INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SDNet: Stream-attention and Dual-feature Learning Network for Ad-hoc Array Speech Separation

Honglong Wang[1], *Chengyun Deng*[2],*Yanjie Fu*[1], *Meng Ge*[3,*], *Longbiao Wang*[1,*],
Gaoyan Zhang[1], *Jianwu Dang*[1], *Fei Wang*[2]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Beijing Xiaoju Technology Co., Ltd., Beijing, China
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

hlwang@tju.edu.cn, gemeng@tju.edu.cn, longbiao_wang@tju.edu.cn

Abstract

Considerable progress has been made in multi-channel speech separation for fixed arrays. In this paper, we aim to develop a robust system for ad-hoc arrays to deal with uncertainties of microphone locations and numbers. Previous works commonly used the averaging method for ad-hoc arrays, overlooking the diversity of microphones in various positions. Some studies suggest that microphones with high signal-to-noise ratio(SNR) are more helpful in improving speech quality. Motivated by this, we propose stream-attention and dual-feature learning network called SDNet. The key points are as follows: 1) We propose a dual-feature learning block with fewer parameters to learn the long-term dependency better. 2) Based on this high-quality speech representation, we further propose stream attention that effectively handles microphone variability and allocates more attention to microphones with higher SNR. Experiments show that our proposed model outperforms other advanced baselines.

Index Terms: speech separation, ad-hoc, deep learning, stream attention

1. Introduction

With the development of artificial neural networks, deep-learning-based speech separation has achieved notable results. This technique is usually used as a front end for automatic speech recognition (ASR) or to improve auditory perception in humans. Multiple microphones are often necessary for effective speech separation in real-world scenarios, but fixed arrays may not always be practical due to spatial constraints.

Speech signal can be modeled in the time-domain or time-frequency (T-F) domain, depending on whether short-time Fourier transform (STFT) is required. Many monaural approaches perform well irrespective of the need for STFT [1, 2]. Furthermore, many studies [3, 4] have shown that multiple microphones can improve separation performance. Given the good performance of some traditional beamforming methods, some multi-channel speech separation models explore them in combination with deep neural networks [5, 6, 7]. Other methods adopt a data-driven approach and train a model to map multi-channel mixed speech to a clean version relative to a reference microphone [8, 9]. In [10], an embedding module was employed to estimate beamforming weights using a multi-channel approach, which expanded on the single-channel convolutional recurrent network (CRN) [11], also shaped like a U-Net [12].

However, all of these aforementioned multi-channel methods are based on fixed arrays. Processing speech signals using spatially distributed microphones with unknown numbers and positions remains a challenging task. The transform-average-concatenate (TAC) module [13] and its variations [14, 15, 16] offer a simple yet effective solution to the microphone number invariant problem by averaging and concatenating, although this approach may not necessarily be optimal. Different from TAC, [17] adopts a sum operation rather than concatenate. Although the average operation is suitable to tackle the problem of microphone permutation and number variation, it ignores the diversity of microphones in different positions [18]. [19] use an attention mechanism for spatial processing, but the utilization of spatial information is limited by the use of only the magnitude spectrum. [20, 21] utilize a triple-path network in time domain, in which two paths are used for temporal processing, and one path is used for spatial processing.

Some studies [18, 22] have indicated that microphones with higher SNR are more effective in improving speech quality for enhancement tasks. Based on this observation, we propose a framework called: stream-attention and dual-feature learning network (**SDNet**) for distributed microphone array speech separation in T-F domain. The key contributions are as follows: Firstly, instead of placing the sequence model at the bottleneck as most CRN methods do, we propose to use dual-feature learning(DFRNN) blocks in the encoder to better learn long-term dependency. This can help alleviate the issue of inadequate feature learning, because the frequency dimension will gradually decrease in the encoder. Each DFRNN block comprises two RNNs following behind the encoder block, one for frequency learning and the other for feature map channel learning. By learning features independently and iteratively, the speech information of each microphone will be represented better with fewer trainable parameters. The U-Net with the DFRNN block is referred to as **DNet**. Secondly, we consider better utilization of spatial information based on this high-quality speech representation. The distance between microphones and speakers or noises can vary significantly in an ad-hoc array, leading to varying SNR among microphones and different gains. To improve the utilization of microphones that are more useful for speech separation, we use stream attention to assign higher weights to microphones with higher SNR. The experimental results show that our proposed method outperforms the advanced baselines.

2. Signal model and baseline

In the T-F domain, the multi-channel noisy signal X recorded by an M-channel microphone array can be represented as:

$$X_m(f,t) = \sum_{p=1}^{P} S_m(f,t) + N_m(f,t) \tag{1}$$

where $m \in [1, M], f \in [1, F], t \in [1, T]$ and $p \in [1, P]$ respectively denote the indices of microphones, frequencies, time and speakers. This work focuses on estimating P clean signals

* denotes the corresponding author.

10.21437/Interspeech.2023-1766

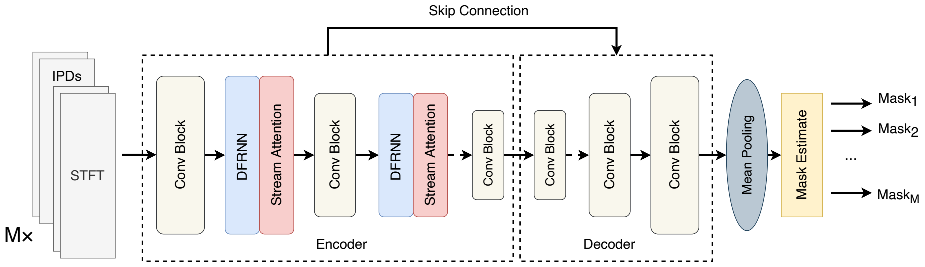

Figure 1: *The flowchart of the proposed framework SDNet. "DFRNN" is used to model the temporal dependencies. "Stream attention" is responsible for processing unknown geometry and reassigning weights to microphones at various distances. M complex masks will be estimated to multiply with the mixtures in a filter-and-sum manner.*

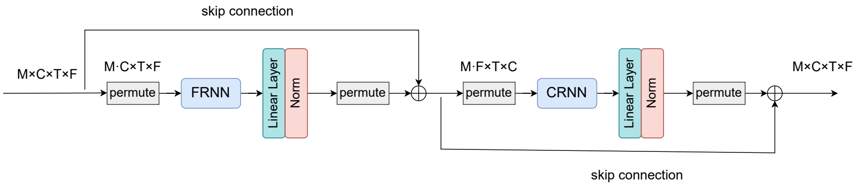

Figure 2: *The flowchart of DFRNN. "FRNN" and "CRNN" mean the input feature is the frequency and feature map channel dimension.*

from noisy and reverberated speech signals. The training target is the reverberant clean speech signal, so reverberation suppression is not considered.

EabNet [10] adopts the U^2-Net [23] architecture in the multi-channel speech enhancement task and achieves good results. Single-channel methods usually receive concatenated real and imaginary parts of STFT spectrogram as input features. EabNet further concatenates spectrograms of all microphones to create the input $X \in \mathbb{R}^{C\times F\times T}$ for multi-channel processing, $C = 2M$. Spatial information can be learned by several convolving kernels during down/up sampling, and eventually represented by an embedding with spectral information. This embedding is used to estimate M complex masks to multiply with the noisy mixtures in a filter and sum manner. This method use only convolutional kernels to learn spatial information may work for a fixed array, but it's not effective for ad-hoc arrays.

3. SDNet Architecture

The overall diagram of the proposed model is shown in Fig. 1. In our pipeline, inspired by [15], we introduce an extra dimension stream (i.e. microphones) as the first dimension. Each stream contains all the information of corresponding microphone. Many speech processing methods for ad-hoc arrays ignore the role of spatial features such as inter-channel phase difference (IPD). In our experiments, we find that better results will be obtained if IPD is used, although the distances between each microphone are relatively large. The number of microphones is zero-padded to an upper bound M. We compute the cosine and sine values of IPDs by the phase difference between the first microphone and other $M-1$ microphones of complex spectrogram. STFT and IPDs are concatenated along the channel dimension. Thus our input is $X \in \mathbb{R}^{M\times C\times F\times T}$, $C = 4$. After the last layer of decoder, we average pool the information from each stream fused by attention and feed it to the mask estimation module. This approach can model the representation of each microphone well and fully learn spatial information.

3.1. Dual-feature Learning RNN

Prior work based on U-Net architecture usually put the sequence modeling module in the middle of the encoder and decoder to learn long-term dependencies. However, due to repeated downsampling in the encoder, the frequency dimension of the input to the bottleneck is greatly reduced, potentially impeding the temporal module's ability to fully capture frequency information. Another option we suggest is to put sequence modeling module after each encoder block. Inspired by DPRNN [24], we propose a dual-feature learning RNN block, this approach has better temporal modeling ability and can leverage more information due to multi-scale learning.

The schema of the DFRNN is shown in Fig. 2. Each DFRNN block contains two RNN layers for frequency and feature map channel learning iteratively across time (frame) dimension. More specifically, the stacked two RNN layers expect to learn long-term dependencies respectively from the frequency dimension and the feature map channel dimension. The first RNN layer called FRNN can learn abundant information from frequency bins to distinguish different voices. The frequency dimension is not the sole dimension in the latent domain due to the increase in the number of channels caused by convolution operation. The second RNN layer, CRNN, is essential for learning the relationships among feature maps. After each RNN layer, a linear layer and a layer normalization layer are appended to make the network robust.

3.2. Stream attention

Paper [25] uses channel attention to determine the importance of each feature map. Similar to EabNet, the input to the channel attention (CA) units are 3D tensors with shape $X \in \mathbb{R}^{C\times F\times T}$, $C = 2M$. It should be noted that after the processing of convolution block such as DenseNet [26], the number of feature maps will be changed. We use $'$ to indicate potential variable dimensions due to convolutional resampling. In other words, channel attention units are performed on the tensors of shape $X \in \mathbb{R}^{C'\times F'\times T'}$, and the dimensions of the input are different for CA unit at different blocks. So the number of feature maps will not always be equal to the number of microphones. The calculation of attention weights will make less sense as far as signal theory is concerned.

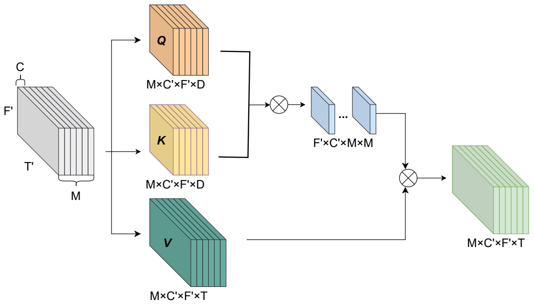

Figure 3: *Illustration of the calculation process of stream attention. Q, K and V are transformed from X by point-wise conv-2d.*

Different from [25] performing self-attention on different feature maps, we propose stream attention as shown in Fig. 3, which is performed on different streams. The output of encoder block $X \in \mathbb{R}^{M \cdot C' \times F' \times T}$ is rearranged to $4D$ tensors $X \in \mathbb{R}^{M \times C' \times F' \times T}$, then stream attention is performed. We first compute $Q, K \in \mathbb{R}^{M \times C' \times F' \times D}$, and $V \in \mathbb{R}^{M \times C' \times F' \times T}$ from X. Note that we call each frequency bin of feature map "c-f bin". For each c-f bin, three latent representations $Q_{c,f} \in \mathbb{R}^{M \times D}$, $K_{c,f} \in \mathbb{R}^{M \times D}$ and $V_{c,f} \in \mathbb{R}^{M \times T}$ can be computed using point-wise 2D convolution (PConv2d) as following:

$$Q_{c,f}, K_{c,f}, V_{c,f} = PConv2d_{1,2,3}(X) \tag{2}$$

We multiply the query and key to calculate the similarity. And then a softmax function is applied to the last dimension(stream) of the attention weights matrix:

$$W_{c,f} = softmax(Q_{c,f}^{\mathrm{T}} \cdot K_{c,f}) \tag{3}$$

After normalizing the weights matrix with the softmax function, we multiply it by V to assign various streams various weights as follows:

$$Y_{c,f} = W_{c,f} \cdot V_{c,f}^{\mathrm{T}} \tag{4}$$

The above equations show that c-f bin weights differ across streams at identical positions. The shape of the whole attention matrix is $W \in \mathbb{R}^{C' \times F' \times M \times M}$. As a result, different feature maps in a stream and frequency bands in a feature map will get varying attention scores. Compared with common attention calculation procedure, this approach not only selects the best streams but also reassigns attention weights to different frequency bands and feature maps within a stream. After that, we rearrange Y into 3D tensors for next encoder block.

4. Experiments

4.1. Dataset

In [13], the authors develop a multi-channel noisy reverberant dataset for ad-hoc array speech separation. Following the same configuration[1], we simulated 20000, 5000, and 3000 4-second long utterances for training, validation, and testing, respectively. The dataset includes an equal distribution of microphone numbers ranging from 2 to 6. Each utterance is composed of two speakers and one nonspeech noise which are randomly selected from the 100-hour Librispeech dataset [27] and the 100 Nonspeech Corpus [28]. The locations of all the microphones, speakers and noises are random. The sampling rate is 16 kHz. As the microphones are assumed to be the same, the signals are synchronized. Please refer to [13] for more details.

4.2. Experimental setup

4.2.1. Network parameters

We use the same convolution blocks as [10] in the encoder and decoder, with the exception of the number of convolution channels. As the entire network has a U^2-Net shape, each convolution block in the en/decoder is also a U-Net network. The number of encoding layers within the U-Net block is {4, 3, 2, 1, 0} and the reverse is for decoding layers. For each down/up sampling block, the kernel and stride sizes are (1, 3) and (1, 2), and the number of convolution channels is 32 except the last one of encoder and decoder which is 64. 3 DFRNN blocks composed of 2 stacked 1-layer-bidirectional LSTM are placed after the first 3 blocks of encoder. In addition, the number of stream attention blocks is set to 4. D is set to 32.

4.2.2. Training details

The input representation is the complex STFT computed using a Hanning window of length 320 samples (20ms) with a hop size of 160 samples. We use the Adam optimization algorithm to train the models with the learning rate of 0.001. We use negative scale-invariant signal-to-noise ratio (SI-SNR) [29] with utterance-level permutation invariant training (uPIT) [30] as the loss function for all models. SI-SNR improvement(SI-SNRi), PESQ [31], and STOI [32] are used to evaluate the separation performance. All results are derived by averaging values from two speakers and across all examples in the test set. Except for the results in Table 1, which are trained for a maximum of 100 epochs, all models undergo training for up to 150 epochs with early stopping after 10 consecutive epochs of no improvement in validation accuracy. Due to space limitations, only the results for 2, 4, and 6 microphones are reported for datasets that include various microphone counts.

Table 1: *Ablation study on the part of the dataset with 6 microphones. "Par" denotes the number of parameters.* $\overline{v}$ *denotes the average of the three experimental results. * denotes using different sequence modeling strategies.*

Model	Par.(M)	SI-SNRi	PESQ
UNet*TCN	1.84	7.92	1.64
+IPDs	1.85	$\overline{9.82}$	$\overline{1.70}$
*DFRNN	0.72	$\overline{10.40}$	$\overline{1.81}$

4.3. Results and discussion

We conduct ablation experiments using a subset of the dataset with only 6 microphones, as shown in Table 1. DNet performs not very well since less spatial information has been learned. But the incorporation of IPDs has resulted in substantial advancement. This indicates that manually calculated IPD can improve separation performance in an ad-hoc array consisting of identical microphones. To solve the problems mentioned in section 3.1, we remove the stacked temporal convolutional network

[1] https://github.com/yluo42/TAC

Table 2: *Experimental results with different multi-channel schemes on the ad-hoc array. A clear trend is that better results can be achieved with more microphones. Due to space restrictions, only the results of 2/4/6 microphones are presented.*

Model	of mics	SI-SNRi	PESQ	STOI
DF-EabNet	2/4/6	12.79/13.00/13.05	2.14/2.17/2.15	0.89/0.89/0.89
+channel attention[25]		12.75/13.24/13.35	2.18/2.25/2.24	0.89/0.90/0.90
DNet		11.40/12.69/12.98	1.96/2.12/2.13	0.86/0.88/0.89
+stream pooling [15]		13.30/14.61/15.12	2.25/2.48/2.53	0.90/0.92/0.93
+stream attetnion		14.84/15.85/16.11	2.49/2.70/2.75	0.92/0.93/0.94

Table 3: *Experimental results compared with advanced baselines on ad-hoc array with various numbers of microphones. $(\cdot)$ denotes the results are from the original paper. Bold indicates the best results.*

Model	Par.(M)	of mics	SI-SNRi	PESQ	STOI
FasNet-TAC [13]	2.9	2/4/6	11.17/12.14/12.43	1.77/1.84/1.83	0.84/0.86/0.86
EabNet [10]	2.8		12.10/12.62/12.58	2.00/2.07/2.05	0.87/0.89/0.89
TPRNN [20]	2.24		(13.25)/(14.27)/(14.65)	-	-
SDNet	0.85		**14.84/15.85/16.11**	**2.49/2.70/2.75**	**0.92/0.93/0.94**

(TCN) in the bottleneck and then add our DFRNN blocks after the convolution block termed as DNet. Compared to putting the temporal module in the bottleneck, DFRNN can fully utilize information from different dimensions. And each dimension only needs to be learned with a small network. Compared with TCN, using DFRNN with fewer parameters yields better results. It demonstrates the viability of DFRNN. This multi-scale learning might be better than learning at bottleneck since the DFRNN block immediately follows the convolution block.

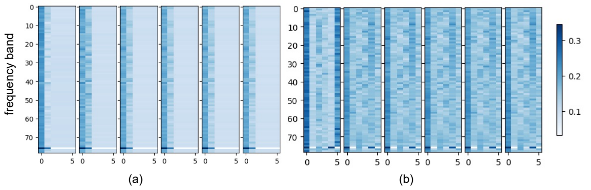

Figure 4: *(a) Attention matrix W from 2 microphones input. (b) W from 6 microphones input.*

Table 4: *Input SNR of one example as Fig. 4 (b). shown.*

SNR	mic0	mic1	mic2	mic3	mic4	mic5
spk0	-9.99	2.47	-4.04	1.13	**10.76**	-13.25
spk1	9.42	-8.83	3.30	-2.47	-12.91	**12.52**
mean	-0.28	-3.18	-0.36	-0.67	-1.07	-0.36

Table 2 shows the results of different spatial information learning methods on the whole dataset. In order to verify the effectiveness of channel attention(CA) [25], we add the CA unit to the EabNet with DFRNN block(termed as DF-EabNet) directly, yet the results barely change. We assume that the feature maps are no longer associated with specific microphones in the latent domain, resulting in biased attention computation. For DNet, we first use stream pooling method which is similar to TAC [13]. Stream pooling improved the results, which also demonstrated the effectiveness of the previous averaging method. However, stream attention improves the performance further. Stream attention pays attention to different groups of feature maps, and the number of groups and microphones are fairly even. We think this mechanism can implicitly select the microphones (stream) that contain more speech information and less noise information. We compute the average of W over the feature map channel dimension and visualize the score of the first stream attention block where streams have not yet been integrated, as shown in Fig. 4. Each subfigure shows $\{W_{F' \times M \times 0}, ..., W_{F' \times M \times M}\}$ from left to right, where $M = 6$. It can only focus on the stream with actual spectrum for input with padding. Since microphone 0 is chosen as the reference microphone, the corresponding score is high. In addition, microphones 4 and 5, which correspond to a speaker with a higher SNR, have also received more attention.

As shown in Table 3, we compare our proposed network with some advanced baselines. FasNet-TAC [13] is a time domain multi-channel speech separation method with a TAC module for microphone permutation and number invariant processing. EabNet [10] is a multi-channel speech enhancement method in T-F domain, we simply increase the number the estimated complex mask for multiple source separation. TPRNN [20] is a recently proposed state-of-the-art method in time domain on the same dataset. Considering only the results on the 6-microphone test dataset, our method achieved a 28.06% relative improvement in SI-SNRi over the baseline, and a 9.96% relative improvement over the state-of-the-art (SOTA) method. Our proposed method also achieves the best results among the list baselines with fewer parameters on other metrics.

5. Conclusion and future work

In this work, we propose DNet to model the long-term dependency signal better by learning features independently and iteratively. Furthermore, we introduce stream attention mechanism for the variant number and permutation problem in unknown geometry, thereby better utilizing spatial information and focusing more on the microphones that are more helpful for separation. Our proposed SDNet reaches a new state-of-the-art on ad-hoc array speech separation task with fewer parameters. Considering the non-real-time nature of our network, future work involves optimizing this aspect to enhance its performance.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176182 and DiDi GAIA Research Collaboration Plan.

7. References

[1] 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.

[2] 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," *arXiv preprint arXiv:2008.00264*, 2020.

[3] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-tasnet: Time-domain audio separation network meets frequency-domain beamformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6384–6388.

[4] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "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.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[6] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6089–6093.

[7] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-spex: Localized target speaker extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7287–7291.

[8] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 260–267.

[9] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1888–1900, 2020.

[10] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6487–6491.

[11] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[12] 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*. Springer, 2015, pp. 234–241.

[13] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6394–6398.

[14] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, and N. Kanda, "Vararray: Array-geometry-agnostic continuous speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6027–6031.

[15] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Wang, Z. Chen, and X. Huang, "One model to enhance them all: array geometry agnostic multi-channel personalized speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 271–275.

[16] T. Yoshioka, X. Wang, and D. Wang, "Picknet: Real-time channel selection for ad hoc microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 921–925.

[17] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, "Scene-Agnostic Multi-Microphone Speech Dereverberation," in *Proc. Interspeech 2021*, 2021, pp. 1129–1133.

[18] Z. Yang, S. Guan, and X.-L. Zhang, "Deep ad-hoc beamforming based on speaker extraction for target-dependent speech separation," *Speech Communication*, vol. 140, pp. 87–97, 2022.

[19] D. Wang, Z. Chen, and T. Yoshioka, "Neural Speech Separation Using Spatially Distributed Microphones," in *Proc. Interspeech 2020*, 2020, pp. 339–343.

[20] X. Xiang, X. Zhang, and W. Xie, "Distributed microphones speech separation by learning spatial information with recurrent neural network," *IEEE Signal Processing Letters*, vol. 29, pp. 1541–1545, 2022.

[21] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Tadrn: Triple-attentive dual-recurrent network for ad-hoc array multichannel speech enhancement," *arXiv preprint arXiv:2110.11844*, 2021.

[22] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Comparison of reference microphone selection algorithms for distributed microphone array based speech enhancement in meeting recognition scenarios," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 316–320.

[23] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.

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

[25] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 836–840.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[27] 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.

[28] G. Hu, "100 nonspeech sounds," http://web.cse. ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[30] 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.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.