



Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech

Keisuke Kinoshita, Marc Delcroix, Naohiro Tawara

NTT Corporation, Japan

keisuke.kinoshita@ieee.org

Abstract

Recently, we proposed a novel speaker diarization method called End-to-End-Neural-Diarization-vector clustering (EEND-vector clustering) that integrates clustering-based and end-to-end neural network-based diarization approaches into one framework. The proposed method combines advantages of both frameworks, i.e. high diarization performance and handling of overlapped speech based on EEND, and robust handling of long recordings with an arbitrary number of speakers based on clustering-based approaches. However, the method was only evaluated so far on simulated 2-speaker meeting-like data. This paper is to (1) report recent advances we made to this framework, including newly introduced robust constrained clustering algorithms, and (2) experimentally show that the method can now outperform competitive diarization methods such as Encoder-Decoder Attractor (EDA)-EEND, on CALLHOME data which comprises real conversational speech data including overlapped speech and an arbitrary number of speakers. By further analyzing the experimental results, this paper also discusses pros and cons of the proposed method and reveals potential for further improvement.

Index Terms: Neural diarization, real data, overlapped speech

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].

Currently, there are mainly two major approaches to the diarization problem, that is, clustering-based approaches [1, 4–6] such as x-vector clustering, and End-to-End Neural Diarization (EEND) approaches [7–9]. The clustering-based approaches first segment a recording into short homogeneous blocks and compute a speaker embedding 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 diarization results [5, 6, 10, 11]. On the other hand, EEND is relatively simple. It receives standard frame-level spectral features and directly outputs a frame-level speaker activity for each speaker. In recent diarization challenges such as DIHARD-III [12], it is revealed that these two approaches are complementary to each other as it will be discussed below, and thus many institutes achieved reliable diarization for real conversational data by performing system combination of these approaches, e.g., [13].

To accomplish reliable diarization for any real conversational speech, the following essential problems have to be addressed:

- (1) overlapped speech (i.e., segments where more than one person is speaking)
- (2) long-form audio (e.g., duration of more than 10 min.),
- (3) an arbitrary number of speakers.

In each of these aspects, the aforementioned two approaches have different properties complementary to each other. The

clustering-based approaches have been studied for a decade and have been shown to work well with long-form audio containing an arbitrary number of speakers [1]. However, by nature of the assumption made in the extraction of speaker embeddings, i.e., single-speaker block assumption, it cannot handle overlapped speech. On the other hand, EEND was first developed to address the overlapped speech problem [7, 8]. Then, recently, it was extended to handle meetings containing an arbitrary number of speakers by introducing speaker counting functionality based on an Encoder-Decoder Attractor architecture (EDA) [9]. However, it was experimentally shown to still have difficulty in dealing with meetings containing a realistically large number of speakers, such as more than 3 speakers [9]. In addition, it was shown that it is difficult to directly apply the EEND systems to long-form audio (e.g., recordings longer than 10 minutes) [14]. Since the original EEND system was designed to operate in a batch processing mode, it inevitably requires a very large computer memory when performing inference with long recordings. Besides, aside from the memory issue, the neural networks (NNs) in EEND have difficulty generalizing to unseen very long sequential data. 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.

Focusing on these different pros and cons of the clustering and EEND approaches, we proposed a simple but effective hybrid diarization approach [14], called EEND-vector clustering, by combining the best of the clustering-based diarization and EEND. The framework allows us to process long-form audio containing overlapped speech and an arbitrary number of speakers. It first split the input long recording into fixed-length blocks. Then it applies a modified version of EEND to each block to obtain the diarization results for a fixed number of speakers as well as global speaker embedding vectors for each of the speakers. This assumes that for each short block the maximum number of speakers will be equal to or less than the number of the output of EEND. Finally, to solve the inter-block label permutation problem, speaker clustering is performed across blocks by using a constrained clustering algorithm. In [14], the EEND-vector clustering framework was shown to significantly outperform the conventional EEND [8] and x-vector clustering when processing *simulated* long-recordings of 2 speakers containing various overlap conditions, noise and reverberation, and thus was proven to be more advantageous in addressing the aforementioned problems (1) and (2).

However, it was not clear from our past studies whether the EEND-vector clustering could be generalized to *real* conversational speech data containing an arbitrary number of speakers (aforementioned problems (1) and (3)). To this end, this paper focuses on (i) application of the EEND-vector clustering to the widely used CALLHOME dataset [15], which consists of the real conversational speech of 2 to 6 speakers, and (ii) its evaluation in comparison with current state-of-the-art systems such as EDA-EEND [9], x-vector clustering [6], and Region-Proposal Network based Speaker Diarization (RPNSD) [16] that can handle CALLHOME data including overlapped speech. We also (iii) introduce practical techniques to increase robustness against real data such as more robust constrained clustering

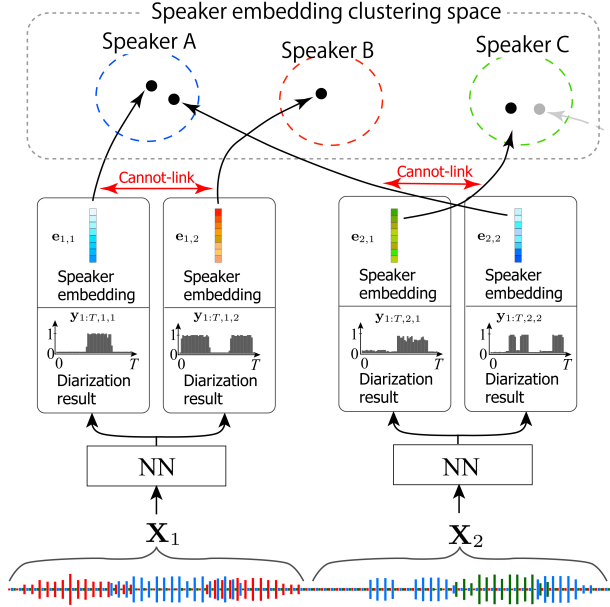


Figure 1: Schematic diagram of the proposed diarization framework. The input contains 3 speakers in total (red, green, and blue speakers shown in the waveform at the bottom), but only at most 2 speakers are actively speaking in each block.

methods and silent speaker detection.

In the remainder of the paper, we first review the proposed EEND-vector clustering approach (in Sec. 2), and then introduce the practical techniques required to deal with real meeting data (in Sec. 3). Finally, with experiments, we show that the EEND-vector clustering can outperform the other state-of-the-art approaches by a large margin, especially when the number of speakers is large.

2. EEND-vector clustering

2.1. Overall framework

Figure 1 shows a schematic diagram of the proposed EEND-vector clustering framework, originally proposed in [14].

It first segments the input recording into blocks and calculates a sequence of the input frame features within each block, as $\mathbf{X}_i = (\mathbf{x}_{t,i} \mid t = 1, \dots, T)$ where i, t and T are the block index, the frame index in the block and the block size, respectively. $\mathbf{x}_{t,i} \in \mathbb{R}^K$ 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 blocks and contains 3 speakers in total. In the following explanation, we assume that we can reasonably fix the maximum number of active speakers in a block, S_{Local} , to 2, for the sake of simplicity¹.

Based on the assumption/hyper-parameter $S_{\text{Local}} = 2$, the neural network NN always estimates, as neural network's outputs, diarization results and associated speaker embeddings for 2 speakers in each block. If a speaker is absent (i.e., there is only one active speaker in that block), the network simply estimates the diarization results of all zeros for that silent speaker. The diarization results are estimated independently in each block². 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. We can solve this permutation problem and es-

timate the correct association of the diarization results among blocks, by clustering the speaker embeddings given the total number of speakers in the input recording, S_{Global} , (3 in this case). Note that the speaker embedding extraction process is optimized 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. Based on the clustering results, we can stitch together the diarization results of the same speaker across blocks to obtain the final diarization output. Note that, while the proposed framework estimates the diarization results for the fixed number of speakers S_{Local} in a block, it can handle a meeting with an arbitrary number of speakers.

To tightly couple the embedding estimation and clustering process, it is beneficial to convey to the clustering algorithm the useful information that NN always estimates speech activities of two *different* speakers in each block. To this end, we suggest using a *constrained/semi-supervised* clustering algorithm which typically allows us to set a cannot-link constraint between a given pair of embeddings to prevent the pair from being assigned to the same speaker cluster.

2.2. Formulation of NN to jointly perform diarization and speaker embedding estimation

NN in Fig. 1 can be formulated as follows. Let us denote the ground-truth diarization label sequence as $\mathbf{Y}_i = \{\mathbf{y}_{t,i} \mid t = 1, \dots, T\} \in \mathbb{R}^{S_{\text{Local}} \times T}$ 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 block i . Similarly, let us denote the ground-truth speaker embedding set as $\mathbf{E}_i = (\mathbf{e}_{i,s} \mid s = 1, \dots, S_{\text{Local}}) \in \mathbb{R}^{S_{\text{Local}} \times C}$ that corresponds to \mathbf{X}_i . $\mathbf{e}_{i,s}$ is C -dimensional speaker embedding vector for the s -th speaker.

Then, the joint estimation of diarization results and speaker embeddings by NN in Fig. 1 is formulated as:

$$\hat{\mathbf{Y}}_i, \hat{\mathbf{E}}_i = \text{NN}(\mathbf{X}_i),$$

NN can be trained with a multitask loss function composed of diarization loss, i.e., binary cross entropy loss, and speaker embedding loss that encourages the embeddings to have small intra-speaker and large inter-speaker euclidean distances, as proposed in [14].

3. Handling real conversational data

This section summarizes techniques that we incorporated into the EEND-vector clustering to cope with real conversational data. While we should be able to improve the NN architecture and training procedure as in much other literature [8, 9] to push up the final performance, we found that improving only the constrained clustering part, which is a unique part of EEND-vector clustering, can already make a big difference in the final performance and help achieve state-of-the-art performance. The following subsections detail modifications we newly introduce to the original framework proposed in [14].

3.1. Improving constrained clustering

In [14], we employed a constrained clustering algorithm called COP-Kmeans [17] which is an extension of k-means clustering, and showed it was effective for handling simulated 2-speaker noisy reverberant meeting data. However, we neither compared it with other constrained clustering algorithms nor confirmed whether it is beneficial to use the cannot-link constraint. To thoroughly investigate the effectiveness of the constrained clustering algorithm in the proposed framework, here we incorporate and evaluate other constrained clustering algorithms in addition to COP-Kmeans. The standard k-means clustering optimizing squared error generally does not well handle

¹In the following experiments, we use $S_{\text{Local}} = 3$.

²For the details of NN and its training procedure, please refer to [14].

non-Gaussian data and/or data containing imbalanced classes, both of which are common for real conversational speech data. On the other hand, it was shown that Agglomerative Hierarchical Clustering (AHC) with an appropriate linkage criterion do not suffer from such limitations and thus has been widely used within clustering-based diarization systems [5, 6]. In addition, Spectral Clustering (SC) is often used for clustering-based diarization systems [10, 18], since it can also handle non-Gaussian data. In accordance with this development in the clustering for diarization, we here introduce the following constrained AHC and constrained SC into the proposed diarization framework.

3.1.1. Constrained Agglomerative Hierarchical Clustering

AHC is a common but effective unsupervised clustering technique used in state-of-the-art diarization systems [5, 6]. Standard AHC requires pairwise distances between all input samples, i.e., speaker embeddings $\hat{\mathbf{E}}_i$ ($i = 1, \dots, I$), resulting in a distance matrix. Then, based on the matrix, it generates a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. Each input sample starts in its own cluster, and pairs of clusters are merged as it moves up the hierarchy. We stop the cluster merging process either (a) when we obtain a required number of clusters, or (b) when the clusters are too far apart beyond a certain threshold to be merged. In diarization systems, the stopping criterion (a) is used to obtain diarization results given prior knowledge on the number of speakers in a meeting (i.e., oracle-number-of-speaker evaluation), while the stopping criterion (b) is used to obtain diarization results with the estimated number of speakers in a meeting (i.e., estimated-number-of-speaker evaluation).

To incorporate the cannot-link constraint to AHC, it is proposed to directly modify the distance matrix such that the distance between samples with the cannot-link constraint becomes larger than any other values in the matrix [19]. Within the proposed framework, the distance matrix $\mathbf{D}^{\text{dist.}} \in \mathbb{R}^{I \cdot S_{\text{Local}} \times I \cdot S_{\text{Local}}}$ we obtain based on $\hat{\mathbf{E}}_i$ ($i = 1, \dots, I$) can be expressed as

$$\mathbf{D}^{\text{dist.}} = \begin{bmatrix} \mathbf{D}_{1,1} & \cdots & \mathbf{D}_{1,I} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{I,1} & \cdots & \mathbf{D}_{I,I} \end{bmatrix} \quad (1)$$

where each element of the above block matrix, $\mathbf{D}_{i,j} \in \mathbb{R}^{S_{\text{Local}} \times S_{\text{Local}}}$, is a symmetric distance matrix calculated based on a set of speaker embedding vectors obtained at i -th block, $\hat{\mathbf{E}}_i$, and j -th block, $\hat{\mathbf{E}}_j$. When incorporating the cannot-link constraint, we insert certain large value κ into the off-diagonal components of matrices $\mathbf{D}_{i,i}$ for all i from 1 to I . After obtaining the modified distance matrix, we can use it in a standard AHC with the aforementioned stopping criteria.

3.1.2. Constrained Spectral Clustering

SC is another common unsupervised clustering technique used in recent diarization systems [10, 18]. SC is a technique with roots in graph theory, and it requires a pairwise similarity score between all input sample, i.e., speaker embeddings $\hat{\mathbf{E}}_i$ ($i = 1, \dots, I$), resulting in a similarity graph. When we have prior knowledge on the number of speakers, S_{Global} , in a meeting, we typically analyze the eigenvectors corresponding to S_{Global} smallest eigenvalues of unnormalized graph Laplacian constructed from the similarity graph. When we would like to estimate the number of clusters, we can use a widely used method called eigengap heuristics [20].

There are several ways to incorporate the cannot-link constraint to SC [21], varying from a simple way [22] to methods that can control the contribution of the constraint in a soft manner [21]. Here we employ a relatively straightforward method proposed in [22], which directly modifies the similarity graph such that the graph edge between samples with the cannot-link

is forced to 0. This can be done by constructing a matrix similar to eq. (1) but with similarity scores, and inserting 0 into the off-diagonal components of matrices $\mathbf{D}_{i,i}$ for all i from 1 to I .

3.2. Silent speaker detection

Before clustering speaker embeddings with a constrained clustering, it is beneficial to detect and exclude embeddings corresponding to silent speakers from a set used for the clustering. It is mainly because, if there are multiple silent speakers in a block, we should not set the cannot-link constraint between those embeddings, i.e., they should belong to the same silent speaker cluster. We found that, in many cases, the diarization results for those silent speakers stay very close to 0 (as we trained NN to do so). Therefore, we propose to detect it by examining whether the mean of the diarization results is sufficiently small; A speaker embedding is judged to be from a silent speaker if $\frac{1}{T} \sum_{t=1}^T \hat{y}_{t,i,s} < \tau$, where τ is a predetermined threshold.

4. Experiments

In this section, we evaluate the effectiveness of the proposed EEND-vector clustering in comparison with state-of-the-art conventional methods, based on CALLHOME dataset [15]. We also evaluate clustering algorithms in the proposed EEND-vector clustering framework to highlight the importance of the constrained clustering. The diarization error rate (DER) with collar tolerance of 0.25s is used as an evaluation metric as in [9].

4.1. Data

For the training, we used simulated mixtures created from Switchboard-2 (Phase I & II & III), Switchboard Cellular (Part 1 & 2), and the NIST Speaker Recognition Evaluation (2004 & 2005 & 2006 & 2008) for speech, and the MUSAN corpus [23] for noise with simulated room impulse responses used in [24], by following the data generation procedure in [7]. We created 3-speaker meeting-like dataset based on the algorithm proposed in [7] with $\beta = 10$.³

For evaluation and adaptation, we used the telephone conversation dataset CALLHOME (CH) [15], i.e., NIST SRE 2000 (LDC2001S97, Disk-8), which has been the most widely used dataset for speaker diarization studies. The CALLHOME dataset contains 500 sessions of multilingual telephonic speech. Each session has 2 to 6 speakers while there are two dominant speakers in each conversation. For evaluation and adaptation purpose, we split the CALLHOME dataset into two subsets according to [9], and performed adaptation on a subset and evaluation of the proposed method on the other subset.

4.2. Conventional methods to be compared with

The proposed method was compared with state-of-the-art methods, namely, x-vector clustering with PLDA scoring and variational Bayesian resegmentation [1, 4, 6], EDA-EEND [9] and Region-Proposal Network based Speaker Diarization (RPNSD) [16]. We did not reproduce their results, but simply borrow the results of x-vector clustering and EDA-EEND from [9], and that of RPNSD from [16].

4.3. Settings of the proposed EEND-vector clustering

4.3.1. NN training and hyper-parameters

In this experiment, the assumed maximum number of speakers in each block S_{Local} was set at 3 in the proposed method, i.e., NN always estimates diarization results and embedding vectors for 3 speakers including silent speaker(s). The block size T was varied from 50 s to 15 s to see how the block size will

³Each mixture contains dozens of utterances per speaker with reasonable silence intervals between utterances of the same speaker's. $\beta = 10$ means that the average duration of the silence interval is 10 s.

Table 1: *DER (%) of EEND-vector clustering with various clustering algorithms for CALLHOME dataset.*

Method	speaker counting	# of speakers in a session					
		2	3	4	5	6	All
Oracle clustering	oracle	7.51	9.66	12.34	12.93	22.01	10.21
Kmeans	oracle	8.54	21.22	24.79	34.16	31.97	18.68
COP-Kmeans [14]	oracle	9.25	19.55	24.49	35.71	41.03	18.74
SC	oracle	7.95	13.07	19.45	23.04	35.54	14.12
Constrained SC	oracle	7.94	12.17	19.56	20.12	35.95	13.73
AHC	oracle	7.96	11.90	15.16	25.01	28.36	12.59
Constrained AHC	oracle	8.08	11.27	15.01	23.14	26.56	12.22
SC	estimated	8.13	16.24	23.13	37.62	35.67	16.75
Constrained SC	estimated	8.71	16.25	23.51	37.62	43.10	17.38
AHC	estimated	9.54	12.49	21.49	24.41	32.18	14.78
Constrained AHC	estimated	7.96	11.93	16.38	21.21	23.10	12.49

have an impact on the diarization performance. To speed up the experiments, we first trained a model with $S_{\text{Local}} = 3$ and $T = 15$ s on the training data for 100 epochs, and adapted the pre-trained model to different block sizes T 's to obtain a model appropriate for each T .

For the NN architecture and training protocol, we basically followed [9], and followed [14] for speaker embedding estimation and multi-task training setting. We used self-attention-based six-layer stacked Transformer encoders with eight attention heads as a backbone of our method. The input for the network was the same as [9], i.e., 345-dimensional log-scaled Mel-filterbank-based features. The threshold τ to detect the silent speaker was set at 0.05.

4.3.2. Clustering algorithms in the proposed method

We evaluate oracle clustering, the following 3 unconstrained and 3 constrained clustering algorithms in the framework of the EEND-vector clustering, namely, k-means, AHC, SC, COP-Kmeans, constrained AHC, and constrained SC.

The oracle clustering corresponds to permutation that can yield a diarization result closest to the true one, based on a diarization result estimated by the network.

Using unconstrained clustering algorithms such as k-means, AHC, and SC, there may be some cases where speaker embeddings from a certain block are clustered into the same cluster. In such case, with an assumption that a certain speaker's speech activity was erroneously split into more than one output, we heuristically merge the diarization results corresponding to those speaker embeddings, by taking maximum across these diarization results at each time frame t .

For the constrained AHC, we set κ at 10000. We performed the AHC clustering such that it minimizes the average of the distances between all observations of pairs of clusters. The distance threshold above which, clusters will not be merged, was set at 1.

4.4. Results

4.4.1. Effect of clustering algorithms

Table 1 shows DERs of the proposed EEND-vector clustering with best-performing condition ($T = 30$ s), with different clustering methods performed with the oracle number of speakers and the estimated number of speakers. First, we can see that the constrained AHC performs the best and stays closest to the oracle clustering. In comparison with its unconstrained counterpart, it performs significantly better for difficult cases such as the number of speakers of 6. Hereafter, in this paper, we use constrained AHC as the clustering algorithm in the proposed method.

Except for AHC, it is not perfectly clear whether the incorporation of the cannot-link constraint is advantageous or not. The obtained results are very much affected by issues such as the imbalanced class problem, poor accuracy in speaker counting, which makes it difficult to make a definitive conclusion.

Table 2: *DERs (%) on CALLHOME diarization results with oracle number of speakers. T values in parentheses are the block sizes in EEND-vector clustering.*

Method	# of speakers in a session					
	2	3	4	5	6	All
x-vector clustering [9]	8.93	19.01	24.48	32.14	34.95	18.98
RPNSD [16]	N/A	N/A	N/A	N/A	N/A	17.06
EDA-EEND [9]	8.35	13.20	21.71	33.00	41.07	15.43
EEND-vector clust. ($T = 50$ s)	7.63	13.14	13.71	22.14	28.82	12.53
EEND-vector clust. ($T = 30$ s)	8.08	11.27	15.01	23.14	26.56	12.22
EEND-vector clust. ($T = 20$ s)	7.97	13.56	15.28	28.52	22.62	13.15
EEND-vector clust. ($T = 15$ s)	8.57	14.33	17.12	30.02	23.17	14.07

Table 3: *DERs (%) on CALLHOME diarization results with estimated number of speakers. T values in parentheses are the block sizes in EEND-vector clustering.*

Method	# of speakers in a session					
	2	3	4	5	6	All
x-vector clustering [9]	15.45	18.01	22.68	31.40	34.27	19.43
EDA-EEND [9]	8.50	13.24	21.46	33.16	40.29	15.29
EEND-vector clust. ($T = 50$ s)	7.18	12.50	16.91	28.04	26.76	12.98
EEND-vector clust. ($T = 30$ s)	7.96	11.93	16.38	21.21	23.10	12.49
EEND-vector clust. ($T = 20$ s)	8.13	12.83	16.75	31.90	23.08	13.40
EEND-vector clust. ($T = 15$ s)	9.77	16.27	16.12	27.21	23.16	14.84

4.4.2. Comparison with state-of-the-art methods

Tables 2 and 3 shows DER of the proposed EEND-vector clustering in comparison with state-of-the-art methods, performed with the oracle number of speakers (Table 2) and with the estimated number of speakers (Table 3). They first reveal that, comparing the best performing configuration of the proposed method ($T = 30$ s) with the other state-of-the-art methods, EEND-vector-clustering largely outperforms the others. It works significantly better than EDA-EEND especially when the number of speakers is large. The proposed method works generally better than x-vector-clustering most probably because it handles overlapped speech well and can estimate speaker embeddings from longer segments. It also outperforms RPNSD that combines NN-based diarization with clustering in a different manner.

However, if we decrease the size of the processing block T , the diarization performance tends to degrade. This issue should be resolved in the future, because, although for telephone speech data such as CALLHOME, speakers tend to speak for a long time and there are relatively few different speakers appearing in relatively long segments, it may not be always the case, especially when dealing with more causal real-life conversations. To deal with such data, we need to use a shorter block size to keep the maximum number of speakers in a block as small as $S_{\text{Local}} = 3$. It is not shown here because of space limitation, but, by examining the DERs more carefully, we found that, as we decrease the block size, missed speech and false alarm do not significantly increase but speaker confusion does. It suggests that if we could improve accuracy of the speaker embeddings or the clustering algorithm, we could bring about significant improvement even when the processing block is short, which will be part of our future works.

5. Conclusions

This paper evaluated the EEND-vector clustering framework based on real conversational speech dataset CALLHOME, and showed that it outperforms significantly the conventional state-of-the-art methods. We also experimentally showed the importance of constrained clustering in our framework. Future work includes investigations to improve the speaker embedding extraction and clustering performance when dealing with short speech segments.

6. 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. Karaiskos, 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] 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.
- [5] 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. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1893>
- [6] 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. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1749>
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] X. Li, Y. Zhao, C. Luo, and W. Zeng, "Online speaker diarization with relation network," 2020, arXiv:2009.08162.
- [12] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third DIHARD diarization challenge," 2021.
- [13] S. Horiguchi, N. Yalta, P. Garcia, Y. Takashima, Y. Xue, D. Raj, Z. Huang, Y. Fujita, S. Watanabe, and S. Khudanpur, "The hitachi-jhu DIHARD III system: Competitive end-to-end neural diarization and x-vector clustering systems combined by DOVER-Lap," 2021.
- [14] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (To appear)*, 2021.
- [15] M. Przybocki and A. Martin, *2000 NIST Speaker Recognition Evaluation (LDC2001S97)*. Philadelphia, New Jersey: Linguistic Data Consortium, 2001.
- [16] Z. Huang, S. Watanabe, Y. Fujita, P. García, Y. Shao, D. Povey, and S. Khudanpur, "Speaker diarization with region proposal network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6514–6518.
- [17] 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.
- [18] H. Aronowitz, W. Zhu, M. Suzuki, G. Kurata, and R. Hoory, "New advances in speaker diarization," in *Proc. Interspeech 2020*, 2018, pp. 279–283.
- [19] I. Davidson and S. S. Ravi, "Using instance-level constraints in agglomerative hierarchical clustering: theoretical and empirical results," *Data Mining and Knowledge Discovery*, vol. 77, no. 18, pp. 257–282, Dec. 2009.
- [20] U. von Luxburg, "A tutorial on spectral clustering," *Statist. and Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
- [21] X. Wang, B. Qian, and I. Davidson, "On constrained spectral clustering and its applications," *Data Mining and Knowledge Discovery*, vol. 28, pp. 1–30, Dec. 2014.
- [22] S. D. Kamvar, D. Klein, and C. D. Manning, "Spectral leawrning," in *Proc. Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, 2003, pp. 561–566.
- [23] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, arXiv:1510.08484.
- [24] 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.