END-TO-END SPEAKER DIARIZATION AS POST-PROCESSING

Shota Horiguchi[1]   *Paola García*[2]   *Yusuke Fujita*[1]   *Shinji Watanabe*[2]   *Kenji Nagamatsu*[1]

[1] Hitachi, Ltd. Research & Development Group, Japan
[2] Center for Language and Speech Processing, Johns Hopkins University, USA

ABSTRACT

This paper investigates the utilization of an end-to-end diarization model as post-processing of conventional clustering-based diarization. Clustering-based diarization methods partition frames into clusters of the number of speakers; thus, they typically cannot handle overlapping speech because each frame is assigned to one speaker. On the other hand, some end-to-end diarization methods can handle overlapping speech by treating the problem as multi-label classification. Although some methods can treat a flexible number of speakers, they do not perform well when the number of speakers is large. To compensate for each other's weakness, we propose to use a two-speaker end-to-end diarization method as post-processing of the results obtained by a clustering-based method. We iteratively select two speakers from the results and update the results of the two speakers to improve the overlapped region. Experimental results show that the proposed algorithm consistently improved the performance of the state-of-the-art methods across CALLHOME, AMI, and DIHARD II datasets.

Index Terms— Speaker diarization, EEND

1. INTRODUCTION

Speaker diarization, which is sometimes referred to as "who spoke when", has important roles in many speech-related applications. It is sometimes used to enrich transcriptions by adding speaker attributes [1], and at the other times, it is used to improve the performance of speech separation and recognition [2, 3].

Speaker diarization methods can be classified roughly into two: clustering-based methods and end-to-end methods. Typical clustering-based methods i) first classify frames into speech and non-speech, ii) then extract an embedding which describes speaker characteristics from each speech frame, and iii) finally apply clustering to the extracted embeddings. Most methods employ hard clustering such as agglomerative hierarchical clustering (AHC) and k-means clustering; as a result, each frame belongs either to one of the speaker clusters or to the non-speech cluster. The assumption that underlies these clustering-based methods is that each frame contains at most one speaker, *i.e.*, they treat speaker diarization as a set partitioning problem. Thus, they fundamentally cannot deal with overlapping speech. Despite the assumption, they are still strong baselines over end-to-end methods on datasets of a large number of speakers, *e.g.*, DIHARD II dataset [4]. This is because they handle multiple speaker problems based on unsupervised clustering without using any speech mixtures as training data. Thus, the methods do not suffer from overtraining due to the lack of the overlap speech especially for a large number of speakers.

On the other hand, some end-to-end methods called EEND treat speaker diarization as a multi-label classification problem. They predict whether each speaker is active or not at each frame; thus, they can deal with speaker overlap. Evaluation of the early models fixed the number of speakers to two [5, 6, 7]. Some extensions are proposed recently to handle unknown number of speaker cases, *e.g.*, encoder-decoder-based attractor calculation [8] and one-by-one prediction using speaker-conditioned model [9]. However, these methods still perform poorly when the number of speakers is large. One reason is the training datasets. Mixtures of a large number of speakers are often rare in various datasets; thus, end-to-end models cannot produce diarization results for large number of speakers because they are overtrained on mixtures of a few number of speakers. Even if the issue on the number of mixtures is solved, the EEND depends on the permutation invariant training [10] so that it is still hard to train the model on a large number of mixtures in terms of the calculation cost. For these reasons, how to handle mixtures that contain overlapping speech of a large number of speakers is still an open problem for both clustering and end-to-end diarization methods.

In this paper, we propose to combine both clustering-based and end-to-end methods effectively to deal with overlapping speech regardless of the number of speakers. We first obtain the initial diarization result using x-vector clustering, which does not produce overlapping results in most cases. We then apply the following steps iteratively: i) frame selection to contain only two speakers and silence and ii) overlap estimation using a two-speaker EEND model. The frame selection is also used to adapt the EEND model to a dataset which contains mixtures of more than two speakers. We evaluate our method using various datasets including CALLHOME, AMI, and DIHARD II datasets.

2. RELATED WORK

2.1. Clustering-based diarization

While some methods provide supervised clustering of speaker embeddings [11], the most common approach is an x-vector clustering in an unsupervised manner (See the systems submitted to DIHARD II Challenge, *e.g.*, [12, 13]). Since naive x-vector clustering results in poor performance, various techniques to improve the performance have been proposed, *e.g.*, probabilistic linear discriminate analysis (PLDA) rescoring [14] and Variational Bayes (VB) hidden Markov model (HMM) resegmentation [15]. In terms of overlap processing, most methods first detect overlapped frames and then assign the second speaker for the detected frames based on heuristics [12, 16] or the results of VB resegmentation [17].

Another direction is based on clustering of overlapped segments [18]. It first extracts overlapped segments using a region proposal network, and then applies clustering for embeddings extracted from each of them. It fundamentally solved the issue of embedding extraction using a sliding window, but its accuracies are not comparable to end-to-end methods described in Section 2.2.

2.2. End-to-end diarization for overlapping speech

One end-to-end approach is called EEND [5, 6, 7]. They calculate multiple speaker activities, each corresponding to a single speaker. Recent models can output a flexible number of speakers' activities by using encoder-decoder-based attractor calculation modules (EDA) [8] or speaker-conditional EEND (SC-EEND) [9]. Another approach is called RSAN, which are based on residual masks in the time-frequency domain to extract speakers one by one [19, 20].

While EEND and RSAN take only acoustic features as input, a variant of these methods also accepts a speaker embedding as input to determine the target-speaker and output his/her speech activities. For example, target-speaker voice activity detection (TS-VAD) uses i-vectors to output the corresponding speakers' voice activities [21], but the number of speakers is fixed by the model architecture. Personal VAD [22] and VoiceFilter-Lite [23], which are based on d-vectors, have not such a limitation, but they assume that each speaker's d-vector is stored in the database in advance; thus they are not suited for speaker-independent diarization.

3. PROPOSED METHOD

3.1. Overview

Given acoustic features $\{\boldsymbol{x}_t\}_{t=1}^T$, where $t \in \{1, \dots, T\} =: [T]$ denotes a frame index, diarization is a problem to predict a set of active frames $\mathcal{T}_k \subseteq [T]$ for each speaker $k \in \{1, \dots, K\} =: [K]$. K is the estimated number of speakers. For simplicity, we use $\mathcal{X}_{\mathcal{T}} := \{\boldsymbol{x}_t \mid t \in \mathcal{T}\}$ to denote the features of selected frames $\mathcal{T} \subseteq [T]$.

Clustering-based methods assume that input recordings do not contain speaker overlap. It formulates diarization as a set partitioning problem, i.e., $\mathcal{T}_k$ for $k \in [K]$ are predicted to be disjoint, i.e., $\forall \{i,j\} \in \binom{[K]}{2}, \mathcal{T}_i \cap \mathcal{T}_j = \emptyset$. In EEND, on the other hand, diarization is formulated as a multi-label classification to handle overlapping speech; thus, they do not have to be disjoint. The formulation of EEND is appropriate for real conversations in which speakers sometimes utter simultaneously. However, it makes the problem too difficult to be solved; when K is large (e.g. 10), it rarely happens that K speakers speak together. Therefore, we assume that at most $K'(< K)$ speakers speak simultaneously, and refine the clustering-based results using an end-to-end model that is trained to process at most K' speakers. In this study, we set $K' = 2$. The detailed algorithm is explained in Section 3.2.

3.2. Algorithm

Given initial diarization results $\{\mathcal{T}_k \mid \emptyset \neq \mathcal{T}_k \subseteq [T]\}_{k=1}^K$, we iteratively select two speakers among K and update the diarization results of the two speakers using an EEND model. The EEND model $f_{\text{EEND}} : \mathbb{R}^{D \times L} \to (0,1)^{2 \times L}$ was trained to estimate posteriors probabilities of two speakers from an L-length sequence of D-dimensional acoustic features. Figure 1 show the flow of the proposed method when $K = 3$.

3.2.1. Processing order determination

To apply the iterative refinement to each pair of speakers, the processing order influences the accuracy of final diarization results. This is because we cannot select frames to include only two speakers based on estimated diarization results because they include diarization errors. For example, if we select frames not containing Speaker 1 in Figure 1, the fourth frame contains Speaker 1 according to the final results. If the ratio of such impurities among the selected

frames is high, the refinement using EEND may not perform well. We found that this problem is simply solved by processing the pairs of speakers in decreasing order of the number of selected frames (Figure 1(i)). For each speaker pair $\{i, j\} \in \binom{[K]}{2}$, we first select a set of frames $\mathcal{P}_{i,j}$ not containing speakers other than i and j as follows:

$$\mathcal{P}_{i,j} = [T] \setminus \bigcup_{k \in [K] \setminus \{i,j\}} \mathcal{T}_k. \quad (1)$$

We then apply the refinement described below for each speaker pair in descending order of $|\mathcal{P}_{i,j}|$ as in Figure 1 (ii-a)–(ii-c).

3.2.2. Iterative update of diarization results

To update the diarization results of speakers i and j, we first reselect a set of frames $\mathcal{P}_{i,j}$ using (1) . This is because the diarization results $\{\mathcal{T}_k\}_{k=1}^K$ are updated at each refinement step so that we cannot reuse the one that is calculated to decide the processing order. Then the corresponding features $\mathcal{X}_{\mathcal{P}_{i,j}}$ are input to the EEND model to obtain posteriors of two speakers (a) and (b) by

$$\left(\left[q_t^{(a)}, q_t^{(b)} \right]^\top \middle| t \in \mathcal{P}_{i,j} \right) = f \left(\mathcal{X}_{\mathcal{P}_{i,j}} \right) \in (0,1)^{2 \times |\mathcal{P}_{i,j}|}, \quad (2)$$

where $q_t^{(a)}$ and $q_t^{(b)}$ denote posteriors of the first and second speakers at frame index t, respectively, and $(\cdot)^\top$ denotes the matrix transpose. We simply apply the threshold value of 0.5 to obtain the indexes of active frames of the two speakers as

$$\mathcal{Q}^{(a)} = \left\{ t \in \mathcal{P}_{i,j} \middle| q_t^{(a)} > 0.5 \right\}, \; \mathcal{Q}^{(b)} = \left\{ t \in \mathcal{P}_{i,j} \middle| q_t^{(b)} > 0.5 \right\}. \quad (3)$$

Note that we have speaker permutation ambiguity between (a)–(b) and i–j, and we solve permutation to find the optimal correspondence between $(\mathcal{T}_i, \mathcal{T}_j)$ and $(\mathcal{Q}^{(a)}, \mathcal{Q}^{(b)})$ as follows:

$$\left(\hat{\mathcal{T}}_i, \hat{\mathcal{T}}_j \right) = \underset{(u,v) \in \{(a,b),(b,a))\}}{\arg \max} s \left(\mathcal{Q}^{(u)}, \mathcal{T}_i \right) + s \left(\mathcal{Q}^{(v)}, \mathcal{T}_j \right), \quad (4)$$

where $s(\mathcal{U}, \mathcal{V})$ is a function to calculate similarity between speech and non-speech activities described by two sets $\mathcal{U}$ and $\mathcal{V}$ defined as

$$s(\mathcal{U}, \mathcal{V}) := \underbrace{|\mathcal{U} \cap \mathcal{V}|}_{\text{speech similarity}} + \underbrace{|([T] \setminus \mathcal{U}) \cap ([T] \setminus \mathcal{V})|}_{\text{non-speech similarity}}. \quad (5)$$

Finally, we update the diarization results of speakers i and j. To confirm that the new results $\hat{\mathcal{T}}_i$ and $\hat{\mathcal{T}}_j$ are calculated for speaker i and j, we check whether they satisfy the following conditions:

$$\frac{\left| \hat{\mathcal{T}}_i \cap (\mathcal{T}_i \cap \mathcal{P}_{i,j}) \right|}{|\mathcal{T}_i \cap \mathcal{P}_{i,j}|} > \alpha, \quad \frac{\left| \hat{\mathcal{T}}_j \cap (\mathcal{T}_j \cap \mathcal{P}_{i,j}) \right|}{|\mathcal{T}_j \cap \mathcal{P}_{i,j}|} > \alpha, \quad (6)$$

where α is a lower limit of the ratio of the intersection between the new results $\hat{\mathcal{T}}_i$ (or $\hat{\mathcal{T}}_j$) and the previous results $\mathcal{T}_i \cap \mathcal{P}_{i,j}$ (or $\mathcal{T}_j \cap \mathcal{P}_{i,j}$). In this study, we set $\alpha = 0.5$. Only if the conditions in (6) are satisfied, we update the results of speakers i and j. When $K = 2$, we simply update the results with the new ones as

$$\mathcal{T}_i \leftarrow \hat{\mathcal{T}}_i \cup ([T] \setminus \mathcal{P}_{i,j}), \quad \mathcal{T}_j \leftarrow \hat{\mathcal{T}}_j \cup ([T] \setminus \mathcal{P}_{i,j}). \quad (7)$$

On the other hand, when $K \geq 3$, such fully-update strategy causes a performance drop due to impurities in the selected frames. Thus, we use the following instead of (7) to update only overlapped frames:

$$\mathcal{T}_i \leftarrow \mathcal{T}_i \cup \left(\hat{\mathcal{T}}_i \cap \hat{\mathcal{T}}_j \right), \quad \mathcal{T}_i \leftarrow \mathcal{T}_j \cup \left(\hat{\mathcal{T}}_j \cap \hat{\mathcal{T}}_j \right). \quad (8)$$
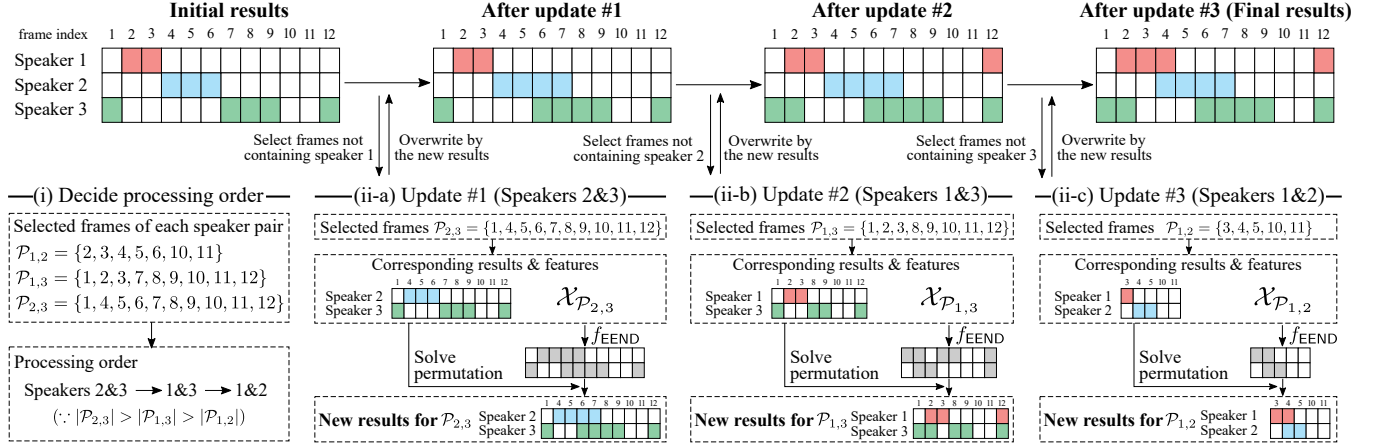
7189

Fig. 1. The flow of the proposed method when the number of speakers is three. Given initial diarization results (top left), our method (i) picks up a pair of speakers iteratively in decreasing order of the number of frames and (ii) refines the diarization results of the two speakers using an end-to-end speaker diarization model.

For the end-to-end model f_{EEND}, we use the self-attentive EEND model with an encoder-decoder attractor calculation module (SA-EEND-EDA) [8]. It consists of a four-layer-stacked Transformer encoder to extract embeddings for each frame and the EDA module to calculate attractors from the extracted embeddings. The EDA includes long short-term memories but we shuffled the order of embeddings just before they are fed into the EDA, which improves the diarization performance. Thus, we can consider that all the components of f_{EEND} are independent of the order of embeddings and therefore the model can treat input features of selected frames even if they are not continuous in time.

3.3. Training strategy of the SA-EEND-EDA model

In the original EEND and its derived methods [6, 8, 9] used *matched* dataset for model adaptation, *i.e.*, only two-speaker subset of the original dataset (*e.g.* CALLHOME[1]) was used to finetune the models in two-speaker evaluations. This strategy cannot be used to finetune two-speaker models when the dataset does not contain two-speaker mixtures (*e.g.* AMI [24]). Even if two-speaker mixtures are included in the dataset, it does not make full use of the datasets, which may cause performance degradation.

To cope with this situation, we adopt the frame-selection technique used in (1) for model adaptation. If the input chunk contains more than two speakers, we first choose two dominant speakers and then eliminate frames in which the other speakers are active as in (1). The model is trained only using the selected frames to output speech activities of the two speakers. This makes it possible to finetune two-speaker models from any kind of multi-speaker datasets without mixture-wise selection.

4. EXPERIMENT

4.1. Settings

Table 1 shows the datasets used for our evaluation. The model was pretrained using simulated two-speaker mixtures *Sim2spk* for 100 epochs. Each mixture was simulated from two single-speaker audios derived from Switchboard-2 (Phase I & II & III), Switchboard Cellular (Part 1 & 2), or NIST Speaker Recognition Evaluation (2004

---

[1] https://catalog.ldc.upenn.edu/LDC2001S97

Table 1. Dataset to train and test our diarization models.

Dataset		#Speakers	#Mixtures	Overlap ratio (%)
Pretrain	Sim2spk	2	100,000	34.1
Adaptation	CALLHOME-2spk	2	155	14.0
	CALLHOME	2–7	249	17.0
	AMI train [24]	3–5	118	19.4
	DIHARD dev [4]	1–10	192	9.8
Test	CALLHOME-2spk	2	148	13.1
	CALLHOME	2–6	250	16.7
	AMI eval [24]	3–4	24	18.6
	DIHARD eval [4]	1–9	194	8.9

& 2005 & 2006 & 2008). Noise sources from MUSAN corpus [25] and simulated room impulse responses [26] are also used to simulate noisy and reverberant environments. The detailed simulation protocol is in our previous paper [6]. For the pretraining, Adam optimizer with the learning rate scheduler proposed in [27] was used. The number of warm-up steps was set to 100,000 following [8].

After the pretraining, the model was adapted on CALLHOME, AMI [24] and DIHARD II [4] datasets for another 100 epochs, respectively. Adam optimizer was also used in the adaptations but its learning rate was fixed to 1×10^{-5} following [8].

We used diarization error rate (DER) and Jaccard error rate (JER) for evaluation. While some studies excluded overlapped regions from evaluation [14, 11], this study scored overlapped region. We also note that our evaluations are based on estimated speech activity detection (SAD), while some studies used oracle segments [17] or only reported confusion errors [11].

4.2. Preliminary evaluation of the training using frame selection

Before the evaluation of the proposed post-processing method, we first evaluated the training strategy explained in Section 3.3 using two transformer-based two-speaker EEND models: SA-EEND [6] and SA-EEND-EDA [8]. They were trained on CALLHOME-2spk in the original papers, but we utilized mixtures that contain more than two speakers in the CALLHOME dataset. Table 2 shows DERs on the CALLHOME-2spk test set. Using the full CALLHOME improved DER of SA-EEND from 9.54% to 9.00% and that of SA-EEND-EDA from 8.07% to 7.84%. According to these results, we show the effectiveness of our training strategy described in Sec-

Table 2. DERs (%) on CALLHOME-2spk. Collar tolerance of $0.25\,\text{s}$ is allowed.

Model	Adaptation	DER
SA-EEND [7]	CALLHOME-2spk	9.54
SA-EEND	CALLHOME + frame selection	9.00
SA-EEND-EDA [8]	CALLHOME-2spk	8.07
SA-EEND-EDA	CALLHOME + frame selection	**7.84**

Table 3. DERs (%) on CALLHOME. All the results include overlapped regions and are NOT based on oracle SAD. Collar tolerance of $0.25\,\text{s}$ is allowed.

Method	#Speakers 2	3	4	5	6	All
SA-EEND-EDA [8]	**8.50**	13.24	21.46	33.16	40.29	15.29
x-vector AHC	15.45	18.01	22.68	31.40	34.27	19.43
x-vector AHC + Proposed	13.85	14.72	18.61	**28.63**	29.02	16.79
x-vector AHC + VB	12.62	16.82	21.27	31.14	31.80	17.61
x-vector AHC + VB + Proposed	9.87	**13.11**	**16.52**	28.65	**27.83**	**14.06**

tion 3.3 based on the frame selection with (1).

4.3. Results

4.3.1. CALLHOME

We first evaluated the proposed method on CALLHOME dataset, which is composed of telephone conversations. As a clustering-based baseline, x-vectors with AHC and PLDA[2] was used with TDNN-based speech activity detection[3]. We also prepared the results for which VB-HMM resegmentation [15] was applied. All the components were implemented in Kaldi recipe.

Table 3 shows the evaluation results. X-vector clustering without and with VB achieved 19.43 % and 17.61 % DERs, respectively, but they didn't outperform the 15.29 % DER scored by SA-EEND-EDA trained to output diarization results on flexible number of speakers. However, we can also observe that the clustering-based methods are better when the number of speakers is larger than four. Applying the proposed post-processing for x-vector clustering baselines achieved 16.79 % and 14.06 % without and with VB, and the latter is 1.23 % better than the SA-EEND-EDA model. In terms of the number of speakers, the proposed method performed well on both large and small number of speakers.

4.3.2. AMI

Second, we evaluated our method on AMI dataset, consisting of meeting recordings. While it includes various types of recordings, we used *Headset mix* recordings for this experiment. We chose the system developed during JSALT 2019 [28] as a baseline. It is based on x-vector clustering followed by VB resegmentation and overlap detection and assignment for the second speaker candidate [29].

Table 4 shows DERs and JERs on AMI eval set. The proposed method reduced DERs of 3.07 %, 3.14 %, and 0.18 % of absolute improvement from the three baselines. Surprisingly, our method im-

Table 4. DERs and JERs (%) on AMI eval. VB: Variational Bayes resegmentation, OVL: Overlap detection and speaker assignment [29]. All the results include overlapped regions and are NOT based on oracle SAD. No collar tolerance is allowed.

Method	DER	JER
x-vector AHC [28]	33.75	45.68
x-vector AHC + Proposed	30.64	43.78
x-vector AHC + VB [28]	32.80	43.72
x-vector AHC + VB + Proposed	29.66	42.63
x-vector AHC + VB + OVL [28]	28.15	41.00
x-vector AHC + VB + OVL + Proposed	**27.97**	**40.57**

Table 5. DERs and JERs (%) on DIHARD II eval. All the results include overlapped regions and are NOT based on oracle SAD. No collar torelance is allowed.

Method	DER	JER
DIHARD II baseline [30]	40.86	66.60
DIHARD II baseline + Proposed	37.90	63.79
BUT system (w/o OVL) [12, 16]	27.26	49.15
BUT system (w/o OVL) + Proposed	26.91	48.49
BUT system (w/ OVL) [12, 16]	27.11	49.07
BUT system (w/ OVL) + Proposed	**26.88**	**48.43**

proved DER and JER of the results in which the overlap detection [29] was already applied.

4.3.3. DIHARD II

Finally, we evaluated the proposed method on DIHARD II dataset, which includes recordings from 10 different domains. We used the official baseline system [4] and the BUT system [12, 16], which is the winning system of the second DIHARD Challenge, to obtain initial diarization results. Both are based on the x-vector clustering, but the BUT system is more polished in that it extracts x-vectors in shorter intervals and uses VB resegmentation and overlap detection and assignment based on heuristics.

The results are shown in Table 5. The proposed method reduced DER and JER of the baseline system by 2.96 % and 2.91 %, respectively. Our method also improved DER and JER of 0.35 % and 0.66 % from the BUT system without overlap assignment and 0.38 % and 0.64 % from that with overlap assignment, respectively. These improvements are small, but it is significantly better than the heuristic-based overlap assignment in [12], which improved DER by 0.15 % (= 27.26 − 27.11) and JER by 0.08 % (= 49.15 − 49.07).

5. CONCLUSION

In this paper, we proposed a post-processing method for clustering-based diarization using an end-to-end diarization model. We iteratively selected two speakers, picked up frames that contain the two speakers, and process the frames by the end-to-end model to update diarization results. Evaluations on CALLHOME, AMI, and DIHARD II datasets showed that our proposed method improves various types of clustering-based diarization results.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE TASLP*, vol. 20, no. 2, pp. 356–370, 2012.

[2] C. Boeddeker, J. Heitkaemper, J. Schmalenstoeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME-5*, 2018.

[3] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "The STC system for the CHiME-6 Challenge," in *CHiME-6*, 2020.

[4] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, task, and baselines," in *INTERSPEECH*, 2019, pp. 978–982.

[5] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *INTERSPEECH*, 2019, pp. 4300–4304.

[6] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *ASRU*, 2019, pp. 296–303.

[7] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, and K. Nagamatsu, "End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification," arXiv:2003.02966, 2020.

[8] S. Horiguchi, Y. Fujita, S. Wananabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *INTERSPEECH*, 2020, pp. 269–273.

[9] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, J. Shi, and K. Nagamatsu, "Neural speaker diarization with speaker-wise chain rule," arXiv:2006.01796, 2020.

[10] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[11] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP*, 2019, pp. 6301–6305.

[12] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, "BUT system for the Second DIHARD Speech Diarization Challenge," in *ICASSP*, 2020, pp. 6529–6533.

[13] Q. Lin, W. Cai, L. Yang, J. Wang, J. Zhang, and M. Li, "DIHARD II is still hard: Experimental results and discussions from the DKU-LENOVO team," in *Odyssey*, 2020, pp. 102–109.

[14] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *SLT*, 2014, pp. 413–417.

[15] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on Bayesian HMM with eigenvoice priors," in *Odyssey*, 2018, pp. 102–109.

[16] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, "BUT system description for DIHARD Speech Diarization Challenge 2019," arXiv:1910.08847, 2019.

[17] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Černocký, "Bayesian HMM based x-vector clustering for speaker diarization," in *INTERSPEECH*, 2019, pp. 346–350.

[18] Z. Huang, S. Watanabe, Y. Fujita, P. García, Y. Shao, D. Povey, and S. Khudanpur, "Speaker diarization with region proposal network," in *ICASSP*, 2020, pp. 6514–6518.

[19] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *ICASSP*, 2018, pp. 5064–5068.

[20] K. Kinoshita, M. Delcroix, S. Araki, and T. Nakatani, "Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system," in *ICASSP*, 2020, pp. 381–385.

[21] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, Podluzhny, et al., "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," in *INTERSPEECH*, 2020, pp. 274–278.

[22] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. L. Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Odyssey*, 2020, pp. 433–439.

[23] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, and A. Gruenstein, "VoiceFilter-Lite: Streaming targeted voice separation for on-device speech recognition," in *INTERSPEECH*, 2020, pp. 2677–2681.

[24] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[25] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," arXiv:1510.08484, 2015.

[26] 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 *ICASSP*, 2017, pp. 5220–5224.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[28] P. Garcia, J. Villalba, H. Bredin, J. Du, D. Castan, A. Cristia, L. Bullock, L. Guo, K. Okabe, P. S. Nidadavolu, et al., "Speaker detection in the wild: Lessons learned from JSALT 2019," in *Odyssey*, 2020, pp. 415–422.

[29] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *ICASSP*, 2020, pp. 7114–7118.

[30] 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 *INTERSPEECH*, 2018, pp. 2808–2812.