP-Means: An efficient Bayesian nonparametric model for speaker diarization

Yijun Gong, Xiao-Lei Zhang

[1] CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
[2] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China
gongyj@mail.nwpu.edu.cn, xiaolei.zhang@nwpu.edu.cn

Abstract

Recently, Bayesian probabilistic model based clustering gets superior performance in speaker diarization, however, it is much more complicated than widely used efficient clustering algorithms, which is not convenient for some real-life scenarios. In this paper, we propose a covariance-asymptotic variant to Dirichlet process mixture models (DPMM), named Dirichlet process means (DP-means) clustering for speaker diarization. Similar to Bayesian nonparametric models (e.g. DPMM), DP-means can constantly generate new clusters during clustering, which is suitable to the speaker diarization problem where the number of speakers is determined on-the-fly. Different from Bayesian nonparametric models, DP-means is a hard clustering that does not need to optimize the variance of mixtures, which is efficient for real-world problems. We further exploited an initialization method to obtain the prior cluster centroids for DP-means. Experimental results on the CALLHOME, AMI and DIHARD III corpora show that the proposed method is more efficient than the state-of-the-art speaker clustering methods with slight performance degradation.

1. Introduction

Speaker diarization is a task of labeling the identities of speakers in conversations with time stamps. It aims to solve the problem of "who spoke when" [1, 2]. It is a key front-end of multi-speaker speech recognition, and finds its applications in many real-life scenarios such as meetings. Generally, it has two research directions. One is the stage-wise approach, and the other is the end-to-end approach [3, 4, 5, 6]. Stage-wise speaker diarization usually consists of four steps. Given a speech recording, it first removes silence segments from the raw recording by voice activity detection (VAD). Then, it partitions the speech recording into segments that are short enough to ensure that only a single speaker exists in most segments. Next, the segments are fed into a feature extractor to obtain speaker embeddings, such as the i-vectors [7, 8], d-vectors [9], or x-vectors [10, 11]. Finally, the sequential embeddings are fed into a clustering algorithm to obtain the final diarization results. An optional resegmentation process is sometimes applied to refine the result. On the other side, end-to-end diarization obtains the diarization result by a single neural network.

This paper focuses on the clustering algorithm of the stage-wise diarization. Many traditional clustering methods have been widely utilized, such as AHC and spectral clustering. AHC is a bottom-up clustering method [7, 12]. First, each segment is assigned into a single cluster. Then, the two closest clusters are merged into a new cluster repeatedly. Spectral clustering (SC) [13] first calculates the pairwise similarity between segments, which generates an affinity matrix. Then, the affinity matrix is decomposed into low dimensional features by Laplacian eigenvalue decomposition for the speaker clustering.

Recently, many advanced clustering and feature extracting algorithms have been applied as well. In [14], Li *et al.* proposed compositional embeddings to represent two speakers or more in a single embedding. In [15], a novel deep model is applied to reduce the noise and small variance of speaker embeddings. In [16, 17, 18], they used neural networks to learn deep similarity matrices between speaker embeddings. When applying the new representation or similarity matrices to clustering, the diarization performance is boosted. Besides, in [19], the authors proposed a Bayesian hidden Markov model (HMM) based clustering method called VBx. It assumes that the input sequence of embeddings is generated by a speaker-specific state distributions, and uses an ergodic HMM with one-to-one correspondence between the HMM states and speakers to extract a context-dependent representation of speaker embeddings. To our knowledge, VBx reaches the state-of-the-art performance. A main problem of the above algorithms is that their computational complexities are high.

To reduce the time complexity of speaker clustering with guaranteed performance, in this paper, we propose to apply a simplified Dirichlet process mixture models (DPMM), named Dirichlet process means (DP-means) [20], to speaker clustering. DPMM is a Bayesian nonparametric model, which determines the number of mixtures on-the-fly. It is suitable to speaker clustering, however, its main computation is on the variance estimation of the mixtures. To reduce its time complexity, we assume that the variance of each cluster approaches to zero. With this variance asymptotic assumption, we obtain a hard clustering algorithm called DP-means [19]. Because DP-means is sensitive to the initial centroids, we exploit an initial clustering method to provide the DP-means robust initial centroids. Experimental results on CALLHOME [21], AMI [22] and DIHARD III [23] demonstrate that the DP-means yields lower diarization error rate (DER) than AHC and spectral clustering baselines, and is more efficient than the state-of-the-art VBx system with slight performance drop.

The rest of the paper is organized as follows. Section 2 introduces DPMM and our proposed method. Experimental results are presented in Section 3. Finally, Section 4 draws a conclusion.

2. Proposed Method

The architecture of the proposed method is shown in Fig. 1. After the feature extraction by e.g. x-vector, we conduct an initial clustering. Then, we add a clusters filtering module between the initial clustering and DP-means, which filters out very small

10.21437/Odyssey.2022-22

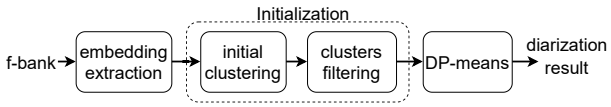

Figure 1: *Architecture of the DP-means based speaker diarization system.*

clusters. Finally, we use the centroids of the obtained clusters to initialize DP-means for the final clustering.

In this section, we first introduce DPMM, then present how to simplify DPMM to DP-means, and finally present the initialization method for DP-means.

2.1. Dirichlet process mixture models

Before DPMM, we first introduce Gaussian mixture models (GMM). We suppose that $\boldsymbol{x}$ is a x-vector, which arises from the distribution:

$$p(\boldsymbol{x}) = \sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \Sigma_c) \tag{1}$$

where $\boldsymbol{\pi}_c$, $\boldsymbol{\mu}_c$ and Σ_c are the mixing coefficient, mean and covariance corresponding to the cth component, and k is the number of the components of GMM.

Then, we place a Dirichlet prior on the mixing coefficients and assume that the covariances in (1) are fixed to σI, where σ is a constant and I is an identity matrix. Moreover, we assume that the means are drawn from the prior distribution. Then, we can build a Bayesian model as follows:

$$\boldsymbol{\mu}_j \sim G_0,\ \forall j = 1, ..., k \tag{2}$$
$$\boldsymbol{\pi} \sim \mathrm{Dir}(k, \boldsymbol{\pi}_0) \tag{3}$$
$$z_i \sim \mathrm{Discrete}(\boldsymbol{\pi}),\ \forall i = 1, ..., n \tag{4}$$
$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \sigma I),\ \forall i = 1, ..., n \tag{5}$$

where G_0 is a prior distribution of the means of speaker clusters, $\boldsymbol{\pi} = \{\pi_1, ..., \pi_k\}$ is the mixing coefficients, n denotes the number of embeddings of the voice segments, z_i indicates that $\boldsymbol{x}_i$ belongs to the z_ith component, $\mathrm{Dir}(k, \boldsymbol{\pi}_0)$ denotes a Dirichlet distribution, and $\mathrm{Discrete}(\boldsymbol{\pi})$ is a discrete distribution. In (5), all $\boldsymbol{\mu}_{z_i}$ are draw from $\boldsymbol{\mu}_j$.

Next, we assume that k tends to ∞, and $\boldsymbol{\pi}_0 = (\alpha/k)\boldsymbol{e}$ where $\boldsymbol{e}$ is an all-one vector. Based on Gibbs sampling utilized in [24], we conduct the inference for [20] which results in the final DPMM as follows[19]:

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{6}$$
$$\phi_i \sim G,\ \forall i = 1, ..., n \tag{7}$$
$$\boldsymbol{x}_i \sim \mathcal{N}(\phi_i, \sigma I),\ \forall i = 1, ..., n \tag{8}$$

where $\mathcal{N}(\phi_i, \sigma I)$ is a Gaussian distribution, $\mathrm{DP}(\alpha, G_0)$ is a Dirichlet process, whose base measure and parameter are G_0 and α respectively, and G is a draw from the Dirichlet process. We can think of a draw from G as choosing one of the infinite means $\boldsymbol{\mu}_c$ drawn from G_0, with the property that the means are chosen with probability equal to the corresponding mixing weights. As a result, each ϕ_i is equal to $\boldsymbol{\mu}_c$ for some c.

2.2. DP-means

We assume that the prior distribution G_0 of the means of DPMM is a zero-mean Gaussian distribution with ρI as the covariance, where ρ is a constant. Then, a parameter λ is applied for expressing α as $(1+\rho/\sigma)^{1/2} \cdot \exp(-\frac{\lambda}{2\sigma})$. Now we can derive the probability of the assignment of an x-vector to a speaker as follows:

$$\hat{\gamma}(z_{ic}) = \frac{n_{-i,c} \cdot \exp(-\frac{1}{2\sigma}\|\boldsymbol{x}_i - \boldsymbol{\mu}_c\|^2)}{\exp(-\frac{\lambda}{2\sigma} - \frac{\|\boldsymbol{x}_i\|^2}{2(\rho+\sigma)}) + \sum_{j=1}^{k} n_{-i,j} \cdot \exp(-\frac{1}{2\sigma}\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2)} \tag{9}$$

$$\hat{\gamma}(z_{i,new}) = \frac{\exp(-\frac{\lambda}{2\sigma} - \frac{\|\boldsymbol{x}_i\|^2}{2(\rho+\sigma)})}{\exp(-\frac{\lambda}{2\sigma} - \frac{\|\boldsymbol{x}_i\|^2}{2(\rho+\sigma)}) + \sum_{j=1}^{k} n_{-i,j} \cdot \exp(-\frac{1}{2\sigma}\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2)} \tag{10}$$

where $\hat{\gamma}(z_{ic})$ and $\hat{\gamma}(z_{i,new})$ are the posterior probabilities of assigning the x-vector to the cth speaker and a new speaker respectively. $n_{-i,c}$ and $n_{-i,j}$ denote the number of the x-vectors classified previously into the c and j-th components respectively.

We see obviously that the above equations are computationally heavy. In order to simplify this model, the core idea of DP-means is to conduct the variance asymptotic approximation, i.e. $\sigma \to 0$, to DPMM. Specifically, an asymptotic approximation is used to the numerator of $\hat{\gamma}(z_{i,new})$ which reformulates the numerator of (10) to:

$$\exp(-\frac{1}{2\sigma}[\lambda + \frac{\sigma}{\rho+\sigma}\|\boldsymbol{x}_i\|^2]). \tag{11}$$

Next, let σ tend to 0, we see that λ dominates (11). As a result, $\hat{\gamma}(z_{ic})$ and $\hat{\gamma}(z_{i,new})$ are only related to the smallest value of $\{\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|^2, ..., \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|^2, \lambda\}$. In this variance asymptotic approximation, only the smallest $\hat{\gamma}$ among $\{\hat{\gamma}(z_{i1}), ..., \hat{\gamma}(z_{ik}), \hat{\gamma}(z_{i,new}))\}$ gets the binary non-zero value. Then, we turn to the posterior means and covariances of the model:

$$\tilde{\boldsymbol{\mu}}_c = (1 + \frac{\sigma}{\rho n_c})^{-1} \bar{\boldsymbol{x}}_c \tag{12}$$
$$\tilde{\Sigma}_c = \frac{\sigma\rho}{\sigma + \rho n_c} I \tag{13}$$

where $\bar{\boldsymbol{x}}_c$ and n_c are the mean and the number of the x-vectors assigned to the cth speaker. When $\sigma \to 0$, $\tilde{\boldsymbol{\mu}}_c$ and $\tilde{\Sigma}_c$ tend to $\bar{\boldsymbol{x}}_c$ and 0 respectively. To this end, we obtain the DP-means algorithm, which is a k-means like algorithm with the parameter k determined on-the-fly. Fig. 2 demonstrates the difference between DPMM and DP-means. We see that DP-means is much simpler than DPMM.

DP-means is optimized by the expectation maximization algorithm. First, it initializes a point as the centroid of the initial cluster. Then, in the E-step, we assign each point to the nearest cluster by calculating the cosine similarity between the point and the centroid of each cluster. If the smallest distance is larger than λ, we create a new cluster. In the M-step, we update the means of each cluster according to the assignment in the E-step. Here, we have to note that when cosine similarity is used to assign x-vectors into clusters, the larger the similarity value is, the closer the two nearest neighbors are. We repeat this algorithm

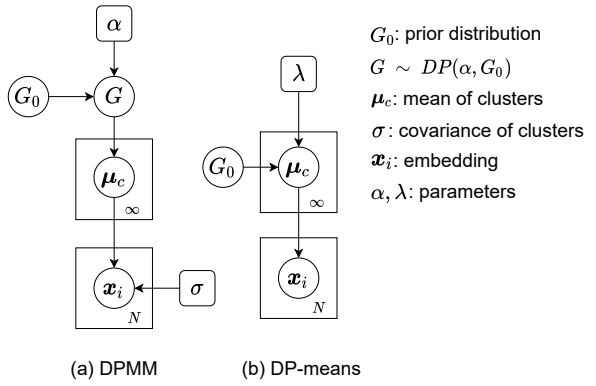

Figure 2: *Comparison between DPMM and DP-means.*

Algorithm 1 DP-means with the ICCF initialization.

Input: $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$: input data, λ: DP-means parameter, p: clusters filtering threshold

Output: Clustering result $\ell_1, ..., \ell_n$

1: Conduct initial clustering: $Y = y_1, ..., y_m$.
2: Remove the clusters with less than p x-vectors: $Y_{init} = y_{(init,1)}, ..., y_{(init,l)}$.
3: Calculate the means of Y_{init}: $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_l$.
4: Initialize $k = l$, $z_i = 1$, $\forall i = 1, ..., n$.
5: Repeat until objective function (14) converges
- For each x-vector $\boldsymbol{x}_i$.
  - Compute $\text{sim}_{ic} = \cos(\boldsymbol{x}_i, \boldsymbol{\mu}_c)$, for $c = 1, ..., k$.
  - If $\max_c \text{sim}_{ic} < \lambda$, set $k = k+1$, $z_i = k$, and $\boldsymbol{\mu}_k = x_i$.
  - Otherwise, set $z_i = \text{argmax}_c \text{sim}_{ic}$.
- Generate clusters $\ell_1, ..., \ell_n$, where $\ell_j = \{\boldsymbol{x}_i | z_i = j\}$.
- For each cluster ℓ_j, compute $\boldsymbol{\mu}_c = \frac{1}{|\ell_c|} \sum_{\boldsymbol{x} \in \ell_c} \boldsymbol{x}$.

until the objective function (14) converges:

$$\sum_{c=1}^{k} \sum_{\boldsymbol{x} \in \ell_c} \|\boldsymbol{x} - \boldsymbol{\mu}_c\|^2. \tag{14}$$

2.3. Initial clustering with cluster filtering for DP-means

DP-means suffers from bad local minimum easily. To overcome this problem, here we propose an initial clustering with cluster filtering (ICCF) method to initialize the cluster centroids of DP-means.

ICCF first generates initial cluster centroids by conventional clustering methods, which behaves like a reliable prior for DP-means. Candidate initial clustering algorithms include AHC and spectral clustering (SC). However, when the number of the initial clusters are too fragile, the final number of speaker clusters may be uncontrolled to be meaninglessly redundant, given that DP-means may generates infinite number of new clusters at the extreme case. To address this problem, a cluster filtering is utilized after the initial clustering, which simply discards the initial clusters that have few number of speaker embeddings. The DP-means algorithm with the ICCF initialization strategy is summarized in Algorithm 1.

3. Experiments

3.1. Experimental setup

We used CALLHOME [21], AMI [22] and DIHARD III [23] as our evaluation datasets. CALLHOME consists of single channel telephone recordings, each of which contains 2 to 7 speakers. The corpus are recorded in Arabic, English, German, Japanese, Mandarin and Spanish. It consists of 500 recordings. The average time of the recordings is about two minutes. Because of the formatting errors of references in a recording, we used 499 recordings in our experiments.

AMI corpus is about 100 hours long. It consists of 171 meeting recordings, each of which contains 4 to 5 speakers and lasts about thirty minutes. We merged the development and evaluation sets as our test set, which occupies about 10% data of the full corpus. Furthermore, AMI was recorded using both headset and far-field microphones array. In our experiments, we tested the recordings from both headset and a random channel from far-field microphones array.

DIHARD III is the third challenge in a series of speaker diarization challenges focusing on "hard" diarization[23]. The data sets consist of 5-10 minute duration samples drawn from 11 domains such as audio books, broadcast interview and clinical conversations. The development set of DIAHRD III was used for evaluation.

We followed the experimental setting in [19] to extract the speaker embeddings. Specifically, we used the oracle VAD to remove silence segments, and extracted 64 log Mel filter bank acoustic features with a frame length of 25ms and frame shift of 10ms. Then, the acoustic features are fed into a ResNet101 [25] backbone neural network to extract 256-dimensional x-vectors. The backbone network contains a 2D convolutional layer, standard ResNet blocks, a statistical pooling layer and a linear transformation. We further used linear discriminant analysis to reduce the dimension of the x-vector to 128.

For the proposed method, AHC and SC are used as the initial clustering tools for DP-means. The similarity measurement between the x-vectors for all clustering algorithms are the cosine similarity. When AHC is used as the initial clustering of our method, its parameter is the same as that for the AHC baseline. The hyperparameter λ was set to 0.275 for CALLHOME, 0.15 for AMI headset, 0.05 for AMI far field, and 0.09 for DIHARD III respectively. The threshold of the cluster filtering was set to 16, 190 and 70 for CALLHOME , AMI corpus and DIHARD III respectively. The hyperparameters were tuned on the validation sets of the corpora, just like VBx does.

We compared with AHC, SC, and VBx [19]. DER is used as the evaluation metric. Similar to previous works on CALLHOME and AMI, a 0.25 second collar was considered for the DER estimation, and no overlap was evaluated. For DIHARD III, in order to follow the evaluation protocol of DIHARD [23], overlap was under consideration, and no forgiveness collar was applied. The time efficiency was evaluated on a Intel(R) Xeon(R) Platinum 8160 CPU server.

3.2. Main result

Table 1 lists the comparison results on CALLHOME, AMI and DIHARD III. From the table, we see that, our method yields

Table 1: *DER (%) and computational time (in seconds) comparison on the CALLHOME, AMI and DIHARD III corpus.*

Data	Method	DER	Time
CALL HOME	AHC	8.46	450
	SC	14.26	186
	VBx	4.42	6041
	ICCF (AHC)+DP-means	5.79	1625
	ICCF (SC)+DP-means	10.76	989
AMI Headset	AHC	5.73	11469
	SC	6.73	2620
	VBx	1.93	10998
	ICCF (AHC)+DP-means	4.17	7961
	ICCF (SC)+DP-means	5.48	8261
AMI Far field	AHC	12.39	11271
	SC	11.15	2781
	VBx	7.97	11650
	ICCF(AHC)+DP-means	10.50	8295
	ICCF (SC)+DP-means	9.35	6886
DIHARD III DEV	AHC	21.70	2027
	SC	24.55	697
	VBx	16.88	11465
	ICCF(AHC)+DP-means	18.85	1383
	ICCF (SC)+DP-means	22.69	3218

Table 2: *DER (%) of DP-means with different initialization methods on CALLHOME.*

Global mean	RS30	RS50	IC (AHC)	ICCF (AHC)
23.41	10.70	10.42	7.13	5.79

lower DER than AHC and SC. Although the proposed methods behaves not as good as VBx in terms of DER, they are much more efficient than VBx, with a 73%, 28%, 41% and 88% time relative reduction over the VBx system on CALLHOME, AMI headset, AMI far field and DIHARD III development corpora respectively. Besides, we sampled a recording to obtain visualized results of different clustering systems in Fig. 3.

3.3. Effect of the initialization of DP-means

To study the effect of different initialization methods of DP-means on performance, we compare ICCF with the following three candidate initialization methods. The first one, named *global means*, initialize DP-means with a single cluster centroid which is the mean of all speaker embeddings. The second one, named *random selection* (RS), selects N embeddings from the embedding sequence as the initial centroids for DP-means. In this experiment, we set N to 30 and 50 respectively, which results in two initializations, denoted as RS30 and RS50 respectively. To pick the best initial centroids, we ran DP-means multiple times, and picks the initial centroid set that results in the minimum objective value. The third one, named *initial clustering without cluster filtering* (IC), feeds all centroids from the initial clustering into DP-means without the cluster filtering.

The evaluation results on CALLHOME are shown in Table 2.

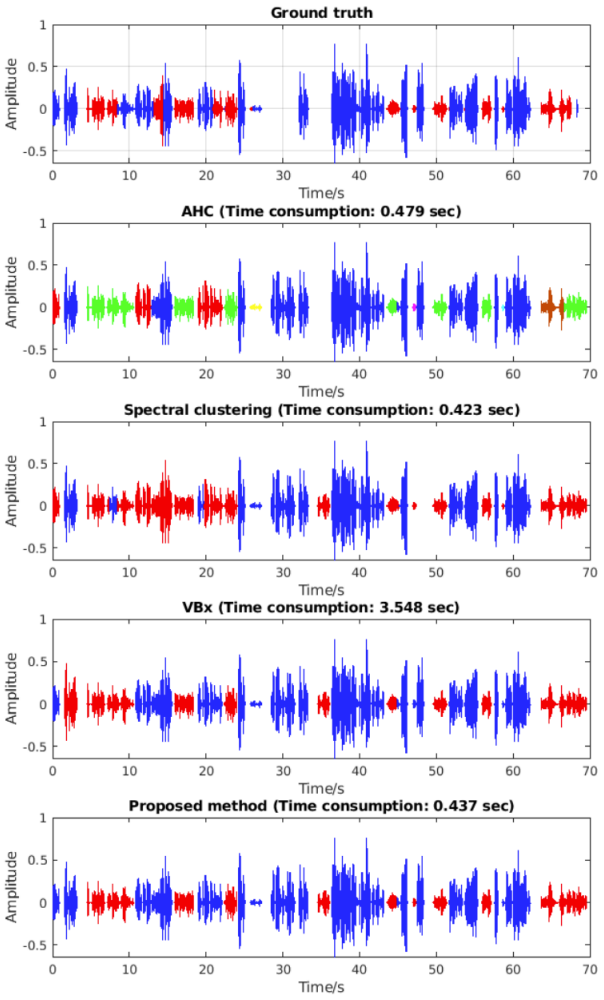

Figure 3: *A comparison example of the diarization result produced from different speaker clustering methods. Different colors represent different speakers*

It can be seen that DP-means without ICCF will easily trap into a local optima. A random initialization reduce the DER of DP-means over the global mean initialization. Moreover, a reliable prior from IC could further improve the performance of DP-means. However, it is still much less effective than the proposed ICCF.

3.4. Effect of hyperparameters of DP-means

The proposed method has two tunable parameters. One is the hyperparameter λ in DP-means; the other is the threshold of clustering filtering p. We tune one of the hyperparameters leaving the other one fixed.

Specifically, for CALLHOME, we set p to 0 and choose λ from 0.2 to 0.32; then, we set λ to 0.275 and choose p ranging from 5 to 17. Similarly, for AMI, we set p to 130, and choose λ from a range of [0.05, 0.18]; then, we select p from [130, 200] with λ set to 0.05. For DIHARD, we set p to 10 and choose λ from 0.08 to 0.26; then, we set λ to 0.09 and choose p ranging from 10 to 80. The result is shown in Fig. 4.

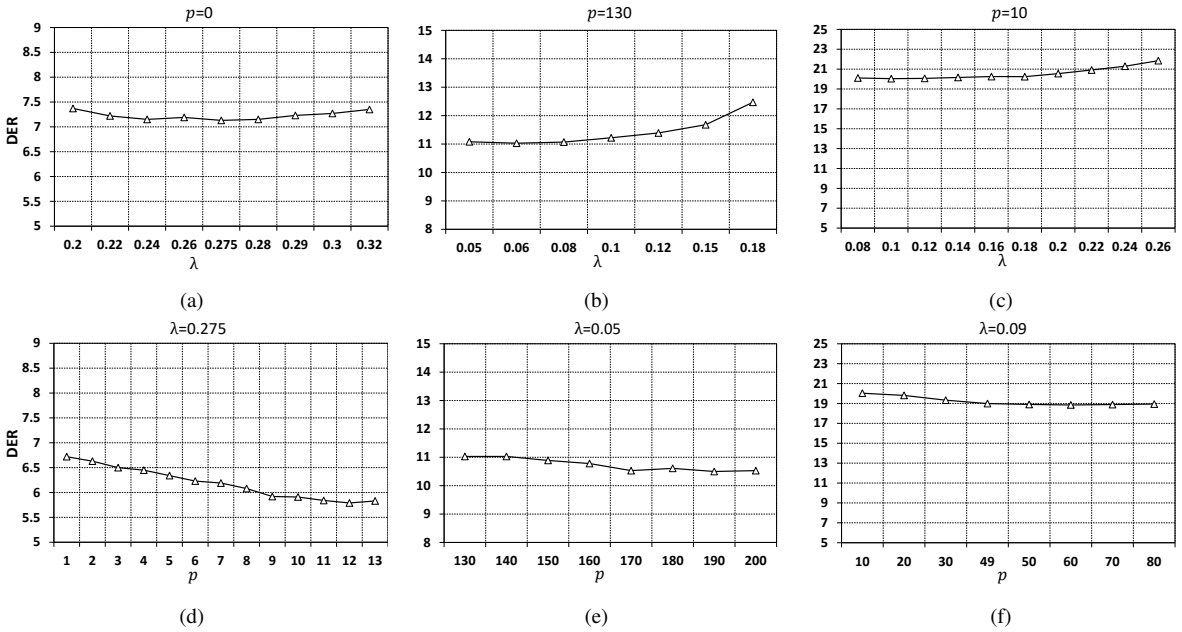

Figure 4: *Effect of the hyperparameters of DP-means. (a) and (d) are the results on CALLHOME. (b) and (e) are the results on the AMI far field corpus. (c) and (f) are the results on DIAHRD III.*

From Figs. 4(a), 4(b) and 4(c), we can see that DER varies from 7.15 to 7.31 on CALLHOME, from 11.08 to 12.47 on AMI, and from 20.11 to 21.85 on DIAHRD III respectively with respect to λ. Figs. 4(d), 4(e) and 4(f) show that DER ranges from 5.79 to 6.72 on CALLHOME, from 10.5 to 11.03 on AMI, and from 18.94 to 20.03 on DIAHRD III respectively with respect to p. The results demonstrate that the proposed method is insensitive to the hyperparameters. Thus, just with a coarse tuning we could also obtain a good performance.

4. Conclusions

In this paper, we proposed a hard clustering algorithm named DP-means for speaker diarization, which could generate new clusters during clustering. DP-means is a simplified Dirichlet process mixture models whose variances asymptotically approach to zero. Because DP-means is relatively sensitive to the initialization, we exploited the ICCF initialization method to provide DP-means robust initial centroids. Comparing with the AHC baseline using AHC initialization, our method achieved 31.6% , 27.2% and 13.1% relative DER reduction on the CALLHOME, AMI and DIHARD III corpora respectively. Moreover, our method could improve the performance and efficiency simultaneously over different initial clustering methods.

5. References

[1] Xavier Anguera, 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.

[2] Sue E Tranter and Douglas A Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 296–303.

[4] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[5] Neil Zeghidour, Olivier Teboul, and David Grangier, "Dive: End-to-end speech diarization via iterative speaker embedding," *arXiv preprint arXiv:2105.13802*, 2021.

[6] Eunjung Han, Chul Lee, and Andreas Stolcke, "Bw-eda-eend: Streaming end-to-end neural speaker diarization for a variable number of speakers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7193–7197.

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

[8] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[9] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[10] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings

for text-independent speaker verification.," in *Interspeech*, 2017, pp. 999–1003.

[11] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[12] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[13] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[14] Zeqian Li and Jacob Whitehill, "Compositional embedding models for speaker identification and diarization with simultaneous speech from 2+ speakers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7163–7167.

[15] Meng-Zhen Li and Xiao-Lei Zhang, "Learning deep representations by multilayer bootstrap networks for speaker diarization," *arXiv preprint arXiv:1910.10969*, 2019.

[16] Tae Jin Park et al., "Multi-scale speaker diarization with neural affinity score fusion," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7173–7177.

[17] Hagai Aronowitz, Weizhong Zhu, Masayuki Suzuki, Gakuto Kurata, and Ron Hoory, "New advances in speaker diarization.," in *INTERSPEECH*, 2020, pp. 279–283.

[18] Qingjian Lin, Yu Hou, and Ming Li, "Self-attentive similarity measurement strategies in speaker diarization.," in *INTERSPEECH*, 2020, pp. 284–288.

[19] Federico Landini, Jn Profant, Mireia Diez, and Luk Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, pp. 101254, 2022.

[20] Brian Kulis and Michael I Jordan, "Revisiting k-means: New algorithms via bayesian nonparametrics," *arXiv preprint arXiv:1111.0352*, 2011.

[21] Alvin F Martin, Mark A Przybocki, et al., "Speaker recognition in a multi-speaker environment.," in *INTERSPEECH*, 2001, pp. 787–790.

[22] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[23] Neville Ryant, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[24] Mike West and Michael D Escobar, *Hierarchical priors and mixture models, with application in regression and density estimation*, Institute of Statistics and Decision Sciences, Duke University, 1993.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.