

INTEGRATING END-TO-END NEURAL AND CLUSTERING-BASED DIARIZATION: GETTING THE BEST OF BOTH WORLDS

Keisuke Kinoshita, Marc Delcroix, Naohiro Tawara

NTT Corporation, Japan

ABSTRACT

Recent diarization technologies can be categorized into two approaches, i.e., clustering and end-to-end neural approaches, which have different pros and cons. The clustering-based approaches assign speaker labels to speech regions by clustering speaker embeddings such as x-vectors. While it can be seen as a current state-of-the-art approach that works for various challenging data with reasonable robustness and accuracy, it has a critical disadvantage that it cannot handle overlapped speech that is inevitable in natural conversational data. In contrast, the end-to-end neural diarization (EEND), which directly predicts diarization labels using a neural network, was devised to handle the overlapped speech. While the EEND, which can easily incorporate emerging deep-learning technologies, has started outperforming the x-vector clustering approach in some realistic database, it is difficult to make it work for *long* recordings (e.g., recordings longer than 10 minutes) because of, e.g., its huge memory consumption. Block-wise independent processing is also difficult because it poses an inter-block label permutation problem, i.e., an ambiguity of the speaker label assignments between blocks. In this paper, we propose a simple but effective hybrid diarization framework that works with overlapped speech and for long recordings containing an arbitrary number of speakers. It modifies the conventional EEND framework to output global speaker embeddings so that speaker clustering can be performed across blocks to solve the permutation problem. With experiments based on simulated noisy reverberant 2-speaker meeting-like data, we show that the proposed framework works significantly better than the original EEND especially when the input data is long.

Index Terms— Speaker diarization, neural networks,

1. INTRODUCTION

Automatic meeting/conversation analysis is one of the essential technologies required for realizing futuristic speech applications such as communication agents that can follow, respond to, and facilitate our conversation. As an important central task for the meeting analysis, speaker diarization has been extensively studied [1–3].

Current state-of-the-art diarization systems that achieve reliable performance in many challenges [1, 2] is based on clustering of speaker embeddings (i.e., speaker identity features) such as i-vectors [4] and x-vectors [5]. Such clustering-based approaches first segment a recording into short homogeneous blocks and compute speaker embeddings for each block assuming that only one speaker is active in each block. Then, speaker embedding vectors are clustered to regroup segments belonging to the same speakers and obtain the diarization results. Various speaker embeddings and clustering techniques have been explored in [6–9]. While these methods can cope with very challenging scenarios [6, 7] and work with an arbitrary number of speakers, there is a clear disadvantage

that they cannot handle overlapped speech, i.e., time segments where more than one person is speaking, because of the way of extracting speaker embeddings. Perhaps surprisingly, even in professional meetings, the percentage of overlapped speech is in the order of 5 to 10%, while in informal get-togethers it can easily exceed 20% [10].

End-to-End Neural Diarization (EEND) has been recently developed [11–13] to address the overlapped speech problem. Similarly to the neural source separation algorithms [14, 15], in EEND, a Neural Network (NN) receives standard frame-level spectral features and directly outputs a frame-level speaker activity for each speaker, no matter whether the input signal contains overlapped speech or not. While the system is simple and has started outperforming the conventional clustering-based algorithms [12, 13], it is difficult to directly apply the EEND systems to *long* recordings (e.g., recordings longer than 10 minutes). The system is designed to operate in a batch processing mode and thus requires a very large computer memory when performing inference with long recordings. Besides, aside from the memory issue, the NNs in EEND has difficulty to generalize to unseen very long sequential data, which also hampers its application to the long recordings. Note that, if we segment the long recordings into small chunks and apply the original EEND model to each chunk independently, the model inevitably suffers from the inter-block label permutation problem, i.e., an ambiguity of the speaker label assignments between chunks. To address this problem (and simultaneously seek for a low-latency solution), [16] proposed an NN-based extension of the EEND to block-online processing. The method in [16] first tries to find single speaker regions, and use them as a guide to assign the speaker labels to the diarization results of future blocks. However, their performance typically does not reach that of the original EEND. Also, more importantly, the method cannot handle an arbitrary number of speakers.

In this paper, we propose a simple but effective hybrid diarization approach, called EEND-vector clustering, by combining the best of the clustering-based diarization and the EEND. A central component of the proposed approach is a modified EEND network that outputs, in each chunk, not only the diarization results but also global speaker embeddings associated with the diarization results. The inter-block permutation ambiguity problem can thus be simply solved by clustering the block-level speaker embedding vectors. This extension thus naturally allows us to combine the advantages of both clustering and the EEND based methods, i.e. it can work with overlapped speech and deal with long recordings including an arbitrary number of speakers. In particular, we confirm experimentally that the proposed EEND-vector clustering significantly outperforms the original EEND system especially when the recordings are long, e.g., more than 5 minutes, while maintaining the same performance as the original EEND system when the recording is short.

The remainder of this paper is organized as follows. We first introduce the proposed framework in section 2 in detail. Then, in section 3, we evaluate its performance in comparison with the original

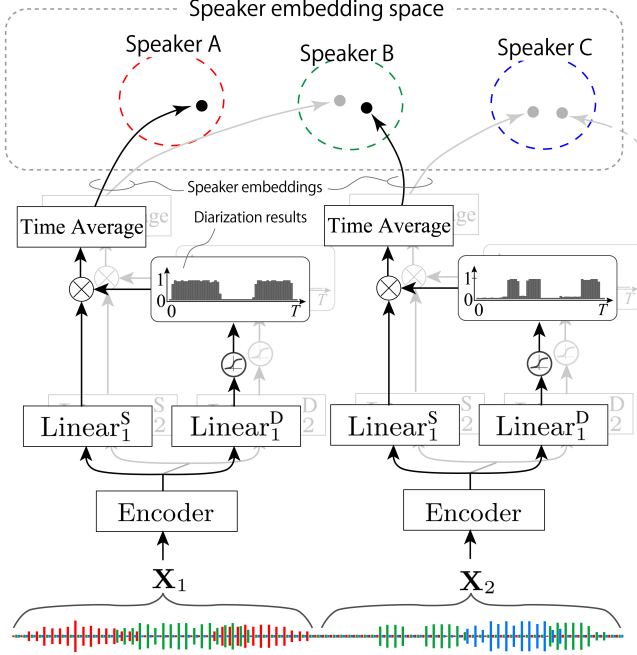


Fig. 1: Schematic diagram of the proposed diarization framework. The input contains 3 speakers in total (red, green, and blue speakers shown in the waveform in the bottom), but only at most 2 speakers are actively speaking in each chunk.

EEND to clarify the advantages of the proposed framework. Finally, we conclude the paper in section 4.

2. PROPOSED DIARIZATION FRAMEWORK: EEND-VECTOR CLUSTERING

2.1. Overall framework

Figure 1 shows a schematic diagram of the proposed EEND-vector clustering framework.

It first segments the input recording into chunks and calculates a sequence of the input frame features within each chunk, as $\mathbf{X}_i = (\mathbf{x}_{t,i} \mid t = 1, \dots, T)$ where i, t and T are the chunk index, the frame index in the chunk and the chunk size¹. $\mathbf{x}_{t,i} \in \mathbb{R}^K$ is the K -dimensional input frame feature at the time frame t . In the example shown in Fig 1, the input recording consists of 2 chunks and contains 3 speakers in total. In the following, we assume that we can fix the maximum number of active speakers in a chunk, S_{Local} , to 2, although the method could be generalized to more speakers or an unknown number of speakers [13]².

Based on the hyper-parameter $S_{\text{Local}} = 2$, the network estimates diarization results for 2 speakers in each chunk. In Fig. 1, the processing for the 1st speaker is drawn with black lines and put in the foreground, while that of the 2nd speaker is drawn with grey lines and put in the background. The diarization results are estimated independently in each chunk through NNs denoted as Encoder and

Linear_s^D ($s = 1, 2$), where s is the speaker index within a chunk. Since it is *not* always guaranteed that the diarization results of a certain speaker are estimated at the same output node, we may have the inter-block label permutation problem in the diarization outputs. As an example, in Fig. 1, the network Linear_1^D estimates the diarization result of ‘speaker A’ in the first chunk, and that of ‘speaker B’ in the second chunk. This means that we cannot obtain an optimal diarization result simply by stitching the diarization results of a specific output node across all the chunks.

To solve this permutation problem, we simultaneously estimate a speaker embedding corresponding to each diarization result in each chunk. The network to estimate the speaker embeddings are denoted as Linear_s^S ($s = 1, 2$) in Fig. 1. The speaker embedding extraction network is optimized through the NN training such that the vectors of the same speaker stay close to each other, while the vectors of different speakers lie far away from each other. This can be seen in the figure by examining how the embeddings are organized in the speaker embedding space. Therefore, after obtaining diarization results for all chunks, by clustering the speaker embeddings given the total number of speakers in the input recording (3 in this case), we can estimate the correct association of the diarization results among chunks. Then, finally, the overall diarization results are obtained by stitching them together based on the embedding clustering result. Note that while the proposed framework estimates the diarization results of the fixed number of speakers in a chunk, it can handle a meeting with an arbitrary number of speakers.

For the clustering, we can use any clustering algorithms. However, it may be preferable if the clustering algorithm is aware of the characteristic of this framework and work with a constraint that the speaker embeddings from a chunk should not belong to the same speaker cluster. In this paper, to incorporate the constraint into the clustering stage, we use a constrained k-means clustering algorithm called COP-k-means [18], which allows us to set cannot-link constraints between a given pair of embeddings to prevent the pair from being assigned to the same speaker cluster.

2.2. Neural diarization with speaker embedding estimation

This subsection details the NN model in EEND-vector clustering to estimate the diarization results and the speaker embeddings.

Let us denote the ground-truth diarization label sequence as $\mathbf{Y}_i = (\mathbf{y}_{t,i} \mid t = 1, \dots, T)$ that corresponds to \mathbf{X}_i . Here, the diarization label $\mathbf{y}_{t,i} = [y_{t,i,s} \in \{0, 1\} \mid s = 1, \dots, S_{\text{Local}}]$ represents a joint activity for S_{Local} speakers. For example, $y_{t,i,s} = y_{t,i,s'} = 1$ ($s \neq s'$) indicates both speakers s and s' spoke at the time frame t in the chunk i .

In the EEND framework, the diarization task is formulated as a multi-label classification problem. Specifically, we estimate the diarization result of the s -th speaker at each time frame, $\hat{y}_{t,i,s}$, as,

$$\begin{aligned} [\mathbf{h}_{1,i}, \dots, \mathbf{h}_{T,i}] &= \text{Encoder}(\mathbf{X}_i) \in \mathbb{R}^{D \times T}, \\ \hat{y}_{t,i,s} &= \text{sigmoid}(\text{Linear}_s^D(\mathbf{h}_{t,i})) \in (0, 1) \\ &\quad (s = 1, \dots, S_{\text{Local}}), \end{aligned} \quad (1)$$

where $\text{Encoder}(\cdot)$ is an encoder such as a multi-head self-attention NN [12], which utilizes all the input features \mathbf{X}_i for inference. $\mathbf{h}_{t,i}$ is a D -dimensional internal representation in the NN, $\text{Linear}_s^D(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^1$ is a fully-connected layer to estimate the diarization result, and $\text{sigmoid}(\cdot)$ is the element-wise sigmoid function.

Now, after estimating the diarization results, for the purpose of solving the inter-block permutation problem, we estimate the

¹The chunk size T for estimating speaker embeddings can be advantageously much longer than the homogeneous blocks used in x-vector clustering since we can handle *heterogeneous* chunks including more than 1 speaker.

²If we select the chunk size carefully, it is not too difficult to set an appropriate maximum number of speakers even for practical use cases [17].

speaker embedding, $\hat{\mathbf{e}}_{i,s}$, corresponding to the diarization result of the s -th speaker as follows.

$$\mathbf{z}_{t,i,s} = \text{Linear}_s^S(\mathbf{h}_{t,i}) \in \mathbb{R}^C, \quad (2)$$

$$\bar{\mathbf{z}}_{i,s} = \sum_{t=1}^T \hat{y}_{t,i,s} \mathbf{z}_{t,i,s} \in \mathbb{R}^C$$

$$\hat{\mathbf{e}}_{i,s} = \frac{\bar{\mathbf{z}}_{i,s}}{\|\bar{\mathbf{z}}_{i,s}\|} \in \mathbb{R}^C \quad (s = 1, \dots, S_{\text{Local}}), \quad (3)$$

where C is the dimension of the speaker embedding, $\text{Linear}_s^S(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^C$ is a fully-connected layer to estimate the s -th speaker's embedding $\mathbf{e}_{i,s}$, and $\|\cdot\|$ is a vector norm. Here we chose to estimate the speaker embeddings as weighted sum of frame-level embeddings $\mathbf{z}_{t,i,s}$ with weights determined by the diarization results $\hat{y}_{t,i,s}$, as in Eq. (2). With these operations, we can estimate diarization results and speaker embeddings for all S_{Local} speakers. This model without the speaker embedding estimator is essentially the same as the conventional EEND [11].

2.3. Training objectives

Now, we will explain a way to train the model to realize the behavior explained in Section 2.1. Since the network estimates both the diarization results and speaker embeddings simultaneously, our natural choice is to use the following multi-task loss.

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{diarization}} + \lambda\mathcal{L}_{\text{speaker}}, \quad (4)$$

where \mathcal{L} is the total loss function to be minimized, $\mathcal{L}_{\text{diarization}}$ is the diarization error loss, $\mathcal{L}_{\text{speaker}}$ is speaker embedding loss, and λ is a hyper-parameter to weight the two loss functions.

2.3.1. Diarization loss

Following [11], the diarization loss in each chunk is formulated as:

$$\mathcal{L}_{\text{diarization},i}^{\phi^*} = \frac{1}{TS_{\text{Local}}} \min_{\phi \in \text{perm}(S_{\text{Local}})} \sum_{t=1}^T \text{BCE}(\mathbf{1}_{t,i}^{\phi}, \hat{\mathbf{y}}_{t,i}), \quad (5)$$

where $\text{perm}(S_{\text{Local}})$ is the set of all the possible permutations of $(1, \dots, S_{\text{Local}})$, $\hat{\mathbf{y}}_{t,i} = [\hat{y}_{t,i,1}, \dots, \hat{y}_{t,i,S_{\text{Local}}}] \in \mathbb{R}^{S_{\text{Local}}}$, $\mathbf{1}_{t,i}^{\phi}$ is the ϕ -th permutation of the reference speaker labels, and $\text{BCE}(\cdot, \cdot)$ is the binary cross-entropy function between the labels and the estimated diarization outputs. ϕ^* is the permutation that minimizes the right hand side of the Eq. (5). This training scheme called permutation-invariant training has shown to be effective for the neural diarization [11], but at the same time, it incurs another problem, i.e., the inter-block label permutation problem since it clearly allows the speaker labels to permute from chunk to chunk. The diarization loss function $\mathcal{L}_{\text{diarization}}$ is formed by collecting B chunks, i.e., $\mathcal{L}_{\text{diarization}} = \sum_{i=1}^B \mathcal{L}_{\text{diarization},i}$, where B is the size of the mini-batch.

Here, as it was mentioned earlier, S_{Local} is a hyper-parameter that has to be appropriately chosen to satisfy (1) $S_{\text{Local}} \leq S_{\text{total}}$ where S_{total} is the total number of speakers in the recording, and (2) S_{Local} is always greater than or equal to the maximum number of speakers speaking in a chunk. With an assumption that S_{Local} is chosen in such a way, the diarization labels in the chunk i , \mathbf{Y}_i , should be formed as a subset of all S_{total} speaker's labels $\mathbf{Y}_i^{\text{total}}$, i.e., $\mathbf{Y}_i \subseteq \mathbf{Y}_i^{\text{total}}$. The subset should be chosen appropriately for each chunk such that it covers all speakers speaking in the chunk i . If the number of speakers speaking in the chunk is smaller than S_{Local} , we fill \mathbf{Y}_i with diarization label(s) of a virtual $(S_{\text{total}} + 1)$ -th always-silent speaker, i.e., $(y_{t,i,S_{\text{total}}+1} \in \{0\} \mid t = 1, \dots, T)$.

2.3.2. Speaker embedding loss

For the speaker embedding training, we use a loss function that encourages the embeddings to have small intra-speaker and large inter-speaker distances. Specifically, we utilize the loss proposed recently in [19], which was shown to be very effective for the speech separation task. For this loss function, we assume that the training data is annotated with speaker identity labels, i.e., indices, based on a finite set of M training speakers. Note, however, that the speaker identity is not required at test time, and that training and test speakers can differ (i.e., open speaker conditions). Let $\sigma_i^* = [\sigma_{i,1}^*, \dots, \sigma_{i,S_{\text{Local}}}^*]$ be the absolute speaker identity indices that correspond to the permutation of the labels that gives minimum value to Eq. (5), i.e., ϕ^* . σ_i^* is a subset of the M speaker identity indices. Then, the speaker embedding loss for chunk i , $\mathcal{L}_{\text{speaker},i}$, is formulated as follows.

$$\mathcal{L}_{\text{speaker},i} = \frac{1}{S_{\text{Local}}} \sum_{s=1}^{S_{\text{Local}}} l_{\text{speaker}}(\sigma_{i,s}^*, \hat{\mathbf{e}}_{i,s}), \quad (6)$$

where

$$l_{\text{speaker}}(\sigma_{i,s}^*, \hat{\mathbf{e}}_{i,s}) = -\ln \left(\frac{\exp(-d(E_{\sigma_{i,s}^*}, \hat{\mathbf{e}}_{i,s}))}{\sum_{m=1}^M \exp(-d(E_m, \hat{\mathbf{e}}_{i,s}))} \right), \quad (7)$$

$$d(E_m, \hat{\mathbf{e}}_{i,s}) = \alpha \|E_m - \hat{\mathbf{e}}_{i,s}\|^2 + \beta, \quad (8)$$

where E is a learnable global speaker embedding dictionary, and E_m is a learnable global speaker embedding associated with the m -th training speaker. Eq. (8) is the squared Euclidean distance between the learnable global speaker embedding and the estimated speaker embedding, which is rescaled with learnable scalar parameters $\alpha > 0$ and β . Eq. (7) is the log softmax over the distances between the estimated embedding and the global embeddings, which can be derived from the categorical cross-entropy loss. The loss function $\mathcal{L}_{\text{speaker}}$ is formed by collecting B chunks, similarly to $\mathcal{L}_{\text{diarization}}$.

By minimizing these loss functions, we expect to estimate diarization results accurately even if there is overlapped speech, and simultaneously estimate speaker embeddings that are suitable for the subsequent clustering process.

3. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed method in comparison with the conventional EEND [12], based on test data including long recordings with a significant amount of overlapped speech. Comparison with the x-vector clustering is omitted since it was already shown in [12] that the conventional EEND works better in case the data contains overlapped speech.

3.1. Data

The training, development, and test data are based on the 16 kHz Librispeech database [20]. To simulate a conversation-like mixture of two speakers, we picked up utterances from randomly selected two speakers, and generated a noisy reverberant mixture containing many utterances per speaker with reasonable silence intervals between utterances. For the simulation, we used the algorithm proposed in [11], and set the average silence interval between utterances at 2 seconds. Noise data was obtained from MUSAN noise data [21]. The signal-to-noise ratio was sampled randomly for each mixture

Table 1: DERs (%) of the conventional EEND and the proposed models for each test set that differs in the duration.

Model	Chunking	Clustering	Test data duration (minutes)			
			3	5	10	20
1. EEND	-	N/A	7.9	8.8	9.2	N/A
2. EEND	✓	N/A	9.9	9.9	10.2	9.9
3. Proposed	-	-	8.0	8.7	9.1	N/A
4. Proposed	✓	-	10.6	10.5	10.9	10.8
5. Proposed	✓	✓	9.1	8.2	7.9	7.7

Table 2: DERs (%) of the conventional EEND and the proposed EEND-vector clustering for each overlap condition.

Model	Chunking	Clustering	Overlap ratio (%)		
			0 - 30	30 - 60	60 - 90
EEND	-	-	10.5	9.4	7.1
Proposed	✓	✓	5.4	8.3	6.6

from 5, 10, 15, and 20 dBs. For reverberation, we used 20000 impulse response data in [22], which simulates various rooms. Consequently, we obtained a set of training, development, and test data that contains various overlapping ratios ranging from 10 to 90 %.

For the training and development data, we randomly selected utterances from 460-hour clean speech training data containing 1172 speakers ($M = 1172$) and generated 40000 and 500 mixtures that amount to 2774 and 23 hours, respectively. For the test data, we generated 4 different sets of data that differ in duration. Each test set contains 500 utterances. The average duration of mixtures in each set is 3, 5, 10, and 20 minutes, respectively. All the test data were generated based on the Librispeech test set containing 26 speakers that were not included in the training and development data.

3.2. NN training and hyper-parameters

For the input frame feature, we extracted 23-dimensional log-Mel-filterbank features with 25 ms frame length and 10 ms frame shift.

For both the proposed method and the conventional EEND, the chunk size T at the training stage was set at 500 (= 50 seconds) as in [12]. Therefore, when the training data is longer than 50 seconds, we split the input audio into non-overlapping 50-second chunks. At the inference stage, the conventional EEND uses an entire sequence for inference without chunking. On the other hand, the proposed method segments the input data into 50-second non-overlapping chunks, and perform diarization as explained in Section 2.1.

For both methods, we used the same network architecture as [12]. For Encoder, we used two multi-head attention blocks with 256 attention units containing four heads ($D = 256$). We used the Adam optimizer with the learning rate scheduler introduced in [23]. The number of warm-up steps used in the learning rate scheduler was 25000. The batch size B was 64. The number of training epochs was 70. The final models were obtained by averaging the model parameters of the last 10 epochs.

For the proposed method, λ was set at 0.01. With an assumption that the maximum number of speakers speaking in each chunk is 2, we set S_{Local} at 2. The dimension of the speaker embedding, C , was set at 256. Since the performance of the proposed method slightly changes due to the initialization of the COP-k-means algorithm, we ran the test inference 10 times with random initialization and obtained the averaged results. The standard deviation of the obtained diarization error rate (DER) was less than 0.2%.

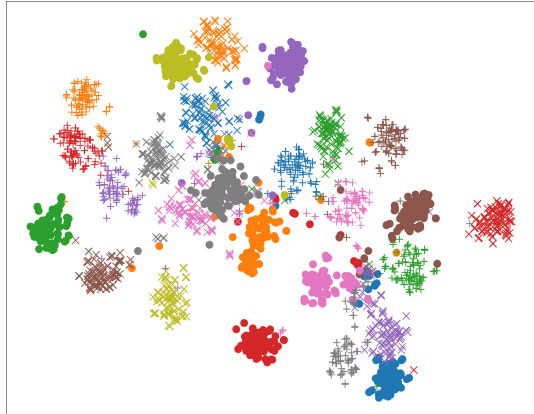


Fig. 2: t-SNE plot of the test speaker's embeddings vector

3.3. Results

Table 1 shows the results of the conventional EEND (1st row) and the proposed method (5th row). The table contains some variants of these methods to clarify the effectiveness of each component in the proposed model.

First, by comparing the 1st row (conventional EEND applied to the entire sequence without chunking) and 5th row (the proposed model that processes chunks and performs clustering, i.e., EEND-vector clustering), we can see that, as the duration of the test data gets longer, the proposed method becomes increasingly advantageous. While the conventional EEND cannot well handle 10- and 20-minute data because of poor generalization to the long data and the CPU memory constraint, EEND-vector clustering can achieve stable diarization performance for such data. Interestingly, it tends to work better (at least for this data) especially when the duration of the data is long. It is probably because the number of embeddings available for the clustering becomes larger as the data gets longer, which helps the clustering algorithm find better cluster centroids.

Now, let us compare the 1st row (EEND without chunking) and 3rd row (the proposed model applied to the entire sequence without chunking). The performance of the proposed model turned out to be almost equal to that of the conventional method in all cases, which indicates that the additional speaker loss did not negatively affect the diarization capability of the model. The results show that the additional speaker loss did not negatively affect the diarization capability of the model.

Next, let us focus on the comparison between 1st/3rd rows (models without chunking) and 2nd/4th rows (models with chunking but without clustering). The performance degradation when using chunking reveals the inter-block label permutation problem. We assume this problem may become even more severe when dealing with more speakers. With this comparison, we could confirm the effectiveness of the clustering-based diarization result stitching.

Overall, we found that, if the test data is shorter than 5 minutes, we can apply either the conventional EEND or the proposed model to the entire sequence (without chunking) to obtain a good diarization performance. On the other hand, if the data is longer than that, it is significantly better to use the proposed framework.

3.4. Detailed analysis

3.4.1. Evaluation in terms of overlapping ratio

Table 2 shows the DERs in each overlap condition. The results were obtained from the test set of 10-minute mixtures. Since each mixture in the test set differs in the amount of overlapped speech, i.e., overlap ratio, we categorized the mixtures into several overlap ratio ranges and obtained DER in each condition. , to better understand the model behavior. The proposed method is shown to largely outperform the conventional EEND in all conditions.

3.4.2. Speaker embedding estimation accuracy

Here we also examine whether the speaker embeddings of the test data is estimated accurately such that they have large inter-speaker and small intra-speaker distances. Figure 2 shows the t-SNE visualization of the speaker embeddings of the 26 test speakers. It clearly shows distinguished clusters for each speaker, which proves that we can estimate the *global* speaker embeddings accurately even if the input data contains a significant amount of overlapped speech.

4. CONCLUSIONS

We proposed a simple but effective diarization framework, EEND-vector clustering, that estimates both diarization results and speaker embeddings. By utilizing the speaker embeddings, we solved the inter-block label permutation problem. Experimental results showed that EEND-vector clustering works significantly better than the original EEND especially when the input data is long. Future work includes application of the proposed framework to more challenging conditions as well as an extension to a scheme that can handle an arbitrary number of speakers within a chunk, e.g., [13].

5. REFERENCES

- [1] X. Anguera, 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] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, *First DIHARD Challenge Evaluation Plan*, 2018, <https://zenodo.org/record/1199638>.
- [3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaikos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, , and P. Wellner, “The AMI meeting corpus: A pre-announcement,” in *The Second International Conference on Machine Learning for Multimodal Interaction*, ser. *MLMI’05*, 2006, pp. 28–39.
- [4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, , and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19(4), pp. 788–798, 2011.
- [5] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, , and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in *Proc. IEEE Spoken Language Technology Workshop*, 2016.
- [6] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, “Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge,” in *Proc. Interspeech 2018*, 2018, pp. 2808–2812.
- [7] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolikova, O. Novotný, K. Veselý, O. Glembek, O. Plchot, L. Mošner, and P. Matějka, “BUT system for DIHARD speech diarization challenge 2018,” in *Proc. Interspeech 2018*, 2018, pp. 2798–2802.
- [8] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, “Fully supervised speaker diarization,” in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6301–6305.
- [9] X. Li, Y. Zhao, C. Luo, and W. Zeng, “Online speaker diarization with relation network,” 2020, arXiv:2009.08162.
- [10] T. von Neumann and S. Araki T. Nakatani R. Haeb-Umbach K. Kinoshita, M. Delcroix, “All-neural online source separation, counting, and diarization for meeting analysis,” in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 91–95.
- [11] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, “End-to-end neural speaker diarization with permutation-free objectives,” in *Proc. Interspeech 2019*, 2019, pp. 4300–4304.
- [12] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, “End-to-end neural speaker diarization with self-attention,” in *Proc. IEEE ASRU*, 2019, pp. 296–303.
- [13] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, “End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors,” 2020, arXiv:2005.09921.
- [14] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct 2017.
- [15] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, “Listening to each speaker one by one with recurrent selective hearing networks,” in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5064–5068.
- [16] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, and K. Nagamatsu, “Online end-to-end neural diarization with speaker-tracing buffer,” 2020, arXiv:2006.02616.
- [17] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitriadis, “Low-latency speaker-independent continuous speech separation,” in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6980–6984.
- [18] K. Wagstaff, C. Cardie, S. Rogers, and S. S. Schroedl, “Constrained k-means clustering with background knowledge,” in *Proc. 18th International Conference on Machine Learning (ICML)*, 2001.
- [19] N. Zeghidour and D. Grangier, “Wavesplit: End-to-end speech separation by speaker clustering,” 2020, arXiv:2002.08933.
- [20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

- [21] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” 2015, arXiv:1510.08484.
- [22] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5220—5224.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5998—6008.