This article has been accepted for publication in a future proceedings of this workshop, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SiPS55645.2022.9919247, 2022 IEEE Workshop on Signal Processing Systems (SiPS)

DE–DPCTnet: Deep Encoder Dual-path Convolutional Transformer Network for Multi-channel Speech Separation

Zhenyu Wang[a,b,d], Yi Zhou[a,b], Lu Gan[c], Rilin Chen[d], Xinyu Tang[a,b], and Hongqing Liu[a,b]
[a] School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] Chongqing Key Laboratory of Signal and Information Processing, Chongqing 400065, China
[c] College of Engineering, Design and Physical Science, Brunel University, London UB8 3PH, U.K
[d] Tencent AI Lab, Beijing, China
zhenyuwangs@163.com, zhouy@cqupt.edu.cn, Lu.Gan@brunel.ac.uk, rilinchen@tencent.com,
xinyutangs@outlook.com, hongqingliu@cqupt.edu.cn

***Abstract*—In recent years, beamforming has been extensively investigated in multi-channel speech separation task. In this paper, we propose a deep encoder dual-path convolutional transformer network (DE–DPCTnet), which directly estimates the beamforming filters for speech separation task in time domain. In order to learn the signal repetitions correctly, nonlinear deep encoder module is proposed to replace the traditional linear one. The improved transformer is also developed by utilizing convolutions to capture long-time speech sequences. The ablation studies demonstrate that the deep encoder and improved transformer indeed benefit the separation performance. The comparisons show that the DE–DPCTnet outperforms the state-of-the-art filter-and-sum network with transform-average-concatenate module (FaSNet-TAC), even with a lower computational complexity.**

***Index Terms*—Speech separation, multi-channel, deep encoder, improved transformer, beamforming**

I. Introduction

The objective of blind source separation (BSS) is to extract the speech signal of individual speaker from single- or multi-channel observed mixed signals, which is also called cocktail party problem [1]. With the developments of neural networks, the deep learning based approaches [2]–[9] have gained much attractions and demonstrated much improved performances over the traditional methods.

The deep learning beamforming (BF) is one of the popular methods for multichannel speech separation with its competitive separation performance. The current deep learning BF approaches can be generally divided into two categories: the output–based beamformers [10]–[15], where a two-step concept is applied, and the DNN–based beamformers [8], [9], [16]–[19], where the BF filters are learned by networks. The first step of the output based beamformers is to apply single-channel separation model to each channel, and then, in the second step, the conventional BF processing is conducted. Note that when the pre-separation model fails, it will lead to the instability of the whole system output. The DNN–based methods are proposed to enable the networks to directly estimate the BF filters in either time domain or frequency domain, and achieve a better performance.

A recently developed system, which employs filter-and-sum network with transform-average-concatenate module (FaSNet-TAC) [9], produces the state-of-the-art performance in DNN–based methods. This model applies dual–path recurrent neural networks (DPRNN) [5] involving a lot of recurrent neural networks (RNNs), which is hard to train. Although the network using temporal convolutional network (TCN) [4] can speed up the training process by parallel computing, its performance is inferior. Besides, the encoder of the FaSNet-TAC is a linear transformation, thus it also limits the system performance due to its linearity. Through comparative experiments, it was pointed out in [20] that the strong performance of the time-domain separation model comes from using extremely small encoder window length, for example, 40 samples (5 ms in 8 kHz) in Tasnet [3], 16 samples (2 ms in 8 kHz) in Conv-tasnet [4], and 2 samples (0.25 ms in 8 kHz) in [5], [6]. This indicates that the tremendous computing resources are required to process with small encoder windows.

To address these disadvantages, in this paper, we propose a deep encoder dual-path convolutional transformer network (DE–DPCTnet) and the core contributions are threefolds:

- The DPCTnet with transformer module is constructed to estimate the BF filters instead of the DPRNN in the FaSNet-TAC. By fully utilizing the power of self-attention mechanism, it not only speeds up the training process by parallel computing, but also achieves a better global information extraction ability.
- The deep encoder with nonlinear activation function is utilized to create the encoder module, which exploits the nonlinear information and completes the separation task better.
- The proposed network with 16 ms encoder window length produces a better performance than the FaSNet-TAC with 4 ms encoder window length, which means that we achieve a better performance with one-third of

This article has been accepted for publication in a future proceedings of this workshop, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SiPS55645.2022.9919247, 2022 IEEE Workshop on Signal Processing Systems (SiPS)

the computational complexity. Moreover, under the same computational complexity, the proposed model attains an absolute improvement of 1.76 dB SDR.

II. Model Description

A. DE–DPCTnet

Figure 1 shows the flowchart of the DE–DPCTnet. The input signals with N channels $\boldsymbol{x}^i, i = 1, ..., N$ are segmentated into Z overlaped frames of L samples with a hop size of $H \in [0, L-1]$ samples. By concatenating W both future and past samples, the final input frame is

$$\boldsymbol{c}_t^i = \boldsymbol{x}^i[tH - W : tH + L + W - 1], \tag{1}$$

where $t = 1, ..., Z$ is the frame index. Without introducing ambiguity, the frame index t will be dropped in the following discussions. The convolution operation is now applied on a context frame $\boldsymbol{c}^i \in \mathbb{R}^{1\times 2W+L}$ for each microphone signal to generate the beamformer output, given by

$$\boldsymbol{y}_j^i = \boldsymbol{h}_j^i \circledast \boldsymbol{c}^i, \tag{2}$$

where $\boldsymbol{h}_j^i \in \mathbb{R}^{1\times 2W+1}, j = 1, ..., C$, which will be estimated by our proposed DPCTnet, are the BF filters of length $2W+1$, C is the number of sources and $\circledast$ represents the convolution operation, $\boldsymbol{y}_j \in \mathbb{R}^{1\times L}$ is the output of filtered signal for source j, which is the sum and average of $\boldsymbol{y}_j^i$.

Next, each subblock will be described including deep encoder and the DPCTnet in the DE–DPCTnet in details.

B. Deep Encoder

As discussed in the introduction, we investigate the possibilities of nonlinear encoders and decoders, by stacking 1-D convolution layers with nonlinear activation functions to construct the encoder, shown in Figure 2. The first layer encodes each context frame $\boldsymbol{c}^i$ to channel embedding by a 1-D convolutional layer with K kernels of size $2W + L$ to generate $\boldsymbol{R}^i \in \mathbb{R}^{1\times K}$ as

$$\boldsymbol{R}^i = \boldsymbol{c}^i\boldsymbol{U}, \tag{3}$$

where $\boldsymbol{U} \in \mathbb{R}^{(2W+L)\times K}$ is the weight matrix. The output $\boldsymbol{R}^i$ is passed to a stack of $M-1$ 1-D convolutional layers, and followed by a parametric rectified linear unit (PReLU), given by

$$\boldsymbol{R}_m^i = PReLU\left(\boldsymbol{U}_m \circledast \boldsymbol{R}_{m-1}^i\right), \tag{4}$$

where $m = 1, ..., M$ denotes the layer index, $\boldsymbol{U}_m \in \mathbb{R}^{K\times 3}$ is the kernel, and $\boldsymbol{R}_m^i$ is the output of the m-th layer.

C. DPCTnet

1) Segmentation: To avoid ambiguity, the microphone index i is omitted in following discussions. First, we extract multi-channel feature information $\boldsymbol{q} \in \mathbb{R}^{(2W+1)\times Z}$ with the normalizad cross-correlation (NCC) feature [9] and generate the final input features $\boldsymbol{F} \in \mathbb{R}^{(2W+K+1)\times Z}$ by concatenating $\boldsymbol{R}$ and $\boldsymbol{q}$. Second, the input features $\boldsymbol{F}$ are converted to channel embedding $\boldsymbol{G} \in \mathbb{R}^{E\times Z}(E < 2W + L)$ by a 1-D convolutional layer for reducing the computation cost and splited into overlapped chunks of length S and hop size $S/2$. Third, all the chunks are concatenated to form a 3-D tensor $\boldsymbol{D} \in \mathbb{R}^{E\times S\times P}$.

2) Dual-path convolutional transformer block: The dual-path convolutional transformer block consists of intra block processing and inter block processing. As shown in Figure 3, 3-D tensor D will be passed to a stack of B DPCT blocks designed to decouple the mixture signal by alternating local and global sequence processing. The output of each DPCT block is connected to TAC module [9], as depicted in Figure 1.

In the task of intra-processing, we use $\boldsymbol{D}_b^{RI}$ and $\boldsymbol{D}_b^{RO}$ to respectively denote the inputs for local LSTM and the corresponding outputs, and the relationship is

$$\boldsymbol{D}_b^{RO} = [\boldsymbol{M}_b \times BLSTM(\boldsymbol{D}_b^{RI}[:, :, p]) + \boldsymbol{e}_b, p = 1, ..., P], \tag{5}$$

where $\times$ is a matrix multiplication and $b = 1, ..., B$ is used to denote the b-th block. $\boldsymbol{D}_b^{RI}[:, :, p] \in \mathbb{R}^{E\times S}$ refers to the local sequence within the p-th chunk. $\boldsymbol{M}_b \in \mathbb{R}^{E\times 2S}$ and $\boldsymbol{e}_b \in \mathbb{R}^E$ are the parameters of the linear layer.

After intra-processing, in the task of inter-processing, we develop an improved transformer, as shown in Figure 4, to capture the global information. The improved transformer mainly utilizes convolutional feed-forward network to model the inter-segment sequences. Instead of directly taking $\boldsymbol{D}_b^{RO}$ as the input to the transformer for inter-processing, $\boldsymbol{D}_b^{RO}$ is summed with the positional encoding $\boldsymbol{PE}_b$ and passed to multi-head attention module (MHA), given by

$$\boldsymbol{D}_b^{MHA} = MHA\left[GLN(\boldsymbol{D}_b^{RO}[:, s, :] + \boldsymbol{PE}_b)\right] + \boldsymbol{D}_b^{RO}, \tag{6}$$

where GLN is the global layer normalization. $\boldsymbol{D}_b^{RO}[:, s, :] \in \mathbb{R}^{E\times P}$, and $s = 1, ..., S$ refers to the global sequence within the s-th chunk. In this work, we use the sinusoid positional encoding to generate $\boldsymbol{PE}_b \in \mathbb{R}^{E\times P}$.

According to layer normalization, after MHA, the convolutional feed-forwad layers are applied to the $\boldsymbol{D}_b^{RO}$ for extracting global features as

$$\boldsymbol{D}_b^{AO} = GLN[FFN(GLN[\boldsymbol{D}_b^{MHA}])], \tag{7}$$

$$FFN(\chi) = GELU[Conv1d(GELU[Conv1d(\chi)])], \tag{8}$$

where $GELU$ is the Gaussian error linerar units, and $Conv1d$ is 1-D convolutional layer with the kernel of size 1. $\boldsymbol{D}_b^{AO}$ is the output of convolutional feed-forward network.

3) Overlap-add: The output $\boldsymbol{D}_{B+1}^{AO} \in \mathbb{R}^{E\times S\times P}$ of last DPCT block is used to learn the filters for each source by 2-D convolutional layer. The filters are transformed back to sequence $\boldsymbol{F}_j^{out} \in \mathbb{R}^{(2W+L)\times Z}$ by the overlap-add method. Finally, it is passed to the last layer for further processing, given by

$$\boldsymbol{h}_j = tanh[Conv1d(\boldsymbol{F}_j^{out})] \cdot \sigma[Conv1d(\boldsymbol{F}_j^{out})], \tag{9}$$

where $\boldsymbol{h}_j \in \mathbb{R}^{(2W+1)\times Z}$ is the BF filter of the j-th source.

This article has been accepted for publication in a future proceedings of this workshop, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SiPS55645.2022.9919247, 2022 IEEE Workshop on Signal Processing Systems (SiPS)

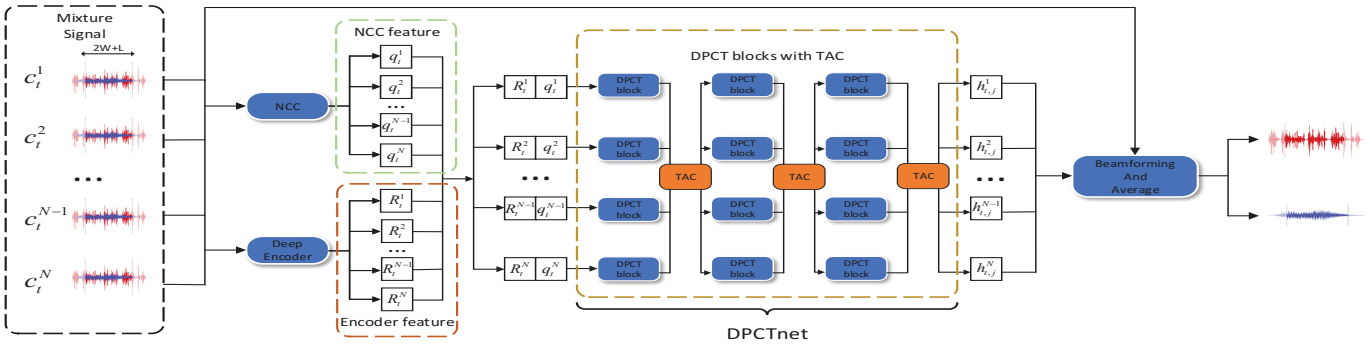

Fig. 1. Flowchart for speech separation with the proposed DE–DPCTnet architecture.

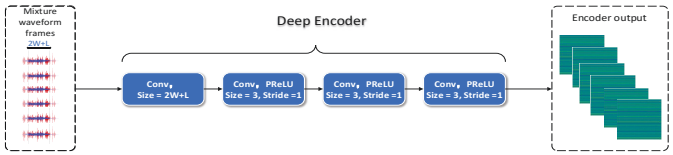

Fig. 2. The deep encoder architecture.

III. Experimental Procedures

A. Dataset

We create a multichannel noisy reverberant dataset of 20000, 5000 and 3000 utterances for training, validation, and test sets, respectively, where two speech signals and one noise signal are randomly selected from the 100-hour Librispeech subset [21] and the Nonspeech Corpus [22], respectively.

We choose a fixed circular six microphones array, where microphones are evenly distributed on a circle with diameter of 10 cm. The speaker locations are uniformly distributed between 0 and 180 degrees. The length and width of testing room are randomly set from 3.0 m and 10.0 m, and the height is randomly sampled between 2.5 m and 4.0 m. The signals are then convolved with room impulse responses generated using gpuRIR toolbox [23]. The T60 reverberation time is randomly sampled from 0.1 to 0.5 s. An overlap ratio between two speakers is uniformly sampled between 0% and 100% for achieving 50% average overlap ratio. The signal-to-noise ratio (SNR) of two speakers is randomly generated from 0 to 5 dB. The SNR between clean signal and noise is randomly sampled between 5 and 15 dB. The noise direction is sampled without further constrains.

B. Model and Training Configurations

In the DE-DPCTnet, the frame size L and the context size W are both set to 16 ms, i.e., 256 samples at 16 kHz sample rate, and the hop ratio H is set to 128, i.e., 50% overlap ratio. The TAC module is the same as in [9]. In the encoder module, the K is set to 256 and the number M of 1-D convolutional layers is set to 3. We use 6 DPCT blocks, i.e., $B = 6$, $E = 64$, $S = 24$ in segmentation stage. Within each DPCT block, local Bi-LSTM is used with 128 hidden units, and the global MHA is set to be 4-head, i.e., $J = 4$.

The models are trained for 100 epochs on 4 s long input sequences with the Adam optimizer [24], and in the process of which, the utterance-level permutation invariant training (uPIT) [25] is also used. The learning rate is initialized to be 0.001 and decays by a factor of 0.98 for every two epochs. The gradient clipping by a maximum gradient ℓ_2-norm of 5 is always applied, and the early stopping is also applied if no best validation model is found for 10 consecutive epochs. The loss function is the scale-invariant source-to-noise ratio (SI-SNR) [26]. The performance is measured by signal-to-distortion ratio improvement (SDRi) [27], as well as the perceptual evaluation of speech quality (PESQ) [28].

IV. Results And Discussions

In this section, we perform an ablation study of the DE-DPCTnet and also compare the performance with representative time-domain end-to-end approaches such as the monaural Conv-Tasnet [4], the monaural DPRNN5, multi-channel Conv-TasNet [29], and the baseline FaSNet-TAC, a representative time-domain beamforming technique extending monaural DPRNN for multi-channel separation. All of them are popular in speaker separation task.

As shown in Table I, we first compared the performance of the methods under different reverberation conditions. It is expected, the performances of all the methods degrade with the increase of T60, since the reverberation distorts the features used for separation, especially for the single-channel system. The multi-channel Conv-TasNet using spatial information ICD yielded more than 2.45 dB SDRi over the reference Conv-TasNet system, suggesting that multi-channel approaches using spatial information can better alleviate the degradation induced by reverberations. By full utilizing the self-attention mechanism and deep encoder, our model has achieved a great improvement in the case of the high reverberations (T60 0.4–0.5s), gaining 2.82 dB SDRi over multi-channel Conv-TasNet and 2.62 dB SDRi over FaSNet-TAC.

Table II shows the experimental results in terms of SDRi. Besides the standard separation measure SDRi, we also analyzed the runtime cost of each model for processing a 4 s

This article has been accepted for publication in a future proceedings of this workshop, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SiPS55645.2022.9919247, 2022 IEEE Workshop on Signal Processing Systems (SiPS)

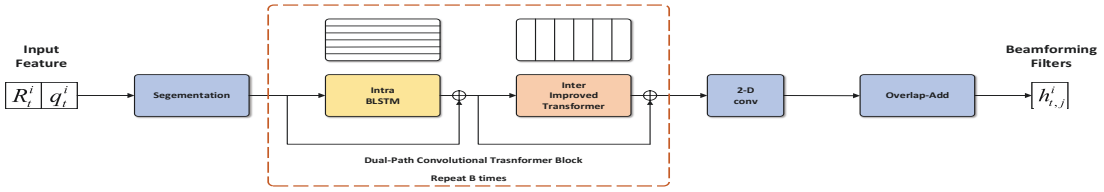

Fig. 3. The framework of dual-path convolutional transformer network.

TABLE I
THE RESULTS OF DIFFERENT NETWORKS IN TERMS OF SDRI (DB) WITH DIFFERENT T60.

Model	Params.	Reverberation(T60)				Average
		0.1–0.2s	0.2–0.3s	0.3–04s	0.4–0.5s	
DPRNN	2.6M	13.33	10.93	9.47	6.05	8.34
Conv-TasNet	5.1M	10.33	8.87	7.95	6.168	7.45
Multi-channel Conv-TasNet	8.8M	14.49	11.98	11.02	7.80	9.90
FaSNet-TAC+16ms	2.9M	15.21	12.57	11.12	7.97	10.2
FaSNet-TAC+4ms	2.9M	15.10	12.63	11.09	8.00	10.91
DPCTnet+16ms	3.2M	16.24	14.00	12.77	9.48	11.62
DPCTnet+4ms	3.2M	16.57	14.41	13.03	9.92	12.02
DPCTnet+DeepEncoder+16ms	3.8M	16.84	14.72	13.22	10.30	12.35
DPCTnet+DeepEncoder+4ms	3.8M	**17.19**	**15.02**	**13.56**	**10.62**	**12.67**

TABLE II
THE RESULTS OF DIFFERENT NETWORKS IN TERMS OF SDRI (DB).

Model	Params.	MACs.	Speaker angle				Overlap ratio				Average
			$< 15°$	15°–45°	45°–90°	$> 90°$	$< 25\%$	25–50%	50–75%	$> 75\%$	
Multi-channel Conv-TasNet	8.8M	28.06G	9.31	9.78	10.07	10.40	14.15	10.75	8.39	6.28	9.89
FaSNet-TAC+16ms	2.9M	14.82G	9.23	10.30	10.51	10.78	14.78	11.03	8.54	6.46	10.2
FaSNet-TAC+4ms	2.9M	55.28G	9.79	11.08	11.28	11.51	15.36	11.74	9.22	7.34	10.91
DPCTnet+16ms	3.2M	15.26G	9.94	11.37	12.29	12.93	15.90	12.54	10.11	7.96	11.62
DPCTnet+4ms	3.2M	60.74G	10.22	11.79	12.76	13.35	16.23	12.92	10.54	8.40	12.02
DPCTnet+DeepEncoder+16ms	3.8M	17.04G	10.19	11.97	13.22	14.07	16.23	13.18	10.94	8.78	12.35
DPCTnet+DeepEncoder+4ms	3.8M	67.64G	10.60	12.33	13.49	14.33	16.84	13.45	11.21	9.2	12.67

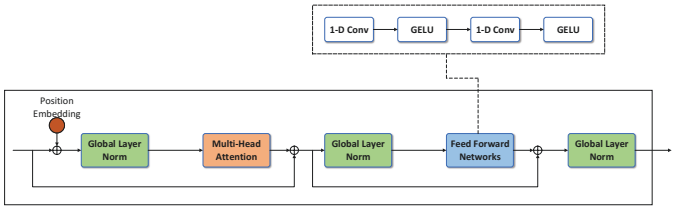

Fig. 4. Architecture of the improved transformer.

mixture input in terms of multi-add operations (MACs) with a third-party module [1], which represents the model efficiency. For FaSNet-based models, the performance improvement of separation is achieved when increasing the encoder window length. For example, the FaSNet-TAC with 4 ms encoder window (55.28G MACs) gains 0.71 dB SDRi over the FaSNet-TAC with 16 ms encoder window (14.82G MACs), which means that the computational complexity of the former is almost four times that of the latter. However, the proposed DPCTnet with 16 ms window (15.26G MACs) achieves a better SDRi than FaSNet-TAC with 4 ms, and the DPCTnet with 4 ms gains absolute 1.11 dB SDRi. This indicates the effectiveness of replacing RNN with transformer modules.

We also perform the ablation experiments with and without nonlinear encoder modules. It is noticed that the proposed network with 16 ms gains 0.71 dB SDRi by using the nonlinear deep encoder, as indicated in Table II. The similar conclusion can be drawn with 4 ms scenario, which indicates that using the nonlinear deep encoder indeed benefits the final separation performance. It is interesting to point out when speaker angle is less than 15°, the proposed model presents an obvious advantage under the same length of encoder window.

In Table III, the PESQ is also utilized as another measure to objectively compare the different models. From Table III, it can be seen the proposed model still achieves a better performance than the FaSNet-TAC, even with a longer window. This further demonstrates the effectiveness of using the transformer module and nonlinear encoder.

[1]https://github.com/sovrasov/flops-counter.pytorch

TABLE III
THE RESULTS OF DIFFERENT NETWORKS IN TERMS OF PESQ.

Model	Params.	MACs.	Speaker angle				Overlap ratio				Average
			$< 15°$	15°–45°	45°–90°	$> 90°$	$< 25\%$	25–50%	50–75%	$> 75\%$	
Multi-channel Conv-TasNet	8.8M	28.06G	2.38	2.40	2.42	2.49	2.82	2.53	2.27	2.06	2.42
FaSNet-TAC+16ms	2.9M	14.82G	2.31	2.41	2.45	2.52	2.82	2.52	2.28	2.08	2.42
FaSNet-TAC+4ms	2.9M	55.28G	2.38	2.49	2.52	2.58	2.88	2.59	2.34	2.16	2.49
DPCTnet+16ms	3.2M	15.26G	2.34	2.46	2.53	2.61	2.88	2.59	2.33	2.13	2.49
DPCTnet+4ms	3.2M	60.74G	2.38	2.51	2.89	2.67	2.92	2.64	2.40	2.19	2.54
DPCTnet+DeepEncoder+16ms	3.8M	17.04G	2.38	2.54	2.64	2.74	2.96	2.67	2.43	2.23	2.57
DPCTnet+DeepEncoder+4ms	3.8M	67.64G	2.43	2.58	2.68	2.78	3.00	2.72	2.48	2.27	2.62

V. CONCLUSIONS

In this paper, we propose a deep encoder dual-path convolutional transformer network for end-to-end multi-channel speech separation. The mixed signal is extracted by the proposed deep encoder and then sent to the DPCTnet to directly estimate the BF filters of each channel. In the experiments, we demonstrate the benefits of using the transformer module and deep encoder module. Compared with other methods, the results also show that the proposed model achieves a better performance than the FaSNet-TAC on noisy reverberant speech separation task in terms of SDRi and PESQ metrics.

REFERENCES

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

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

[3] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*. IEEE, 2018, pp. 696–700.

[4] ——, "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 *ICASSP*. IEEE, 2020, pp. 46–50.

[6] J.-J. Chen, Q.-R. 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.

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

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

[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*. IEEE, 2016, pp. 196–200.

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

[12] X. Xiao, S.-K. Zhao, D. L. Jones, E. S. Chng, and H. H.-Z. Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *ICASSP*. IEEE, 2017, pp. 3246–3250.

[13] Z. Chen, X. Xiao, T. Yoshiokaand, H. Erdogan, J.-Y. Li, and Y.-F. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[14] K.-Z. Qian, Y.Zhang, S.-Y. Chang, X.-S. Yang, D. Florencio, and M. Hasegawa-Johnson, "Deep learning based speech beamforming," in *ICASSP*. IEEE, 2018, pp. 5389–5393.

[15] 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*. IEEE, 2020, pp. 6384–6388.

[16] X. Xiao, S. Watanabe, E. S. Chng, and H.-Z. Li, "Beamforming networks using spatial covariance features for far-field speech recognition," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[17] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, and K. Chin, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[18] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *ICASSP*. IEEE, 2017, pp. 271–275.

[19] M. J. Jo, G.-W. Lee, J. M. Moon, C.-S. Cho, and H. K. Kim, "Estimation of mvdr beamforming weights based on deep neural network," in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[20] J. Heitkaemper, D. Jakobeits, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying tasnet: A dissecting approach," in *ICASSP*. IEEE, 2020, pp. 6359–6363.

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

[22] G. Hu, "100 nonspeech sounds," *Online: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html*, 2006.

[23] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

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

[25] 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.

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

[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] 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.

[29] 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*. IEEE, 2020, pp. 7319–7323.