INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Scenario-Dependent Speaker Diarization for DIHARD-III Challenge

Yu-Xuan Wang[1], *Jun Du*[1,*], *Mao-Kui He*[1], *Shu-Tong Niu*[1], *Lei Sun*[2], *Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]iFlytek Research, Hefei, Anhui, China
[3]Georgia Institute of Technology, Atlanta, GA, USA

yxwang1@mail.ustc.edu.cn, jundu@ustc.edu.cn, hmk1754@mail.ustc.edu.cn,
niust@mail.ustc.edu.cn, sunlei17@mail.ustc.edu.cn, chl@ece.gatech.edu

Abstract

In this study, we propose a scenario-dependent speaker diarization approach to handling the diversified scenarios of 11 domains encountered in DIHARD-III challenge with a divide-and-conquer strategy. First, using a ResNet-based audio domain classifier, all domains in DIHARD-III challenge could be divided into several scenarios by different impact factors, such as background noise level, speaker number, and speaker overlap ratio. In each scenario, different combinations of techniques are designed, aiming at achieving the best performance in terms of both diarization error rate (DER) and run-time efficiency. For low signal-to-noise-ration (SNR) scenarios, speech enhancement based on a progressive learning network with multiple intermediate SNR targets is adopted for pre-processing. Conventional clustering-based speaker diarization is utilized to mainly handle speech segments with non-overlapping speakers, while separation-based or neural speaker diarization is used to cope with the overlapping speech regions, which is combined with an iterative fine-tuning strategy to boost the generalization ability. We also explore post-processing to perform system fusion and selection. For DIHARD-III challenge, our scenario-dependent system won the first place among all submitted systems, and significantly outperforms the state-of-the-art clustering-based speaker diarization system, yielding relative DER reductions of 32.17% and 28.34% on development set and evaluation set on Track 1, respectively.

Index Terms: speaker diarization, scenario-dependent processing, speech separation, neural speaker diarization, speech enhancement

1. Introduction

Speaker diarization is the task of determining "who spoke when" in the given audio signal [1, 2]. It can also act as the pre-processing for automatic speech recognition (ASR) [3, 4] in many realistic application domains.

The NIST Rich Transcription evaluation first focused on the diarization performance on meeting speech [5]. The traditional speaker diarization system consists of several modules, including speech activity detection (SAD), speech segmentation, speaker embedding extraction, and clustering. Because there was no unified task, research on different single domains such as telephone [6, 7], broadcast [8] and meeting [9] continued. This made it hard to compare performance, and when it turned to other real-world challenging domains, such as restaurant and web videos, the performance degraded dramatically.

To attract researchers' interest on more challenging domains, DIHARD-I challenge [10] was held, bringing the dataset drawn from a diverse range of domains. The difficulty for the challenge comes from the variety of the scenarios and the overlapping speech. And the traditional system performed poorly [11, 12]. For the following DIHARD-II challenge [13], Brno University of Technology (BUT) won the first place by the fit-all-domain clustering-based speaker diarization system [14]. The Gaussian Mixed Model (GMM) based speaker embedding i-vector [15] is replaced by neural network based x-vector [16] for the better performance. Two-stage clustering is performed with the first-stage agglomerative hierarchical clustering (AHC) [17] providing an initial under-clustering result and the second-stage Variational Bayes Hidden Markov Model (VB-HMM) [18] refining the initial result. Some teams [19, 20] also tuned specific thresholds when clustering to perform domain-dependent processing.

The methods mentioned above are all based on unsupervised clustering, which considers each frame contains at most one speaker, so they can not deal with overlapping speech. Even combined with the overlap detection and resegmentation, the performance boost is small. In order to directly solve the overlap problem and minimize the diarization errors, the neural network-based speaker diarization, such as end-to-end neural speaker diarization (EEND) [21, 22] and target-speaker voice activity detection (TS-VAD) [23] were proposed. They judge each speaker's activeness for each frame, so they can fundamentally estimate multiple speakers at the same time. But the limitation lies on the total number of speakers is fixed. Further research was taken on handling unknown number of speakers on EEND [24]. But it doesn't work well when the number of speakers is higher than four. Besides, conformer-based continuous speech separation (CSS) [25] was taken as the pre-processing for the diarization to help to improve performance.

In this paper, we propose a novel scenario-dependent speaker diarization pipeline for DIHARD-III challenge [26, 27]. Different from the domain-dependent approach that adjusts suitable parameters for specific domain [19, 20], we combine different techniques according to the scenarios grouped based on the domain classification results. Not only scenario-dependent processing can improve the diarization performance more efficiently, but also it can reduce the risk of performance degradation compared with adopting the same strategy on all the scenarios. For overlapped scenarios, clustering-based speaker diarization first provides the initial diarization results, then separation-based or neural speaker diarization further estimates the remaining speakers embedded in current speech. Besides, speech enhancement is applied on specific noisy scenarios to enhance the speech quality, and post-processing is performed to further improve the overall performance. The DIHARD-III system description [28] of our pipeline intuitively describes all the technical details, so in this study we focus on explaining the motivation of scenario-dependent strategy and iterative fine-tuning strategy, the details can refer to [28].

* corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-516

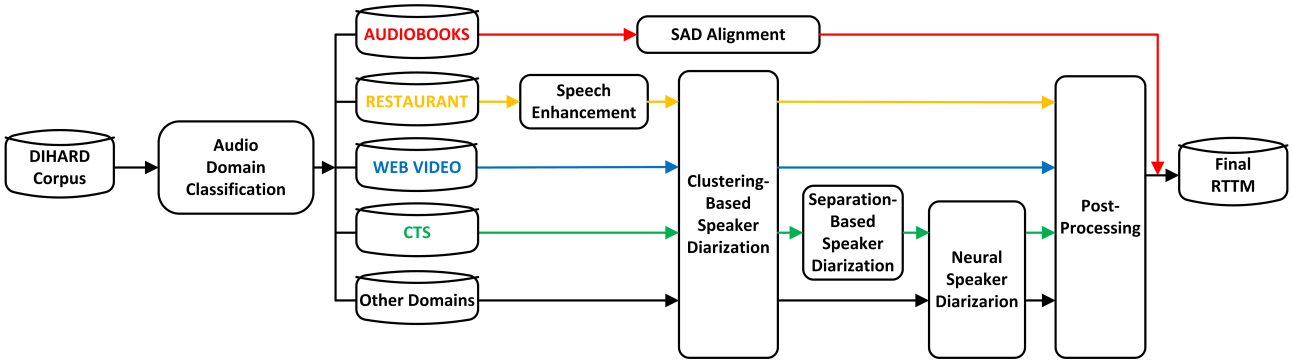

Figure 1: *The overall structure of our proposed pipeline for the DIHARD-III challenge.*

2. Motivation

The main scenario impact factors for each domain on the full development set of DIHARD-III challenge are listed in Table 1 to show how the domains differ from each other and how hard the challenge is. Overlapping speech interferes with the diarization process, as we assign a frame with only one speaker during clustering. And the overlap ratio can exactly represent the proportion of overlapping speech time to total scored speaker time. Besides, the speaker number can also reflect the diarization difficulty, because the more speakers embedded in one session, the harder it is for the clustering to distinguish between speakers. As observed in Table 1, the overlap ratio for most domains is relatively low, among which AUDIOBOOKS has non overlapping speech with only one speaker embedded in one session. But in specific domains like MEETING, WEB VIDEO and RESTAURANT, the overlapping speech is extremely frequent with a wide range of speaker numbers. What's more, loud background noise existing in some domains like RESTAURANT can also obscure speaker's identity, accordingly degrade the diarization performance. The diversified scenarios drive us to develop different solutions specifically and efficiently, so the scenario-dependent strategy is adopted.

Table 1: *Summary of average overlap ratio (%) and speaker number per session for each domain on full development set.*

Domain	Overlap Ratio (%)	Speaker Number
AUDIOBOOKS	0.00	1
BROADCAST.	1.18	3-5
COURTROOM	1.90	5-10
MAP TASK	2.93	2
CLINICAL	4.55	2
SOC.LAB	4.78	2
SOC.FIELD	7.53	2-6
CTS	11.97	2
MEETING	22.43	3-7
WEB VIDEO	21.70	1-7
RESTAURANT	25.20	4-7

Figure 1 illustrates the overall structure of our proposed scenario-dependent pipeline for the DIHARD-III challenge. It consists of audio domain classification, speech enhancement, clustering-based speaker diarization (CSD), neural speaker diarization (NSD), separation-based speaker diarization (SSD), and post-processing. We will elaborate the details of each module in the following section.

3. Proposed Scenario-Dependent Pipeline

3.1. Audio Domain Classification

We first adopt a residual neural network (ResNet) [29] with 17 convolutional layers to classify the diversified eleven domains. All the following modules needed are performed based on the results from audio domain classification. Each recording session from the whole DIHARD corpus is assigned with a specific domain label predicted by the audio domain classification module during testing, and then all the sessions belonging to the same label are grouped together. Domains with similar scenarios are further grouped for the following processing.The accuracy achieved on the leave-one-out cross validation development set was 100%. The domain labels for the evaluation set were released after the challenge, we directly evaluate our classification results and get an accuracy of 80.7% for the full evaluation set, and 64% of the errors come from the confusion between CLINICAL and SOCIOLINGUISTIC LAB data, which both are two-speaker mixed domains with quite low overlap ratios, so can be regarded as the same scenario.

3.2. Speech Enhancement

Many sessions contain background noise due to the complex recording environments, among which all the sessions belonging to the RESTAURANT domain have loud babble noise in common, including background speech from neighboring tables (sometimes at levels close to that of the target speakers), clinking silverware, moving chairs or tables, and loud music. The existence of background noise interferes with the clustering process, contributes to the increasement of diarizing difficulty, so we perform speech enhancement on the RESTAURANT data. For other domains, not all sessions are with persistent low SNRs. Accordingly, speech enhancement is not applied to reduce the risk of degrading the diarization performance in high-SNR sessions. Here we employ the progressive multi-target network based speech enhancement model in [30]. The enhanced speech can be recovered from each intermediate layer by progressively enhanced log-power spectra (PELPS) or progressively ratio mask (PRM) features, and we finally adopt the speech recovered by PELPS from the first hidden layer to achieve the best performance.

3.3. Clustering-Based Speaker Diarization

Here we directly refer to the clustering-based speaker diarization (CSD) system of BUT in DIHARD-II [14], which can be regarded as the scenario-independent system treating each do-

main equally, and has shown its poweful generalization ability on different domains. The x-vectors are first extracted with a time delay neural network (TDNN) [31] for each speech segment divided by SAD, and then clustered by means of AHC with similarity metric based on probabilistic linear discriminant analysis (PLDA) log-likelihood ratio scores [32], followed by VB-HMM based clustering to create the diarization results.

3.4. Neural Speaker Diarization

The performance of CSD is good enough to deal with most of the domains, but it can't well handle overlapping speech. Here we adopt neural speaker diarization (NSD) to cope with overlapped regions in multi-speaker mixed domains. TS-VAD can predict the per-frame speech activities for all the speakers simultaneously, and we further improve the original TS-VAD. Considering that the TS-VAD can only process sessions with a fixed number of speakers, so we improve the TS-VAD to accommodate the situation that the number of speakers varies from session to session. Moreover, to improve the generalization ability of TS-VAD for diversified mismatched domain data, an iterative fine-tuning strategy is designed to optimize the current session. The NSD using TS-VAD is built on CSD because during the fine-tuning stage we first decode the TS-VAD pre-trained model with i-vectors extracted from CSD results. More details about iterative fine-tuning procedure and the strategy for variable speakers can refer to [28].

3.5. Separation-Based Speaker Diarization

For the two-speaker mixed domains, we can apply separation-based speaker diarization (SSD) as the first-stage overlapping processing, followed by NSD as the second-stage overlapping processing to achieve a better performance. We also combine traditional speech separation and iterative fine-tuning strategy together, the former separates the original mixed waveform into two single-speaker streams using a fully convolutional time-domain audio separation network (ConvTasNet) [33], and then detects speaker presence using a DNN-based SAD, and the latter contributes to the generalization ability. The fine-tuning procedure is also taken based on CSD results, which first provides the pre-trained model with the speaker priors to simulate two-speaker conversational data. The more details about SSD with iterative fine-tuning procedure can be found in [28].

3.6. Post-Processing

We first apply Dover-lap [34] on fusing different module results mentioned above, including the results from CSD, different iteration results of different fine-tuning stages of NSD or SSD. We then select the best result among different fusion and single module results (as the fusion result is not always the best) based on the performance of the development set, which is replayed on the evaluation set. Finally, we detect the laughter segments using the speech recognition information where multiple speakers laugh at the same time but we only find out one speaker most of the time. Here we artificially attach the neighborhood speakers. Until now, we can get the final results in the form of Rich Transcription Time Marked (RTTM) files.

3.7. Scenario-Dependent Processing

As shown in Table 1, for the non-overlapping scenario AUDIOBOOKS, each session consists of a single speaker, the data detected just need to be segmented according to the SAD information, and then all segments are assigned to the same speaker. We ignore any subsequent clustering or separating process, which can avoid unnecessary calculations.

For the overlapped scenarios, we perform CSD to deal with the non-overlapped regions, it can also provide quite accurate and abundant prior information to NSD to cope with overlapping speech with the exception of scenarios with high overlap ratios. The existence of the large number of overlapped regions seriously interferes with the clustering process in CSD, making the initial diarization results quite inaccurate. Accordingly NSD leads to performance degradation due to the less accurate priors, and the wrong assignment results in the error accumulation during simulating data in the fine-tuning stage. But we use AMI [35] corpus for TS-VAD model training, so TS-VAD performs well on the approximately in-domain MEETING data. Besides, we point out that all sessions of RESTAURANT data contain loud noise, and speech enhancement can really improve the speech quality and do good to diarization.

We perform SSD as the first-step overlapping speech processing on CTS (Conversational Telephone Speech). CTS is a telephone conversation domain consisting of two English speakers per session. Due to the informal conversation contents and non face-to-face form, quite a few overlapped speech are embedded in each session. SSD can further significantly improve the performance on the overlapped segments compared with only using NSD overlapping processing.

All the domains go through the post-processing module in the end except AUDIOBOOKS. The post-processing needs the audio domain classification information as the fusion and selection are conducted in a domain-dependent manner.

4. Experiments and Result Analysis

4.1. Experimental Setup

For speech enhancement module, we employed a 3-layer long short-term memory (LSTM) [36] network with 1024 cells in each layer. The clean speech were from WSJ0, AIShell-1, THCHS-30, and Librispeech. The noise included 115 types of noise and MUSAN corpus. The total duration of the training set was 1000 hour. For the NSD module, we adopted Switchboard-2, AMI Meeting Corpus, Voxconverse dev set and simulated data by Librispeech to train the TS-VAD pre-trained model, totally 2500-hour data. And we simulated about 4-hour data for each session during fine-tuning stage. For the SSD module, the 250-hour 2-speaker simulated data using Librispeech was adopted for the ConvTasNet pre-trained model. And 2 to 3-hour audios were simulated for each session for fine-tuning. More details about the training data sources, the network structure for each sub-module, as well as the configuration of hyper-parameters can be seen in [28]. Two partitions of the DIHARD-III datasets are defined, that is, core dataset and full dataset, respectively corresponding to the dataset with roughly the equal duration and unbalanced duration on each domain. The diarization performance was assessed by the primary metric diarization error rate (DER) in this paper, where DER (%) is the total percentage of reference speaker time that is not correctly attributed to a speaker. And we only introduced the performance on Track 1 with oracle SAD here due to the lack of space.

4.2. Effect of Speech Enhancement

For the noisy scenario RESTAURANT, we apply speech enhancement (SE) first of all, then perform speaker diarization using CSD. The performance comparison can be obtained in Table 2. We achieve a significant DER reduction of 5.68% by denois-

ing on the development set and 3.94% on the evaluation set. The corresponding reduction on the evaluation set is slightly less than that on the development set, partly due to the audio domain classification errors.

Table 2: *DER(%) performance comparison on full development set and full evaluation set of RESTAURANT data among different systems.*

Domain	CSD		SE + CSD	
	Dev	Eval	Dev	Eval
RESTAURANT	43.82	43.20	38.14	39.26

4.3. Effect of Scenario Categorization

For the remaining eight domains except AUDIOBOOKS and WEB VIDEO, which don't need overlapping processing for avoiding unnecessary calculations and performance degradation, respectively, we use CSD to diarize the non-overlapped speech segments, and NSD to diarize the overlapped speech regions. As shown in Table 3, NSD has brought performance improvement for all the domains on the development set, but audio domain classification errors and scenario changes lead to performance degradation, of which the domain classification errors contribute to the mismatched technique combinations on the single sessions and mismatched processing, and the scenario changes including the obvious changes of overlap ratios and SNR levels further interfere with the classification results, so affect the performances of the first three domains of BROADCAST INTERVIEW, COURTROOM and MAP TASK on the evaluation set. For the next three domains, consistent DER reductions have been achieved on both development and evaluation sets. For the CTS domain, the ratio of the overlapping speech increases when compared with the other former domains according to Table 1. Therefore the improvement brought by NSD is more significant, achieving an absolute reduction of 5.48%, or a 33.8% relative reduction from the scenario-independent CSD system on the development set. Owing to the AMI training data, ND also brings a performance boost to the MEETING domain with a high overlap ratio.

Table 3: *DER(%) performance comparison on full development set and full evaluation set of eight domains data among different systems.*

Domain	CSD		CSD + NSD	
	Dev	Eval	Dev	Eval
BROADCAST.	2.60	4.22	2.37	4.46
COURTROOM	2.95	3.07	2.46	3.07
MAP TASK	5.02	3.41	2.27	3.20
CLINICAL	10.97	11.09	9.83	10.03
SOC.LAB	7.97	6.04	5.17	3.81
SOC.FIELD	11.87	8.05	10.74	7.10
CTS	16.22	14.19	10.74	9.71
MEETING	26.41	33.20	23.05	28.17

4.4. Factors Affecting SSD Performances

The CTS data occupies a quite large proportion of the full dataset, so we pay more attention to this domain. We find that further improvement can be achieved by using SSD first, the performance is listed in Table 4. The telephone recording environment contains little noise, so SSD performs well on CTS domain. But the performance improvement on the other two-speaker mixed domains with some background noise brought by SSD is limited. The SSD brings a DER reduction of 7.46%, and NSD can further bring a 1% reduction, totally 52.2% relative reduction on the development set. DER also reduces from 10.74% to 7.76% compared with the combination of CSD and NSD in Table 3. A huge reduction is also shown on the evaluation set. The performance improvement on CTS really contributes to the significant improvement on the full data set.

Table 4: *DER(%) performance comparison on full development set and full evaluation set of CTS data among different systems.*

Domain	CSD		CSD + SSD		CSD + SSD + NSD	
	Dev	Eval	Dev	Eval	Dev	Eval
CTS	16.22	14.19	8.76	7.85	7.76	7.03

4.5. Effect of Post-Processing

In post-processing, we also fuse CSD systems with different AHC and VB-HMM thresholds for RESTAURANT and WEB VIDEO domains. The best single-system results before post-processing can be obtained in Table 2-4. And we get the final results in Table 5. The scenario-independent system refers to CSD, and the scenario-dependent system refers to our proposed system after post-processing. Almost all of the domains yield better results after post-processing, and we finally obtain the DER of 11.07% on development set and 11.30% on the evaluation set, with great relative reductions of 32.17% and 28.34%, respectively, when compared with results obtained with the powerful scenario-independent system.

Table 5: *DER(%) performance comparison on full development set and full evaluation set among different systems.*

Domain	Scenario-Independent		Scenario-Dependent	
	Dev	Eval	Dev	Eval
AUDIOBOOKS	2.37	0.43	0.00	0.00
BROADCAST.	2.60	4.22	2.15	4.18
COURTROOM	2.95	3.07	1.31	2.82
MAP TASK	5.02	3.41	1.35	1.58
CLINICAL	10.97	11.09	8.71	8.46
SOC.LAB	7.97	6.04	4.31	3.70
SOC.FIELD	11.87	8.05	9.40	6.45
CTS	16.22	14.19	7.50	6.55
MEETING	26.41	33.20	21.63	24.53
WEB VIDEO	35.02	37.30	33.25	34.33
RESTAURANT	43.82	43.20	37.85	38.29
Ave.	16.32	15.77	11.07	11.30

5. Conclusions

In this paper, we propose a scenario-dependent speaker diarization framework to deal with the diverse scenarios encountered in DIHARD-III challenge. Results show that our proposed system performs better than those obtained with the conventional scenario-independent systems even with potential domain classification errors. Our system ranked first among all the submitted systems, in both evaluation tracks and on both evaluation data sets. In future work, we plan to pay more research attention to those scenarios with higher speech overlap ratios.

6. Acknowledgements

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.

7. References

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

[2] 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, 2012.

[3] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[4] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "The stc system for the chime-6 challenge," in *CHiME 2020 Workshop on Speech Processing in Everyday Environments*, 2020.

[5] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[6] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *ICASSP*, 2017, pp. 4930–4934.

[7] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *ICASSP*, 2018, pp. 5239–5243.

[8] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *INTERSPEECH*, 2013, pp. 1477–1481.

[9] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial neural network features for speaker diarization," in *SLT*, 2014, pp. 402–406.

[10] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," *2018, tech. Rep.*, 2018.

[11] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, "Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge." in *INTERSPEECH*, 2018, pp. 2808–2812.

[12] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolíková, O. Novotnỳ, K. Veselỳ, O. Glembek, O. Plchot *et al.*, "But system for dihard speech diarization challenge 2018." in *INTERSPEECH*, 2018, pp. 2798–2802.

[13] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "Second dihard challenge evaluation plan," *Linguistic Data Consortium, Tech. Rep*, 2019.

[14] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotnỳ *et al.*, "But system for the second dihard speech diarization challenge," in *ICASSP*, 2020, pp. 6529–6533.

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

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[17] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

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

[19] M. Sahidullah, J. Patino, S. Cornell, R. Yin, S. Sivasankaran, H. Bredin, P. Korshunov, A. Brutti, R. Serizel, E. Vincent *et al.*, "The speed submission to dihard ii: Contributions & lessons learned," in *INTERSPEECH*, 2019, pp. 999–1002.

[20] Z. Zajíc, M. Kunešová, M. Hrúz, and J. Vaněk, "Uwb-ntis speaker diarization system for the dihard ii 2019 challenge," in *INTERSPEECH*, 2019, pp. 993–997.

[21] 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.

[22] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[23] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. 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.

[24] 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," in *INTERSPEECH*, 2020, pp. 269–273.

[25] X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, Y. Zhao, G. Liu, J. Wu, J. Li, and Y. Gong, "Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2010.11458*, 2020.

[26] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[27] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[28] Y. Wang, M. He, S. Niu, L. Sun, T. Gao, X. Fang, J. Pan, J. Du, and C.-H. Lee, "Ustc-nelslip system description for dihard-iii challenge," *arXiv preprint arXiv:2103.10661*, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] L. Sun, J. Du, X. Zhang, T. Gao, X. Fang, and C.-H. Lee, "Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization," in *ICASSP*, 2020, pp. 7099–7103.

[31] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015, pp. 3214–3218.

[32] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[33] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[34] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," *arXiv preprint arXiv:2011.01997*, 2020.

[35] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88. Citeseer, 2005, p. 100.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.