DPCCN: DENSELY-CONNECTED PYRAMID COMPLEX CONVOLUTIONAL NETWORK FOR ROBUST SPEECH SEPARATION AND EXTRACTION

Jiangyu Han[1,3], Yanhua Long[1,2], Lukáš Burget[3], Jan Černocký[3]*

[1]Shanghai Normal University, Shanghai, China
[2]Unisound AI Technology Co., Ltd., Beijing, China
[3]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

ABSTRACT

In recent years, a number of time-domain speech separation methods have been proposed. However, most of them are very sensitive to the environments and wide domain coverage tasks. In this paper, from the time-frequency domain perspective, we propose a densely-connected pyramid complex convolutional network, termed DPCCN, to improve the robustness of speech separation under complicated conditions. Furthermore, we generalize the DPCCN to target speech extraction (TSE) by integrating a new specially designed speaker encoder. Moreover, we also investigate the robustness of DPCCN to unsupervised cross-domain TSE tasks. A Mixture-Remix approach is proposed to adapt the target domain acoustic characteristics for fine-tuning the source model. We evaluate the proposed methods not only under noisy and reverberant in-domain condition, but also in clean but cross-domain conditions. Results show that for both speech separation and extraction, the DPCCN-based systems achieve significantly better performance and robustness than the currently dominating time-domain methods, especially for the cross-domain tasks. Particularly, we find that the Mixture-Remix fine-tuning with DPCCN significantly outperforms the TD-SpeakerBeam for unsupervised cross-domain TSE, with around 3.5 dB SISNR improvement on target domain test set, without any source domain performance degradation.

Index Terms— DPCCN, Mixture-Remix, cross-domain, speech separation, unsupervised target speech extraction

1. INTRODUCTION

Speech separation (SS) aims to separate each source signal from mixed speech. Traditionally, it has been done in the time-frequency (T-F) domain [1–3]. Recently, a convolutional time-domain audio separation network (Conv-TasNet) [4] was proposed, and has shown significant performance improvements over previous T-F based techniques. Since then, several works have focused on the time-domain methods, such as the DPRNN [5], DPTNet [6], SepFormer [7], etc. Due to the success of speech separation, researchers started to focus on the target speech extraction (TSE) — a sub-task of speech separation requiring additional target speaker clues to extract only a single speaker speech. Similarly, most of these works are also time-domain related, such as the TD-SpeakerBeam [8], SpEx+ [9] and channel-decorrelation [10].

In the current wave of time-domain based speech separation research, we may however ask "is it really the best to separate speech in time-domain?" Reviewing the latest time-domain results on WSJ0-2mix [1], a benchmark dataset of speech separation, the scale invariant signal-to-noise ratio (SISNR) [11] has reached 22.3 dB [7]. This is an excellent result and if we could achieve it on real speech separation, this SISNR value would be already good enough for real applications. However, authors in [12] show that, when evaluating the Conv-TasNet on a noisy and reverberated WSJ0-2mix dataset, the separation performance is significantly reduced. Besides, results in [13] also show that, using time-frequency domain techniques can deliver superior separation performance under complicated reverberant conditions. This may due to the fact that, for the time-domain speech separation systems, all speech transforms are learnt directly from the raw input waveforms. These transforms may be very sensitive to the environment variations, especially for the cross-domain tasks, in which the test conditions deviate far from the conditions of training dataset.

In addition, most state-of-the-art speech separation systems are supervised ones, they are trained to separate the sources from simulated mixtures created by adding up isolated ground-truth sources [4, 5]. This reliance on ground-truth precludes scaling to widely available real-world mixture data and limits the progress on open-domain tasks. Therefore, few latest works start to focus on the unsupervised/semi-supervised speech separation, such as [14], mixture invariant training [15] and its extensions [16, 17]. How to well exploit the real-world unsupervised mixtures to boost the current separation and extraction systems becomes very fundamental, important, and challenging.

In this paper, we focus on the noise and reverberation, cross-domain robustness for both speech separation and extraction from the time-frequency perspective. A densely-connected pyramid complex convolutional network (DPCCN) is first proposed. It is motivated by the network architecture in [18] (we call it DenseUNet for simplicity) – a U-Net [19] based structure that combines temporal convolutional network (TCN) [20] and DenseNet [21] to enhance the separation ability. Our improvements are: 1) the mixture magnitude is discarded, only its complex spectrum is taken as input to the network; 2) the decoder outputs are waveforms instead of the real and imaginary (RI) components of separated speech spectrum, and we replace the RI related loss function to the negative SISNR; 3) a pyramid pooling layer [22] is added at the end of the decoder to exploit more discriminative global information. Besides, we design a new speaker encoder to generalize the DPCCN to extract only the target speech for TSE task. Furthermore, to improve the system robustness for unsupervised cross-domain TSE, a novel and effective Mixture-Remix approach is proposed to adapt the trained source model to the target domain acoustic characteristics.
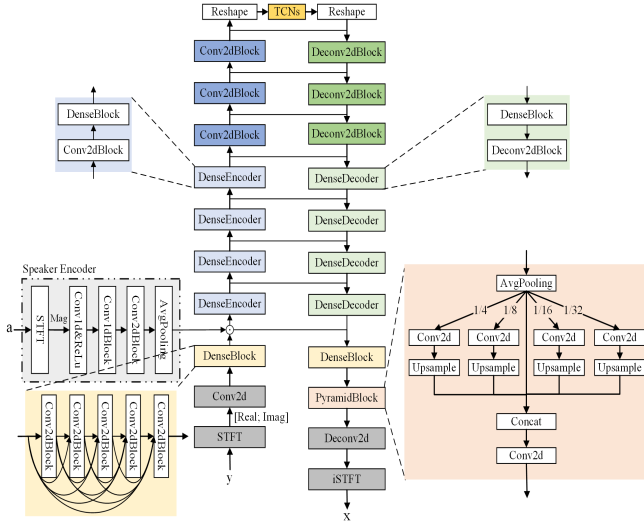
Fig. 1. Structure of DPCCN with speaker encoder (sDPCCN) for target speech extraction.

On the noisy and reverberant in-domain LibriSpeech dataset [23], the proposed DPCCN achieves more than 1.4 dB absolute SISNR improvement over all listed state-of-the-art time-domain speech separation methods. For the cross-domain speech separation and extraction tasks, we evaluate the proposed approaches on clean Libri2Mix [24] and Aishell2Mix that we created from Aishell-1 [25] corpus. Results show that the DPCCN-based systems are more robust and achieve significantly better performance than baselines. In particular, on the unsupervised cross-domain TSE tasks, our proposed Mixture-Remix with DPCCN fine-tuning significantly outperforms the TD-SpeakerBeam [8], a 3.5 dB absolute improvement in SISNR is obtained on target domain test set without any performance degradation on source domain.

2. PROPOSED METHODS

2.1. DPCCN

DPCCN follows a U-Net style to encode mixture spectrum into a high-level representation, then decodes it into the clean speech. In DPCCN, DenseNet is used to alleviate the vanishing-gradient problem and encourage the feature reuse; TCN is clamped between the codec to leverage long-range time information; A pyramid pooling layer is introduced after decoder to improve its global modeling ability. This architecture is used for speech separation. To generalize it for target speech extraction, an additional speaker encoder is designed to encode the target speaker information to guide the network in extraction of the target speech. The whole structure of DPCCN with the speaker encoder (termed sDPCCN) is shown in Fig. 1, where **y** is the mixture, **x** is estimated speech, and **a** is target speaker enrollment speech for TSE task.

The design of DPCCN is inspired by DenseUNet [18]. We improve it in the following ways: First, we discard the mixture magnitude and only use its real and imaginary (RI) parts as inputs to the network. This is because the RI parts inherently contain the magnitude information. Second, the outputs of DPCCN are waveforms instead of the RI parts of separated speech spectrum. Correspondingly, we replace the original RI related loss function to the standard negative SISNR, which has been proved to be effective for speech
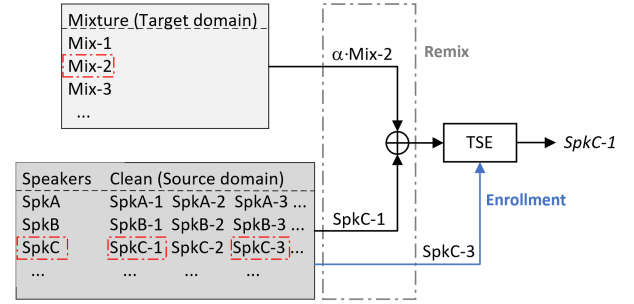
Fig. 2. Illustration of Mixture-Remix fine-tuning. Red dashed box means random selection. α is a scaling factor converted from SNR between the selected mixture and the clean source.

separation. Third, we introduce a pyramid pooling layer at the end of the decoder, considering that the limited receptive field of convolutional network may make the U-Net unable to sufficiently incorporate the useful global information. In the pyramid layer, different levels of global information are obtained by averaging the feature map at different scales, followed by bilinear interpolation upsampling and concatenation operations to form the final feature representation that ultimately carries both local and global context information.

In addition, we design a special speaker encoder to generalize the DPCCN for extracting only the target speech. In the speaker encoder, the magnitude spectrum of enrollment speech is first processed through a 1-D convolution followed by rectified linear unit (ReLU) function. Then a standard 1-D convolution block with channel-wise layer normalization [4] is used to model the temporal information. Next, it output is reshaped into a 4-D tensor and processed by the following 2-D convolution block. Finally, the average pooling operation along the time axis is used to aggregate the information of target speaker to guide DPCCN to extract the target speech through element-by-element multiplication.

2.2. Mixture-Remix Fine-tuning

The Mixture-Remix fine-tuning approach aims to improve the cross-domain robustness and performance of TSE systems under real scenarios, where only real-world target domain mixture data, and simulated source domain mixtures with isolated ground-truth source signals are available. We do not count on having any isolated ground-truth of target domain training mixtures. We call this scenario as "unsupervised cross-domain TSE", and the target domain mixtures are "unsupervised mixtures". Since most state-of-the-art TSE systems still heavily rely on the simulated training data, there is a huge performance gap between in-domain and cross-domain test conditions, due to acoustic mismatch between the source and target domain signals. Therefore, how to well utilize the available real mixtures to improve the TSE system is a challenging open question.

In Fig. 2, an efficient and simple Mixture-Remix approach is proposed. It generates new supervised TSE training data by remixing the real-world unsupervised mixtures of target domain with the isolated sources of source domain. In this way, the acoustic characteristics of both target and source domain can be captured. Then, the new supervised training data is used to fine-tune the TSE model that already well trained on the simulated source domain data. By doing so, the cross-domain acoustic mismatch can be automatically alleviated. The remixing is performed as illustrated in Fig.2: given a random target domain unsupervised mixture Mix-2, we randomly select a speaker SpkC with utterance SpkC-1 and SpkC-3 from

the source domain. Then, one of the selected utterance `SpkC-1`, is used to remix with `Mix-2` to generate a three-speakers mixture, and `SpkC-3` is taken as the target speaker enrollment speech to the new three-speakers mixture for supervision to extract `SpkC-1`. During the remixing, each unsupervised mixture is weighted by scaling factor α to produce more discriminative training mixtures as follows:

$$\alpha = \sqrt{\frac{E_s}{10^{\frac{SNR}{10}} \cdot E_m}}, \tag{1}$$

where E_s and E_m are energy of the selected isolated source signal and the unsupervised mixture belonging to the source and target domains, respectively.

3. TASK CONSTRUCTION

In-domain SS task: we use the noisy reverberant dataset proposed in [23]. This simulated dataset sampled at 16 kHz contains 20k, 5k and 3k utterances in training, validation and test sets, respectively. Two speakers from the 100-hour Librispeech [26] and one noise from 100 Nonspeech Sounds [27] are mixed to generate each mixture. Each utterance is convolved with room impulse responses generated by the image method [28] with reverberation time ranging from 0.1s to 0.5s. More details can be found in [23].

Cross-domain SS and TSE tasks: the English Libri2Mix [24] and Mandarin Aishell2Mix are used as the supervised source domain and unsupervised target domain dataset, respectively. Each mixture in Aishell2Mix is generated by mixing two speakers' utterances from Aishell-1 [25]. These utterances are randomly clamped to 4 seconds and rescaled to a random relative SNR between 0 and 5 dB. This simulated Aishell2Mix is taken as the available real-world unsupervised mixture data as described in section 2.2. For TSE task, the first speaker in the mixture of Libri2Mix is taken as the target speaker; its enrollment speech is randomly selected such that it differs from the one in the mixture. While in the Aishell2Mix, only the test set is provided with the ground-truth speech, the others are all unsupervised mixtures. We resample all the data to 8kHz. Details of the cross-domain tasks are shown in Table 1. Both experimental datasets and their simulation code are publicly available[1].

Table 1. Setup of cross-domain SS and TSE tasks. "Oracle" and "Enroll" indicate whether the oracle (ground-truth) data and enrollment speech are available.

Dataset	Type	#Spks	#Utts	Hours	Oracle	Enroll
Libri2Mix (Source Domain)	train	251	13900	58	✓	✓
	dev	40	3000	11	✓	✓
	test	40	3000	11	✓	✓
Aishell2Mix (Target Domain)	train	340	10000	11	-	-
	dev	40	3000	3	-	-
	test	20	3000	3	✓	✓

4. EXPERIMENTS

4.1. Configurations

Each `Conv2dBlock` in DPCCN includes a 2-D convolution, exponential linear units (ELU), and instance normalization (IN). Each `Deconv2dBlock` includes a 2-D deconvolution, ELU and IN. The `TCNs` contain 2 layers TCN, and each includes 10 TCN blocks

---

[1] https://github.com/jyhan03/icassp22-dataset

Table 2. SISNR/SISNRi performance of different speech separation systems on the noisy and reverberant mixed LibriSpeech dataset.

System	Domain	SISNR / SISNRi
Conv-TasNet [4]	Time	8.3 / 8.72
DPRNN [5]	Time	9.0 / 9.42
DPTNet [6]	Time	8.1 / 8.52
Sudo rm -rf [31]	Time	6.8 / 7.22
DCCRN [32]	Frequency	7.2 / 7.62
DenseUNet [18]	Frequency	8.78 / 9.19
Proposed DPCCN	Frequency	**10.40 / 10.82**
w/ magnitude spectrum	Frequency	10.36 / 10.78
w/o pyramid layer	Frequency	9.88 / 10.30

composed by IN, ELU and 1-D convolution with dilation factors $1, 2, 4, ..., 2^9$. In `PyramidBlock`, all convolution kernel and stride sizes are set to 1, the input/output channels of 2-D convolution before upsampling and after concatenation are 32/8 and 64/32, respectively. All other kernel and stride sizes, paddings, and convolution channels are the same as the DenseUNet in [18].

For (i)STFT layer in DPCCN, we use the square root of Hanning window with FFT size of 512 and hop size of 128. Global mean-variance normalization is applied to all input features. We train all models with Adam [29] optimizer, 100 epochs on 4-second long segments, and the initial learning rate is set to 0.001. Negative SISNR loss function is used for both SS and TSE system training. Utterance-level permutation invariant training (uPIT) [3] is used for SS to address the source permutation problem. SISNR [11] and its improvement, SISNRi, are used to evaluate the performance. All fine-tuning related experiments are finished within 20 epochs.

4.2. Results and discussion

4.2.1. In-domain task

We first compare robustness of our proposed DPCCN to noise and reverberation with several state-of-the-art separation methods for standard in-domain speech separation in Table 2, where the DPRNN, DPTNet and Sudo rm -rf results are the same as reported in [30], and the results of Conv-TasNet, DCCRN and DenseUNet are produced by ourselves on the same dataset. It is clear that our proposed DPCCN significantly outperforms all other systems. Particulary, the DPCCN achieves a 1.4 dB absolute SISNR improvement over the best time-domain system DPRNN [5]. Such a performance gain indicates that the DPCCN is a good candidate for robust speech separation under complicated conditions. In addition, adding the pyramid layer can bring 0.5 dB SISNR improvement, and the magnitude spectrum does not bring any benefits. We also trained the DPCCN with the loss function reported in DenseUNet [18], and we found that using the negative SISNR loss function can bring a 1.1 dB improvement in SISNR.

4.2.2. Cross-domain tasks

Another way to verify the system robustness is to investigate its performance behavior on cross-domain test sets. Table 3 compares the performance of our proposed DPCCN and sDPCCN with two strong baselines, the Conv-TasNet [4] and the TD-SpeakerBeam [8] for SS and TSE tasks, respectively. We also calculate an "ST-Gap" to show the cross-domain performance gap more clearly. It is defined as the SISNR difference between source and target domain test sets divided by the source domain SISNR, the lower the better.

Table 3. SISNR/SISNRi performance of speech separation and target speech extraction on Libri2Mix and Aishell2Mix. "TSB" means TD-SpeakerBeam. "ST-Gap" means the relative SISNR degradation between the source domain and the target domain, the lower the better. Systems are all trained on Libri2Mix.

Task	System	Libri2Mix	Aishell2Mix	ST-Gap
SS	Conv-TasNet	11.98 / 11.98	2.08 / 2.08	↓ 82.6%
	DPCCN	13.05 / 13.04	5.09 / 5.09	↓ 61.0%
TSE	TSB	12.20 / 12.23	1.85 / -0.65	↓ 84.8%
	sDPCCN	11.57 / 11.61	5.78 / 3.28	↓ 50.0%

In Table 3, all systems are trained on the source domain Libri2Mix but tested on both in-domain (Libri2Mix) and out-of-domain (Aishell2Mix) evaluation sets. Three findings are observed: 1) huge cross-domain performance gap exists in both SS and TSE tasks, either from the absolute SISNR numbers, or from the ST-Gap values; 2) DPCCN always shows much better speech separation performance than Conv-TasNet under both in-domain and cross-domain conditions; 3) significant target domain performance improvement (3.9 dB SISNR) is obtained by sDPCCN over TD-SpeakerBeam, even with slightly worse results on in-domain test set. All these findings confirm that the state-of-the-art time-domain methods, Conv-TasNet and TD-SpeakerBeam, are very sensitive to domain deviations, even for Libri2Mix and Aishell2Mix that are both clean speech but in different languages. Our proposed (s)DPCCN shows more robustness, although the absolute SISNR in target domain is still relative low.
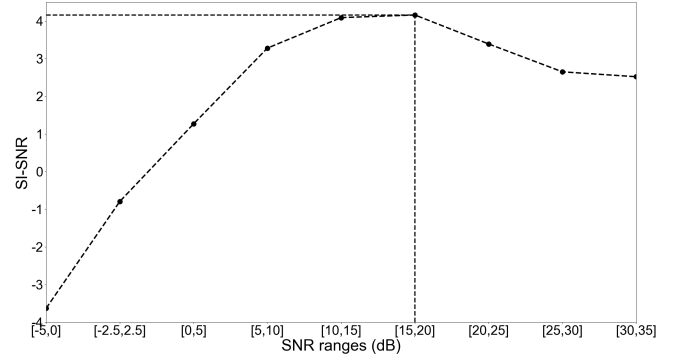
4.2.3. Mixture-Remix fine-tuning

Table 4 presents the target domain performance improvements using our proposed Mixture-Remix fine-tuning for target speech extraction. The 1st line results are upper bound (oracle) performance of sDPCCN trained from Aishell2Mix with ground-truth, and they also demonstrate the large cross-domain performance gap as observed in Table 3.

Table 4. SISNR/SISNRi performance of Mixture-Remix fine-tuning on TSE tasks. "A-L-3Mix" is the three-speakers mixture dataset generated by remixing Aishell2Mix and LibriSpeech. "TSB" refers to the TD-SpeakerBeam.

System	Train	Fine-tune (A-L-3Mix)	Evalation Libri2Mix	Aishell2Mix
sDPCCN	Aishell2Mix	-	2.72 / 2.75	9.44 / 6.93
	Libri2Mix	-	11.57 / 11.61	5.78 / 3.28
		✓	0.55 / 0.58	4.16 / 1.66
		+ Libri2Mix	**11.83 / 11.87**	**6.61 / 4.11**
TSB	Libri2Mix	-	12.20 / 12.23	1.85 / -0.65
		+ Libri2Mix	12.56 / 12.59	3.13 / 0.63

The 3rd line results are from the sDPCCN system first trained on the source domain Libri2Mix training set, and then fine-tuned using only the A-L-3Mix data generated by our Mixture-Remix approach. We find that the SISNR on Libri2Mix is extremely low (only 0.55 dB), which indicates that the English sDPCCN source model has been well adapted to the target Mandarin Aishell2Mix acoustic environments by the Mixture-Remix fine-tuning, and the adapted target model deviates significantly from the source model. Meanwhile, it is interesting to see that on Aishell2Mix, the fine-tuned performance, 4.16 dB is also much worse than the original 5.78 dB, without any fine-tuning operations. This may be due to the fact that, the remixed

Fig. 3. SISNR performance of sDPCCN system varying with different SNR ranges during remixing for Mixture-Remix fine-tuning.

A-L-3Mix mixtures contain three speakers whose acoustic conditions for target speaker are mismatched with the two-speakers mixed training and evaluation sets. Such a phenomenon indicates that the current speech extraction process is very sensitive to the number of speakers in the mixtures.

Therefore, we choose to add the two-speakers mixed source domain training data Libri2Mix, new generated A-L-3Mix mixtures, and their corresponding enrollment speech together to fine-tune the whole source model. The results are shown in the 4th line. To our surprise, both the in-domain and target domain TSE performance are improved: for sDPCCN system, the SISNR is improved from 11.57 to 11.83 dB, and 5.78 to 6.61 dB on Libri2Mix and Aishell2Mix test sets, respectively. And these improvements are consistent for the time-domain TD-SpeakerBeam system. All results indicate that, without sacrificing the performance on source domain test sets, the proposed Mixture-Remix fine-tuning not only can significantly improve the target domain performance for unsupervised cross-domain TSE scenarios, but also generalizes well for time-domain systems.

In addition, it is worth investigating how Mixture-Remix fine-tuning strategy is influenced by the SNR range during remixing. Fig.3 shows the SISNR performance on Aishell2Mix test set of sDPCCN system varying with different SNR ranges during the remixing stage. In the figure, the sDPCCN is only fine-tuned using the A-L-3Mix data and we see that a higher SNR range than that of the evaluation set (0 to 5 dB) is required to obtain better performance. This may because it is more difficult to extract the target speech from a three-speaker mixture than from a two-speaker mixture. Therefore, for all the fine-tuning experiments in Table 4, the best SNR range, from 15 to 20 dB is used.

5. CONCLUSION

In this study, we propose two novel methods for robust speech separation and extraction. One is the densely-connected pyramid complex convolutional network (DPCCN), and the other is a Mixture-Remix fine-tuning approach. Their robustness are verified not only in noisy and reverberant in-domain speech separation task, but also examined in cross-domain speech separation and extraction tasks. Experiments show that the proposed (s)DPCCN has much better performance and robustness over other state-of-the-art separation and extraction methods. Moreover, the Mixture-Remix fine-tuning is proved to be effective to improve the target speech extraction system under real-world unsupervised cross-domain scenarios. Our future work will focus on generalizing the Mixture-Remix fine-tuning to different unsupervised speech separation tasks.

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246–250.

[3] 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.

[4] 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.

[5] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[6] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020, pp. 2642–2646.

[7] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[8] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. ICASSP*, 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," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[10] J. Han, W. Rao, Y. Wang, and Y. Long, "Improving channel decorrelation for multi-channel target speech extraction," in *Proc. Interspeech*, 2021, pp. 1847–1851.

[11] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. ICASSP*, 2019, pp. 626–630.

[12] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 696–700.

[13] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. ICASSP*, 2020, pp. 6359–6363.

[14] Y. Hoshen, "Towards unsupervised single-channel blind source separation using adversarial pair unmix-and-remix," in *Proc. ICASSP*, 2019, pp. 3272–3276.

[15] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Advances in Neural Information Processing Systems*, 2020.

[16] S. Wisdom, A. Jansen, R. J. Weiss, H. Erdogan, and J. R. Hershey, "Sparse, efficient, and semantic mixture invariant training: Taming in-the-wild unsupervised sound separation," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021.

[17] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "Teacher-student MixIT for unsupervised and semi-supervised speech separation," in *Proc. Interspeech*, 2021, pp. 3495–3499.

[18] Z.-Q. Wang and D. Wang, "Count and separate: Incorporating speaker counting for continuous speaker separation," in *Proc. ICASSP*, 2021, pp. 11–15.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[20] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 4700–4708.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. CVPR*, 2017, pp. 2881–2890.

[23] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. ICASSP*, 2020, pp. 6394–6398.

[24] 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.

[25] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[27] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Y. Luo, C. Han, and N. Mesgarani, "Group communication with context codec for lightweight source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1752–1761, 2021.

[31] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.

[32] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.