A Deep Analysis of Speech Separation Guided Diarization Under Realistic Conditions

Xin Fang*†, Zhen-Hua Ling*, Lei Sun†, Shu-Tong Niu*, Jun Du*, Cong Liu† and Zhi-Chao Sheng†
* University of Science and Technology of China, Hefei, China
† iFlytek Research, Hefei, China

***Abstract*—Recently, with the development of voice-print technology, such as x-vectors, the performance of speaker diarization has made a great progress. However, the allocation of overlapping speech segments is still a difficult problem. At the same time, great results have been achieved in the field of speech separation, especially the end-to-end time-domain audio separation network (TasNet). Speaker diarization and speech separation have strong similarities in task definition, one is to give the existence of each speaker in the time dimension, and the other is to give separated speech signals of different speakers. In this paper, we take advantage of the complementarity between the two tasks and propose a speech separation guided diarization (SSGD) approach. To our knowledge, this is the first deep analysis about combining both speaker diarization and speech separation methods. Moreover, we compare the architectures of various common speech separation models, and analyze the robustness and generalization ability of the proposed method. By incorporating this method, the overall system achieved the first place among all submitted systems in the DIHARD-III challenge.**

I. Introduction

The task of speaker diarization is to segment speaker homogeneous regions given an arbitrary audio recording [1]. An ideal speaker diarization system should work without any prior information, such as the number of speakers, the dialog styles and environmental factors. To better evaluate the performance and generalization of current diarization methods, the DIHARD challenge [2], [3], [4] was proposed where the datasets are well-designed and drawn from a diverse set of challenging domains. The DIHARD dataset includes real-world application scenarios, including free-form conversation styles, noisy environments, various speaker numbers, and so on. Generally, it places very strict requirements for the state-of-the-art speaker diarization systems.

In recent years, the technology of speaker diarization has become gradually mature under realistic conditions. For the front-end processing, previous study [5] has shown that deep learning based denoising method has stronger potentials in coping with realistic noisy environments than traditional enhancement approaches. For the back-end system, the bottom-up clustering based methods still take the majority. Powerful speaker representations are extracted and prepared for clustering, such as i-vectors [6], d-vectors [7] and x-vectors [8], [9], [10] and etc. Recently, variational Bayesian hidden Markov model was introduced with x-vectors (VBx) [11] , and achieved the best performance in the DIHARD-II challenge [3], [12].

However, conventional diarization techniques still cannot well handle overlapping speech. There are many reasons, for example, the speaker representations are trained with single speaker speech and cannot generalize to multi-speaker data. And the detection performance of overlapping speech under realistic conditions yields little improvements. It can be observed that almost all top teams [12], [13], [14] in DIHARD-II challenge did not reduce the missed speech error rate in Track 1 where all missed speech came only from overlapping speech.

Speech separation is one of the most straight-forward ways to deal with overlapping speech. Most speech separation approaches have been formulated in the time-frequency (T-F) domain. Such as models include feed-forward neural networks [15], recurrent neural networks (RNNs) [16], [17], and generative adversarial networks (GANs) [18]. Recently, time-domain based networks, such as fully-convolutional time-domain audio separation network (Conv-TasNet) [19] and dual-path RNN TasNet[20], have shown good results in speech separation. To adopt speech separation techniques in speaker diarization task, it can be used to separate source signals from different speakers.

In this study[1], we will introduce the proposed separation guided diarization (SSGD) method, which was also mentioned in the best system in DHARD-III[21], [22]. In Section IV, besides general concepts, we will give deep analysis about the robustness and universality of the proposed method. By incorporating this method, the overall system achieved the first place among all submitted systems in the DIHARD-III challenge. In Section V and VI, we will discuss some unsolved problems and the future work.

II. Task Overview

To connect the two tasks, we firstly describe their definitions, purposes, and current optimal systems.

A. Speaker Diarization

Most current speaker diarization systems use the bottom-up approach, which adopts agglomerative hierarchical clustering (AHC) as the main part. The whole recording is first divided into smaller segments where each segment ideally comes from only one speaker. Speaker representation features

[1]This work was supported in part by the National Natural Science Foundation of China under Grants No.62171427 and the National Key Research and Development Program of China (2020AAA0103600).

can be extracted on each short segment, such as traditional acoustic features [1], i-vectors [23], [24], x-vectors [8] and etc. Similar segments are clustered iteratively according to some similarity measures, including Bayesian Information Criterion (BIC) [25], probabilistic linear discriminant analysis (PLDA) [26], [27], cosine distance and so on. With the development of deep learning based speaker representations, x-vectors have shown superior performance in speaker recognition and diarization. Based on it, researchers in BUT introduced Variational Bayes Hidden Markov Model (VB-HMM) [11] to further refine the assignment of x-vectors to speaker clusters, and got the best performance in DIHARD-II challenge[12]. Besides, overlapped speech detection was often used to assign the same speech region to different speakers [28], [29]

B. Speech Separation

Supervised speech separation approaches have shown great abilities. Most previous studies have explored it in time-frequency domain [30], [31]. When separating multiple sources, the permutation problem [32] causes great difficulty in model training. Accordingly, permutation invariant training (PIT)-based methods [32] were proposed to address the such permutation problem.

Since end-to-end method has become one of the mainstream approaches, lots of researches focus on time-domain speech separation. In [33], researchers introduced an encoder-decoder framework, namely time-domain audio separation network (TasNet), to directly modeled the mixture waveform which totally replaced traditional time-frequency manipulations. TasNet based methods [19] not only achieved better performance than Ideal Ratio Mask (IRM), but also improved the model size and computation cost. In [34], dual-path recurrent neural network (DPRNN) was used to model extremely long sequences without increasing model size. Since then, more and more approaches have emerged following this encoder-decoder framework, such as DPTNet [35], SepFormer [36] and etc.

Typically, a time domain end-to-end speech separation model can be optimized directly by scale-invariant source-to-noise ratio (Si-SNR):

$$\text{Si-SNR} = 10\log_{10}\frac{||\mathrm{s}_{\text{target}}||^2}{||\hat{\mathrm{s}} - \mathrm{s}_{\text{target}}||^2} \tag{1}$$

where $\mathrm{s}_{\text{target}} = \frac{<\hat{\mathrm{s}},\mathrm{s}>\mathrm{s}}{||\mathrm{s}||^2}$. $\hat{\mathrm{s}}$ and s are the estimates and targets respectively. For multiple outputs, the overall loss is often formed using permutation invariant training (PIT) based learning objective as:

$$L = \frac{1}{N}\sum_{i=1}^{N} l(\hat{s}_i - s_\phi) \tag{2}$$

where N is the number of speakers, l is the error between the network output and the target, $\hat{s}_i$ denotes the i-th predicted speech, s_ϕ denotes the reference speech with the permutation ϕ that minimizes training objective L.

III. The proposed speech separation guided diarization (SSGD) system

By definition, the two tasks are highly complementary. Speaker diarization is to give the existence of each speaker in the time dimension, while speech separation yields separated speech signals of different speakers. As illustrated in Figure 1, we show how the two tasks can help each other in a conversation audio. The speaker number is set to 2.

Given a dialogue:

1) A conventional speaker diarization (CSD) system includes speech segmentation, x-vector extraction, AHC, and VB-HMM resegmentation. The generated speaker distribution shows regions where each speaker speaks, but does not include overlapping part.

2) A speech separation (SS) system can process the audio directly and output different channels, each representing an unique speaker identity. Using the VAD system, speaker distribution can also be generated, including information about overlapping part.

3) The above two systems have their own pros and cons. We propose a speech separation guided diarization (SSGD) system to combine them together. The first CSD system performs more stably, but is unable to process overlapping data. The second SS system can process overlapping speech, but the performance fluctuates greatly, especially under realistic conditions. To combine those advantages, we difine a relative DER between two system results, measuring the degree of deviation between two systems:

$$DER_{Relative} = \frac{\sum_{s=1}^{S} d(s)\cdot(\max(K_{CSD}(s), K_{SS}(s)) - K(s))}{\sum_{s=1}^{S} d(s)\cdot K_{CSD}(s)} \tag{3}$$

where S is the number of speaker segments in which both CSD results and SS results contain the same speaker (or speakers), while $d(s)$ correponds to the duration of a single segment s. $K_{CSD}(s)$ and $K_{SS}(s)$ denote the speaker number in speech segment s of CSD and SS results respectively. $K(s)$ means the number of speakers in speech segment s that are correctly matched between CSD and SS results. This formula is equivalent to DER [37] metric, the only difference is the group truth is set to CSD results. If the value of relative DER is smaller than a pre-defined threshod, SS results will be selected. Otherwise, the selection is reversed. According to our results, the proposed selection strategy can effectively detect outrageous SS results on realistic data where the speech sepration module fails.

4) To further enhance overall performance, the pre-trained speech separation model in SSGD system can be fine-tuned using speaker priors derived from CSD system. In this way, the above selection strategy can be dispensed, and the results from fine-tuned speech separation system can be used directly.

IV. Experiments And Analysis

In this paper, we focus on two-speaker conversations, because speaker diarization and speech separation techniques are

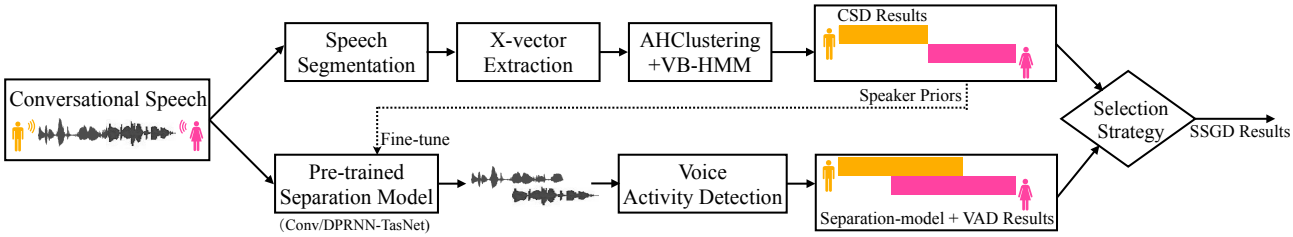

Fig. 1. The overall diagram of the speech separation guided speaker diarization.

well established for this kind of condition. For evaluation data, we select the realistic conversational telephone speech (CTS) dataset from both development set and evaluation set of the DIHARD-III challenge [4], containing a total of 61 items. Each utterance is spoken by two native English speakers and lasts about 10 minutes. Statistically, the overall overlap ratio is about 12%.

The conventional speaker diarization (CSD) system follows both the official baseline [4] and the VBx system published in [38]. The CSD results are taken as the baseline which represents the most mainstream system performance. Diarization error rate (DER) [37] was used as evaluation metric in our experiments, which consists of three parts, namely speaker error, false alarm error and missed error. Oracle VAD was used in our experiments, and no forgiveness collar was used during evaluation.

For speech separation model training, we used the asteroid toolkit [39] which brought lots of model variants. We used Librispeech [40] dataset to simulate speaker mixtures. As for VAD part, we chose WebRTC VAD module [41] to process the separated audio data. In selection strategy, the threshold in Eq.3 was set to 25 %.

A. The Effects of Speech Separation Model Architecture

In the field of speech separation, methods are evaluated on simulated data, such as the commonly used WSJ0-2Mix dataset [30]. As reported in [42], [34], the performance was measured by SI-SNR improvement (SI-SNRi). The numbers are listed in Table-I, all TasNet-based models are better than IRM, which means the separated audio is basically clean and complete.

Separation Methods	Si-SDRi
IRM	12.2
BLSTM-TasNet	13.2
Conv-TasNet	15.3
DPRNN-TasNet	18.8

TABLE I
THE PERFORMANCE COMPARISON OF DIFFERENT SPEECH SEPARATION METHODS ON WSJ0-2MIX DATASET[34], [42].

To migrate such conclusion in speaker diarization, we used two representative model architectures, namely Conv-TasNet and DPRNN-TasNet. We simulated 250 hours speech mixtures to pre-train separation models. During training, all audios were grouped into 3-second long segments, and the batch-size was set to 6.

When testing, the entire utterance was directly processed by the pre-trained model, generating two output channels. We used WebRTC VAD to detect speech segments in each channel. Then, all segments and their channel properties were combined together to produce an entire RTTM file, namely SSGD result. Compared with the CSD result, a final rttm was determined by the procedures described in Section III.

Method	Miss	FA	Spk_Err	DER
CSD-Baseline	12.0	0.0	4.2	16.22
SSGD(Conv-TasNet)	7.60	2.6	2.7	12.95
SSGD(DPRNN-TasNet)	11.4	0.4	4.1	15.91

TABLE II
THE DIARIZATION RESULTS COMPARISON OF DIFFERENT METHODS. ORACLE VAD WAS USED.

As shown in Table II, the following points can be observed. First, when using oracle VAD, the main component of DER in CSD results comes from missed speech. Second, because the separation models separate two speakers, the SSGD methods can decrease the missed speech error. Correspondingly, the false alarm error increases because many regions are misclassified to overlapping speech. Third, relative performance difference between different models in speech separation does not keep the same in speaker diarization task. For example, DPRNN-TasNet outperforms Conv-TasNet with a obvious margin on simulated data in terms of SI-SNRi, as listed in Table I. However, the conclusion is the exact opposite in Table II. It indicates that there are huge mismatches between simulated data and realistic data. DPRNN-TasNet has stronger modeling capabilities than Conv-TasNet and produces much better metrics on a limited-size dataset. But it does not guarantee better performance under realistic conditions, and may be limited to over-fitting. Finally, the SSGD system based on Conv-TasNet yields a relative 20.16% reduction of DER, from 16.22 to 12.95.

B. The Effects of The Proposed Selection Strategy

Ideally in two-speaker situations, a speech separation model with a precise VAD module can be a perfect replacement of a conventional speaker diarization system. Currently, the generalization ability of separation models can not be fully guaranteed for realistic speech dialogues. It's the essential

motivation why we design a selection strategy to eliminate those outrageous cases produced by SSGD system. With the selection strategy, we group all utterances into two categories in Table III.

Model	Selected(33/61)	Others(28/61)
CSD-Baseline	16.26	16.12
Conv-TasNet+VAD	10.30	33.47
Conv-TasNet+Fine-tuning+VAD	9.47	9.84

TABLE III
THE PERFORMANCE COMPARISON BETWEEN TWO GROUPS WHICH ARE DETERMINED BY THE PROPOSED SELECTION STRATEGY.

It can be seen that without the selection strategy, pure results from Conv-TasNet+VAD vary a lot on different testing cases, while the CSD-baseline maintains a stable performance in both groups. It's largely due to the pre-trained speech separation model which lacks great stability. To fully utilize all information, we used the speaker information generated by the CSD system as speaker priors to fine-tune the speech separation model. We randomly made 5,000 speech mixtures by using pseudo speaker labels from the CSD system, and conducted a very lightweight fine-tuning on the pre-trained model. As shown in the bottom row in Table III, all results improves a lot, especially for those outrageous cases, from 33.27 to 9.84. It demonstrates that speech separation is a good solution for the speaker diarization task on overlapping speech data, inferring that its generalization ability on different speakers can be improved.

C. The Effects of Training Data

In this part, we explore the effects of training data in our proposed SSGD diarization system. As listed in Table IV, we use three different data size, including 50 hours, 100 hours and 250 hours. It's clear that increasing training data size gradually improves the overall performance.

It's important to note that all simulated data are produced with a 100% overlap ratio which is not always compatible to real conversations. We made additive sparse mixtures for several different amounts of speech overlap: from 0% to 80%, following the method in [43]. The total amount of these sparse data is about 160 hours. After adding them into training, the DER slightly improves from 12.95 to 12.81.

Method	Training Data	Miss	FA	Spk_Err	DER
SSGD (Conv-TasNet)	50h	9.1	2.0	3.2	14.28
	100h	8.7	1.8	3.1	13.68
	250h	7.6	2.6	2.7	12.95
	250h +sparse data 160h	7.7	2.4	2.7	12.81

TABLE IV
THE EFFECTS OF TRAINING DATA IN SPEECH SEPARATION.

V. DISCUSSIONS

Through utilizing the complementarities between speech diarization and speaker diarization, we propose a speech separation guided diarization system which shows great performance in terms of DER reduction. However, there are still some unsolved problems worth discussing.

A. Relationship with Other Methods

A similar analysis was carried out among speech separation, diarization, and recognition in [44], here we only focused on two of them and built a practical system. In essence, the SSGD diarization system uses a VAD module to produce final results, though the front separation model is trained by end-to-end. Other methods like end-to-end neural speaker diarization (EEND) based approaches [45], [46] neglect the explicit VAD process by predicting classification directly. These methods can benefit from the direct optimization of the final DER metric, and also achieve great results. Compared with them, the advantage of SSGD system is that the separated audios can also be used for an ASR task like CHiME [47]. In our view, joint optimization of the regression (separation) task and the classification task will be a promising direction in the future.

B. The Problem of Generalization Ability

At present, there are still many problems about the robustness of the proposed algorithm, mainly including the characteristics of speakers, the number of speakers, environmental interference factors and so on. We have demonstrated in Section IV-B that lightweight fine-tuning is very effective in alleviating the mismatches. In the future, the following directions can be considered as having potentials to solve these problems, including more realistic data simulation, more powerful model architecture design, multi-task training and so on.

VI. CONCLUSIONS

In this paper, we demonstrate the effectiveness of the proposed speech separation guided diarization system, which effectively reduce the missed speech error rate on the realistic DIHARD-III dataset. Moreover, we conduct a detailed analysis of several aspects of the proposed method, including model architecture, simulated training data and the generalization ability. In the future, we will explore combining this method with the EEND based methods, and include ASR task as a part of overall evaluation.

REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *arXiv preprint arXiv:2101.09624*, 2021.
[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," in *https://zenodo.org/record/1199638*, 2018.
[3] ——, "Second dihard challenge evaluation plan," *Linguistic Data Consortium, Tech. Rep*, 2019.
[4] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.
[5] L. Sun, J. Du *et al.*, "A novel lstm-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *ICASSP*, 2018.
[6] 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.

[7] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

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

[9] X. Fang, T. Gao, L. Zou, and Z. Ling, "Bidirectional attention for text-dependent speaker verification," *Sensors*, vol. 20, no. 23, p. 6784, 2020.

[10] X. Fang, L. Zou, J. Li, L. Sun, and Z.-H. Ling, "Channel adversarial training for cross-channel text-independent speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6221–6225.

[11] M. Diez, L. Burget, F. Landini, and J. Černockỳ, "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 355–368, 2019.

[12] 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 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 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," *arXiv preprint arXiv:2002.12761*, 2020.

[14] 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 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7099–7103.

[15] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[16] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[17] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[18] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 711–715.

[19] 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.

[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[21] 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.

[22] Y. Wang, J. Du, M. He, S. Niu, L. Sun, and C.-H. Lee, "Scenario-dependent speaker diarization for dihard-iii challenge," *Accepted to Interspeech*, 2021.

[23] 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, 2011.

[24] S. Madikeri, I. Himawan, P. Motlicek, and M. Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," Idiap, Tech. Rep., 2015.

[25] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, 1978.

[26] 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.

[27] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," *CRIM, Montreal, Technical Report*, 2008.

[28] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4353–4356.

[29] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7114–7118.

[30] J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[31] C. Weng, D. Yu, M. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1670–1679, 2015.

[32] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[33] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[34] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[35] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[36] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[37] "The 2009 (rt-09) rich transcription meeting recognition evaluation plan," *http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf*, 2009.

[38] "Vbx toolkits," *https://github.com/BUTSpeechFIT/VBx/tree/v1.0*, Brno University of Technology, 2020.

[39] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas *et al.*, "Asteroid: the pytorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[41] "Google webrtc," *https://webrtc.org/*, 2016.

[42] 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.

[43] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[44] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 897–904.

[45] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[46] 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 preprint arXiv:2003.02966*, 2020.

[47] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.