Peer Learning via Shared Speech Representation Prediction for Target Speech Separation

Xusheng Yang, Zifeng Zhao and Yuexian Zou*
School of Electronic and Computer Engineering, Peking University, Shenzhen, 518055, China
E-mail: yangxs@stu.pku.edu.cn, zifeng_z@foxmail.com, zouyx@pku.edu.cn

***Abstract*—Target speech separation approaches widely use the encoder-separation-decoder framework. Generally, the crucial idea of the framework is how to map the mixture of speech signals to a distinctive representation. However, existing methods fail to learn discriminative speech representations because they ignore the distinctive characteristics (tone and timbre) related to each target speaker, which are shared within the corresponding mixture samples. To acquire shared speaker characteristics and discriminative representations, we propose a novel "peer learning" method for this framework. Specifically, we construct a pair of mixture samples involving the same target speaker and design a two-branch speech representation prediction module. Then, the more discriminative representation can be determined via computing the signal-to-noise ratio and exchanged in the prediction module to improve separation performance. Experiments on the Libri2Mix dataset demonstrate the effectiveness of our method.**

I. Introduction

Target speech separation (TSS) is to separate the voice of a speaker of interest from an overlapped mixture of speech signals. It can be applied to hearing aids, mobile communication, and speech recognition systems [1]. Humans can naturally extract relevant information from the target speaker in noisy environments. Researchers have identified the importance of essential cues to a target speaker [2], [3], so the popular approaches usually transform the mixture into a separable space leveraging acoustic, visual and spatial information [4]–[7]. Recently, TSS methods of encoder-separation-decoder have gained more attention. Their separation performance heavily relies on mapping the mixture of speech signals to a distinctive representation, which is a still challenging research problem.

For a specific target speaker, the information in the different mixture signals (tone and timbre) helps learn the distinctive representations. By analyzing these characteristics, the model can learn to recognize the specific frequency patterns and harmonic structures that are unique to the target speaker. Recent works [8]–[10] on target speech separation retain the temporal architecture [11], [12] of the encoder-separation-decoder (speech extraction network) and incorporate a convolutional network (auxiliary network) for extracting the information of target speaker. The speech extraction network first encodes the mixture speech of signals by a 1-D convolutional layer, and subsequently separates and decodes them by several convolutional blocks and a decoder layer in representation space. Finally, it outputs a single signal corresponding to the target speech, effectively isolating a clean signal from the mixture.

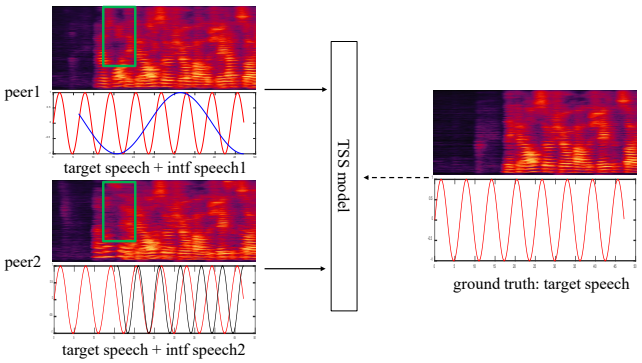

Fig. 1. The intuition of peer learning samples. The red sinusoidal signal represents the target speech, while the blue signal and black signal are interference speech at different frequencies. Peer1 represents a mixture of red sinusoidal signal and blue signal. Peer2 represents a mixture of red sinusoidal signal and black signal. The green-boxed regions indicate harmonic structures and frequency patterns where two samples can mutually enhance each other.

However, when speakers have similar voice characteristics, existing methods become challenging to identify a specific target speaker from a mixture of signals. These methods use each mixture sample of the target speaker in the training independently, so they cannot share enough useful characteristics on speech qualities (tone, timbre, and pitch) resulting in non-discriminative representations. Moreover, the conventional encoder-separation-decoder framework, while adept at distinguishing the target speech from a mixture of signals by estimating masks, is inherently constrained in capturing shared speech representations across diverse mixture samples. This limitation stems from the architecture's primary focus on differentiating the target signal rather than extracting commonalities that may exist within the mixture.

In this work, we introduce a novel learning paradigm termed "peer learning" to extract shared attributes and learn discriminative representations from diverse speech mixtures. Our approach is underpinned by the creation of a novel mixing dataset that pairs mixture samples centered around the target speaker. Our core concept is to explore the potential for two mixture samples to engage in a reciprocal learning process, akin to peers sharing insights, hence the term "peer learning" samples. As depicted in Figure 1, we present two mixture

*Corresponding author.

samples, peer1 and peer2, each a blend of the target speaker's speech with that of a different, non-target speaker or interfering voice (for instance, peer1 is a mix of Speaker 1 and Speaker 2, while peer2 combines Speaker 1 with Speaker 3). Notably, we propose that the harmonic structures and frequency patterns within the green-boxed regions of these mixture samples can be interchangeably leveraged to bolster the accuracy of target speech prediction. To optimize the utilization of "peer learning" samples, we have engineered a dual-branch architecture for speech representation prediction designed to impart distinctive characteristics. In this bifurcated approach, we derive predicted speech via both the representation prediction module and the decoder, subsequently quantifying the signal-to-noise ratio (SNR) based on the predicted speech and the actual target speech. The representation associated with a superior SNR is selected, as it is indicative of a higher-quality speech representation. This enhanced representation is then fed back into the other branch to refine the original, less discriminative representation. Through this mechanism, we facilitate the exchange of discriminative representations between branches, thereby enhancing the overall speech separation performance. Experimental results show that our peer learning framework outperforms the performance of baselines for the target speech separation on the Libri2Mix dataset [13].

II. Related Work

Target speech separation. TD-SpeakerBeam [8] acquires more discriminative speaker embedding vectors using an auxiliary convolutional network and a multi-task loss with speaker identification loss. It follows the Conv-TasNet to construct multiple convolutional filters to transform the temporal signals within a time slot into a learnable representation. SpEx+ [9] shares the same weight in the latent domain by twin speech encoder, importantly it encodes multi-scale information in a uniform latent feature space. Besides, the refining neural network [14] demonstrates that high-order embedding space can leverage the discriminative representation for speech separation. However, in a mixture speech, the same target speaker can be mixed with any speaker. These works do not consider shared knowledge such as different frequency patterns and harmonic structures by constructing richer training pairs.

The paradigms of learning. Self-supervised learning [15]–[20] constructs a pair of positive and negative samples by selecting one speaker's utterance as the anchor, and selecting speech from the same person but a different utterance as the interfering speech. To capture speech characteristics, they introduce a speaker consistency loss on the speech embeddings. The teacher-student network [21], [22] can unidirectionally transfer knowledge from the teacher model to the student model. Mutual learning [23], [24] approaches can use a group of students to learn and share knowledge simultaneously with each other during the training phase, but they require the selection of two models when the testing phase and the parameter of the threshold must be manually set. However, the mainstream frameworks focus on masking interference speech while ignoring learning the shared presentations to improve separation performance.

III. Methods

In this section, we first give a notation of target speech separation. We adopt the mainstream framework of encoder-separator-decoder as the backbone and introduce the proposed peer learning (PL) method to obtain shared attributes and learn discriminative representations. Fig.2 shows the overview of our PL method. Then we explain the speech representation prediction mechanism in detail.

A. Problem Formulation

Notations. Target speech separation is to isolate the speech of a target speaker from a mixture of multiple overlapping speakers and optionally an additional noise. First, X denotes the mixed speech signals, and S is reference clean sources. The adaptation utterance about the target speaker will be denoted r. We aim to separate the predicting source $\hat{S}$ for each speaker from X by leveraging the target speaker information r. To make it simple, a two-speaker setup is considered.

B. Peer Learning Method

Some advanced learning methods have provided performance improvements for target speech separation. However, these methods have certain limitations. These works ignore more discriminative representation can be beneficial to separation performance further. The works mentioned above do not leverage shared harmonic structures and frequency patterns in a pair of mixture samples of the same target speaker. Our intuition, as previously highlighted, posits that the introduction of a pair of mixture samples pertinent to the target speaker can unlock superior representations. This strategy is predicated on the notion that by harnessing the intrinsic relationships between samples, we can distill representations that are more conducive to effective speech separation.

Constructing "peer learning" samples. We construct pairs of training data that contain a mixture sample of the same target speaker and different interference speakers. Generally, there are n speakers. Suppose we have one speech segments of target speaker $s_i(t)$ for speaker i, two speech segments of interference speaker $s_j(t)$ and $s_k(t)$ for speaker j and k. Note that j and k may be the same speaker. We create two mixture signals $x_1(t)$ for speaker i and speaker j, $x_2(t)$ for speaker i and speaker k, which is defined in:

$$\begin{aligned} x_1(t) &= s_i(t) + s_j(t), (i \neq j, i, j \in n) \\ x_2(t) &= s_i(t) + s_k(t), (i \neq k, i, k \in n) \end{aligned} \tag{1}$$

Shared two-branch Framework. We adopt the general framework of neural target speech separation, mainly including a mixture encoder, a fusion layer, and a target extractor. Specifically, the peer mixture speech signals are then transformed into the embedding space. Then we use the fusion layer to introduce the information of the target speaker. Finally, we use the extractor module to estimate target speech. The entire pipeline of peer learning with shared speech representation

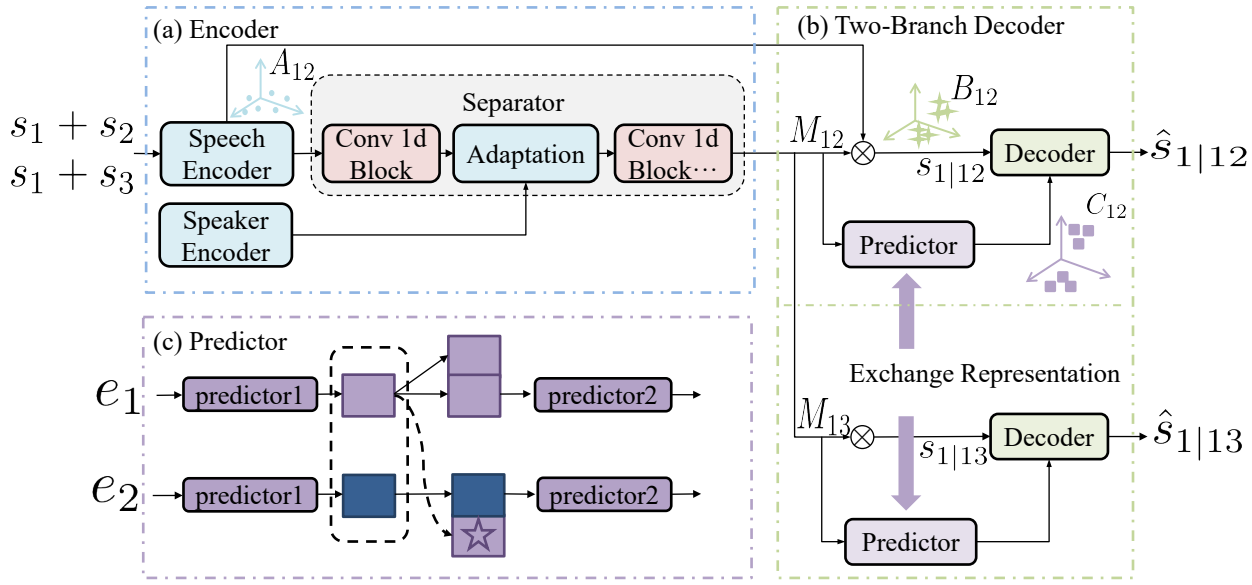

Fig. 2. The pipeline of peer learning via shared speech repressentation prediction. The purple bidirectional arrow illustrates the sharing discriminative representation in this space. (a) The encoder of our method with shared weights for "peer learning" samples (different mixture samples of the same target speaker). (b) The two-branch decoder for "peer learning" samples. (c) The detail of the predictor (the purple block).

prediction is as shown in Fig.2, which is designed based on the famous framework of TD-speakerBeam [8]. The proposed two-branch network employs an encoder with shared weights, indicating that identical network structures are deployed to handle different input datasets. This approach ensures that the same underlying model processes various inputs, promoting efficiency across the network's operations. The speech encoder denotes extracting the mixture representations by 1-D convolutional filters, speaker encoder uses the same 1-D convolutional encoder to accept the enrollment utterance of the target speaker. Then the separator transforms further the representation into the separable representation space incorporating the adaptation layer and the stacked 1-D convolutional blocks. They perform the following equations:

$$\begin{aligned} X_{emb} &= \text{Enc}(x; \theta^{enc}) \\ Spk_{emb} &= \text{SpkEnc}(r; \theta^{Spkenc}) \\ E_{emb} &= \text{Sep}(X_{emb}, Spk_{emb}; \theta^{Sep}) \end{aligned} \tag{2}$$

where Enc and SpkEnc respectively represent the convolutional network and each network with parameters θ^{Spkenc} and θ^{Spkenc}. For $X_{emb} \in R^{D \times L}$, D denotes dimension, L is the encoded temporal length. And $E_{emb} \in R^{D^s \times L}$ is the output of separator Sep in gray modules in Fig.2. Then, we define the mask branch of the representations in green space B, and the purple space C denotes the predictor branch of the representations. We discuss the details of the prediction branch in the following paragraphs.

Shared Speech Representation Prediction. Motivated by self-supervised learning methods BYOL [25], this work employs two different networks and adopts a stop-gradient strategy to introduce asymmetry for representation learning. Therefore, we also attempt to introduce asymmetry in the two-branch network to obtain discriminative representations. From this perspective, the reason why directly adding a speech representation prediction module is that the better representation can guide the worse one, which plays a role in improving the worse representation. We directly add a speech representation prediction module to guide the original representation with the better one. For a clearer presentation, suppose we have one speech segment of target speaker $s_1(t)$ for speaker 1, two speech segments of interference speaker $s_2(t)$ and $s_3(t)$ for speaker 2 and 3. We create two peer mixture signals: $x_1(t) = s_1(t) + s_2(t)$ and $x_2(t) = s_1(t) + s_3(t)$. We have obtained the E_{emb} from the shared two-branch network, at this point, so we denote them e_1 and e_2 for two peer learning mixtures. After the separator module, we generate a speech representation using a prediction network with the convolutional layer. Note that this representation is near to the ground truth of the target speech before the decoder, so the discriminative speech representation can be shared.

More discriminative representations can be understood as features that, after being mapped by the predictor, more closely align with the actual target speech. This enhanced alignment signifies a higher fidelity in the representation, enabling more accurate speech processing. Fig.2 (b) illustrates a predictor that introduces a novel mapping of the representation prior to the decoder. Since the representation at this stage is already quite close to the actual speech, it can be leveraged as new knowledge to share. In other words, a more discriminative speech representation is effectively evaluated through this comparison, highlighting the enhanced fidelity of the predictor's output, as shown in Fig.2 (c). The structure of the predictor is the linear neural network. The predicted output is then fed into

the decoder as input and compared with the true labels to calculate the SNR, which serves as a metric for assessing the quality of the speech representation.

We elaborate in detail on the speech representation prediction module in Fig.3. We first build a prediction branch to generate the representations $e_{11}, e_{22} = \mathrm{Pred}_1(\mathrm{e}_1), \mathrm{Pred}_1(\mathrm{e}_2)$. Further, we decide on the more discriminative representation by calculating SI-SNR with the target speech. Specifically, the representation through predictor 1 and input the decoder to acquire the predicted speech. We compare the SNR values of the two representations by predicted speech and the target speech. In the Fig.3, when SI-SNR1 is bigger than SI-SNR2, the better representation e_{11} with star marker is used to formulate for the other branch. Moreover, we concatenate the star marker representation with the original blue representation, after that, we can acquire the $e_1^{'}$, $e_2^{'}$ using a predictor 2.

$$e_1^{'}, e_2^{'} = \mathrm{Pred}_2(\mathrm{e}_{11}, \mathrm{e}_{11}), \mathrm{Pred}_2(\mathrm{e}_{22}, \mathrm{e}_{11}) \tag{3}$$

Then, we integrate them into the mask branch for recovering target speech combining with e_1 and e_2. In such a way, it can be viewed as a speech representation augmentation.

Training and Inference. During the training, we easily construct peer learning samples of the same target speaker. However, in the validation and inference, it is usually not feasible to create these samples. Therefore, during the phase of inference, we directly concatenate the mixture's representation itself. It means that we directly double the representation after predictor 1 to match the dimension.

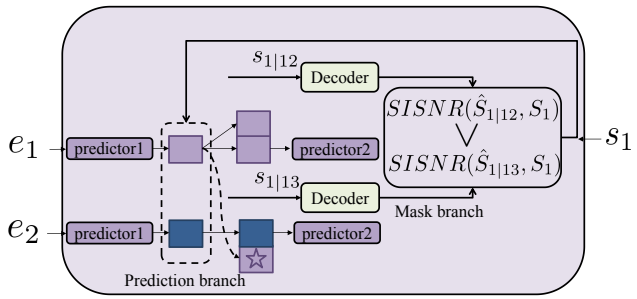

Fig. 3. The mechanism of the speech representations prediction by comparing the SNR.

IV. Experiments

A. Experimental Setup

Datasets. The Libri2Mix[13] dataset extends the utility of the LibriSpeech corpus into the domain of speech separation. Libri2Mix dataset [13] is introduced to address the generalization issue in WSJ0-2mix dataset[26]. The train-100 subset total has 13900 utterances with 58 hours of data from 251 speakers. The development and test subsets have 3,000 mixture utterances with 11 hours of data respectively, which is from 40 unseen speakers. The sampling rate of speech audio is $8\,\mathrm{kHz}$. We generate all mixtures using 'minimum' mode. We only construct peer samples in training data. During the mixing of data, the specific details have been thoroughly explained in the section III-A. The first speaker is chosen as the target speaker, the second speaker is regarded as the interference speaker. The reference speech of the target speaker must be a randomly chosen utterance distinct from the one in the mixture. The other peer sample has the same first speaker as mixtures and chooses a different utterance from the interference speaker. The development and test subsets remain consistent with the original dataset setting.

Baselines and Setup. We compare it with baseline approaches of target speech separation, including sDPCNN [27], TD-SpeakerBeam [8], SpEx+ [9]. Some of the results are obtained from the reported papers. Our all hyper-parameters selections follow the original TD-SpeakerBeam. The training process uses chunks with a 3.0 seconds duration. We adopt Adam optimizer with a learning rate of 1e-3. The batch size is 8. The epoch is set to 200. Early stopping is used in the training process when the validation loss doesn't decrease in 20 consecutive epochs. A gradient clip with a maximum L2-norm is used to avoid gradient exposure.

Evaluation Metrics. We train all time-domain models using negative SISNR loss. For evaluation metrics, we employ the scale-invariant SDR (SI-SDR) and PESQ metrics. SI-SDR assesses from a signal perspective, while PESQ evaluates them from perceptual quality. The higher evaluation value represents the better quality of speech separation.

B. Results Analysis

1) Performance Comparison: We explore the performance of different target speech separation methods on the Libri2Mix test dataset, the overall results are shown in Table.I. We denote the peer learning method as PL. In the last two rows of Table.I, our approach outperforms the sDPCNN method in the time-frequency domain, as well as the TD-SpeakerBeam method in the time domain. Our approach (PL) improves the performance to 1.0 dB and 0.5 dB SI-SDRi respectively compared with TD-SpeakerBeam and SpEx+. In the meantime, we obtain 0.13 and 0.32 PESQ improvement compared with TD-SpeakerBeam and SpEx+ respectively. Furthermore, it can be concluded that our approach can be integrated into existing time-domain frameworks easily. These findings suggest that PL can augment performance by learning more discriminative representation.

TABLE I
PERFORMANCE COMPARING ON THE LIBRI2MIX DATASETS.

Methods	SI-SDR	PESQ
Mixture	0.001	1.603
sDPCCN	11.65	2.74
TD-SpeakerBeam	12.86	2.75
SpEx+	13.41	2.94
TD-SpeakerBeam + PL	**13.94**	**3.07**

2) Speech Representation Prediction Analysis: We further investigate the usage strategy of the speech representation prediction. Specifically, there are three strategies. First, **cat itself**: both branches double their representation without knowledge

sharing. Secondly, **cat anyway**: both branches use a random representation from the other. Finally, **cat better**: we utilize a better representation by speech representation prediction module. We train the three strategies using SISNR as the same loss function. It can be observed that the performance progressively is boosted as shown in Table.II, indicating the effectiveness of our strategy via the speech representation prediction module.

TABLE II
RESULTS ON THE DIFFERENT CONFIGURATIONS.

Model	Operations	SI-SDR	PESQ
Mixture	-	0.001	1.603
TD-SpeakerBeam	-	12.86	2.75
TD-SpeakerBeam + PL	cat itself	12.84	2.74
TD-SpeakerBeam + PL	cat anyway	12.89	2.77
TD-SpeakerBeam + PL	**cat better**	**13.94**	**3.07**

3) Analysis of training process: To further validate how to share speech representation, we analyze the training process of the different representations. As shown in Fig.4, after adopting the asymmetric training method of peer learning, we have 4 outputs: the orange line represents the better peer, and the gray one represents the better mask. Besides, the yellow line denotes the worse peer while the blue one is a worse mask. Firstly, for the results of the mask branch, the better peer outperforms the worse peer by around 1.2 to 1.8 dB. For the map branch, we find that the gap between the representations of the better peer and the worse peer gradually decreases during the training process, from around 1.3 dB to 0.7 dB. The green line represents a remapped result obtained by using the better representation. We can see that its performance is not only better than the mapping performance of the worse peer but also surpasses that of the better peer itself. Upon closer examination, we observed that the representations of the two peers indeed have a synergistic effect.

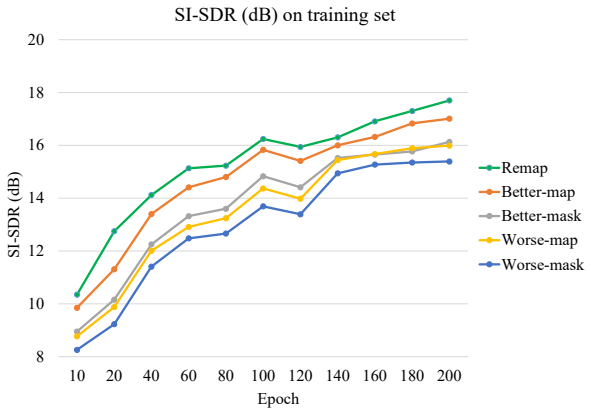

Fig. 4. SI-SDR value of through the five representations on training set in the training phase.

C. Complexity analysis

We calculate the increase in computational load brought by the additional predictive branch, as can be seen in the Table.III, there is only an increase of 0.7 M parameters, which is completely acceptable.

TABLE III
COMPLEXITY ANALYSIS ON THE DIFFERENT CONFIGURATIONS.

Model	Trainable Paras (M)	Total Paras (M)
TD-SpeakerBeam	7.1	18.3
TD-SpeakerBeam + PL	**7.8**	**19.0**

V. CONCLUSION

In this paper, we propose a "peer learning" method to acquire the shared characteristics and the discriminative representations through a two-branch speech representation prediction module. This scheme first constructs the two mixture samples related to the same target speaker for peer learning. A separable speech representation is generated by a predictor module, which is a discriminative representation for predicting the target speech waveform. Notably, the enhanced representation from one branch is leveraged to bolster the separation performance of its counterpart. We verify that the speech representations can be exchanged knowledge of each other in two branches. Extensive experiments have demonstrated the effectiveness of the proposed peer learning method. In the future, we plan to extend our methods to multi-channel and more speakers for target speech separation.

ACKNOWLEDGMENT

This paper was partially supported by NSFC (No:62176008).

REFERENCES

[1] R. Masumura, M. Ihori, A. Takashima, T. Tanaka, and T. Ashihara, "End-to-end automatic speech recognition with deep mutual learning," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 632–637.

[2] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[3] S. Kindt, J. Thienpondt, and N. Madhu, "Exploiting speaker embeddings for improved microphone clustering and speech separation in ad-hoc microphone arrays," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[4] J. Chen, W. Rao, Z. Wang, *et al.*, "Mc-spex: Towards effective speaker extraction with multi-scale interfusion and conditional speaker modulation," *arXiv preprint arXiv:2306.16250*, 2023.

[5] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "Facefilter: Audio-visual speech separation using still images," *arXiv preprint arXiv:2005.07074*, 2020.

[6] R. Gu, S.-X. Zhang, L. Chen, *et al.*, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7319–7323.

[7] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. Wang, "Multi-input multi-output complex spectral mapping for speaker separation," in *Proc. Interspeech*, 2023, pp. 1070–1074.

[8] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 691–695.

[9] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," *arXiv preprint arXiv:2005.04686*, 2020.

[10] V. A. Kalkhorani, A. Kumar, K. Tan, B. Xu, and D. Wang, "Time-domain transformer-based audiovisual speaker separation," in *Proc. Interspeech*, 2023.

[11] 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.

[12] 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.

[13] 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.

[14] Z. Yao, W. Pei, F. Chen, G. Lu, and D. Zhang, "Stepwise-refining speech separation network via fine-grained encoding in high-order latent domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 378–393, 2022.

[15] A. Mohamed, H.-y. Lee, L. Borgholt, *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[16] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*, PMLR, 2020, pp. 9929–9939.

[17] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2021, pp. 15 490–15 500.

[18] J.-B. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[19] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[20] H. Nakamura, M. Okada, and T. Taniguchi, "Representation uncertainty in self-supervised learning as variational inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 484–16 493.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] R. Aihara, T. Hanazawa, Y. Okato, G. Wichern, and J. Le Roux, "Teacher-student deep clustering for low-delay single channel speech separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 690–694.

[23] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4320–4328.

[24] H. M. Tan, D.-Q. Vu, C.-T. Lee, Y.-H. Li, and J.-C. Wang, "Selective mutual learning: An efficient approach for single channel speech separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3678–3682.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*, PMLR, 2021, pp. 12 310–12 320.

[26] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 31–35.

[27] J. Han, Y. Long, L. Burget, and J. Černockỳ, "Dpccn: Densely-connected pyramid complex convolutional network for robust speech separation and extraction," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7292–7296.