MULTICHANNEL SPEECH ENHANCEMENT WITHOUT BEAMFORMING

Ashutosh Pandey[1*], *Buye Xu*[1], *Anurag Kumar*[1], *Jacob Donley*[1], *Paul Calamia*[1] *and DeLiang Wang*[2]

[1]Facebook Reality Labs Research, USA
[2]Department of Computer Science and Engineering, The Ohio State University, USA

ABSTRACT

Deep neural networks are often coupled with traditional spatial filters, such as MVDR beamformers for effectively exploiting spatial information. Even though single-stage end-to-end supervised models can obtain impressive enhancement, combining them with a traditional beamformer and a DNN-based post-filter in a multi-stage processing provides additional improvements. In this work, we propose a two-stage strategy for multi-channel speech enhancement that does not require a traditional beamformer for additional performance. First, we propose a novel attentive dense convolutional network (ADCN) for estimating real and imaginary parts of complex spectrogram. ADCN obtains state-of-the-art results among single-stage models. Next, we use ADCN with a recently proposed triple-path attentive recurrent network (TPARN) for estimating waveform samples. The proposed strategy uses two insights; first, using different approaches in two stages; and second, using a stronger model in the first stage. We illustrate the efficacy of our strategy by evaluating multiple models in a two-stage approach with and without a traditional beamformer.

Index Terms— multi-channel, two-stage, waveform mapping, complex spectral mapping, fixed array

1. INTRODUCTION

Multi-channel speech enhancement is the task of removing noise, intereference and reverberation from a degraded speech signal by utilizing recordings from multiple microphones. Traditional approaches use linear spatial filters, such as those from a minimum-variance distortionless-response (MVDR) optimization, to preserve signal from the target source and suppress all other signals in the space [1]. In recent years, supervised speech enhancement using deep neural networks (DNNs) has become the mainstream methodology for speech enhancement [2].

For multi-channel processing, DNNs are generally incorporated with traditional spatial filters [3]–[5], where the role of DNN is to provide better estimates of speech and noise statistics for the spatial filter. Another general approach is to train DNNs with spatial features, such as inter-channel phase, time, and level differences [6], [7]. A DNN trained with spatial features is expected to exploit spatial cues for improved discrimination between target and interference. In recent times, end-to-end supervised approaches without any explicit spatial filtering have obtained impressive results [8]–[13]. The goal of end-to-end approaches is to make the spatial filtering an implicit part of supervised learning.

Even though these end-to-end supervised approaches have shown impressive enhancement performance, they are yet to be widely accepted. This is due to a confounding empirical finding that an end-to-end supervised model when combined with a traditional spatial filter, such as an MVDR beamformer, and a DNN-based post-filter, provides superior results compared to a DNN-only single-stage or multistage processing [8], [14]–[16]. For example, study in [8] obtained impressive performance by training a dense convolutional recurrent network (DCRN) for multi-channel complex spectral mapping. However, the performance was further improved by using an MVDR beamformer along with a following DCRN as the post-filter.

The effectiveness of an MVDR beamformer even with strong DNN models can be attributed to the fact that DNNs introduce non-linear distortions in the enhanced speech, which are removed or reduced when they are combined with a distortionless beamformer. As a result, many of the approaches based on end-to-end learning have been inspired from traditional beamformers [17], [18]. The computation of beamformer weights requires matrix inversion, which makes end-to-end learning unstable. A widely accepted strategy to avoid training instability is diagonal loading [19]. A recent study used recurrent neural networks for directly estimating the matrix inverse [18].

In this work, we argue that the use of a traditional beamformer with a DNN is not necessary to obtain distortionless speech enhancement. We propose a novel two-stage approach where both stages are based on neural networks. Our two-stage scheme for multi-channel speech enhancement uses two key strategies. The first strategy is to use two different approaches for speech enhancement in two stages. For instance, one stage might rely on complex spectral mapping, an approach that estimates real and the imaginary parts of complex spectrogram, and the other stage may use waveform mapping where direct waveform to waveform enhancement is done. We believe that complex spectral mapping and waveform mapping complement each other in terms of removing the overall model bias. In other words, they can get rid of some component of each other's distortions, and hence provide an overall system with fewer distortions.The second strategy is to use the stronger approach out of the two, in the first stage of the two-stage processing.

To this end, we first propose a novel attentive dense convolutional network (ADCN) for multi-channel complex spectral mapping. Similar to a waveform mapping based model in [20], ADCN is an encoder-decoder based UNet architecture where layers within the encoder and decoder are augmented with dense blocks and attention blocks for context aggregation. ADCN obtains state-of-the art results among single-stage systems.

Next, we evaluate different models in a two-stage scheme with and without an MVDR beamformer. We empirically study two complex spectral mapping models: ADCN and DCRN from [8], and one waveform mapping model triple-path attentive recurrent network (TPARN) recently proposed in [12].

[*]Work done during internship at Facebook Reality Labs Research.

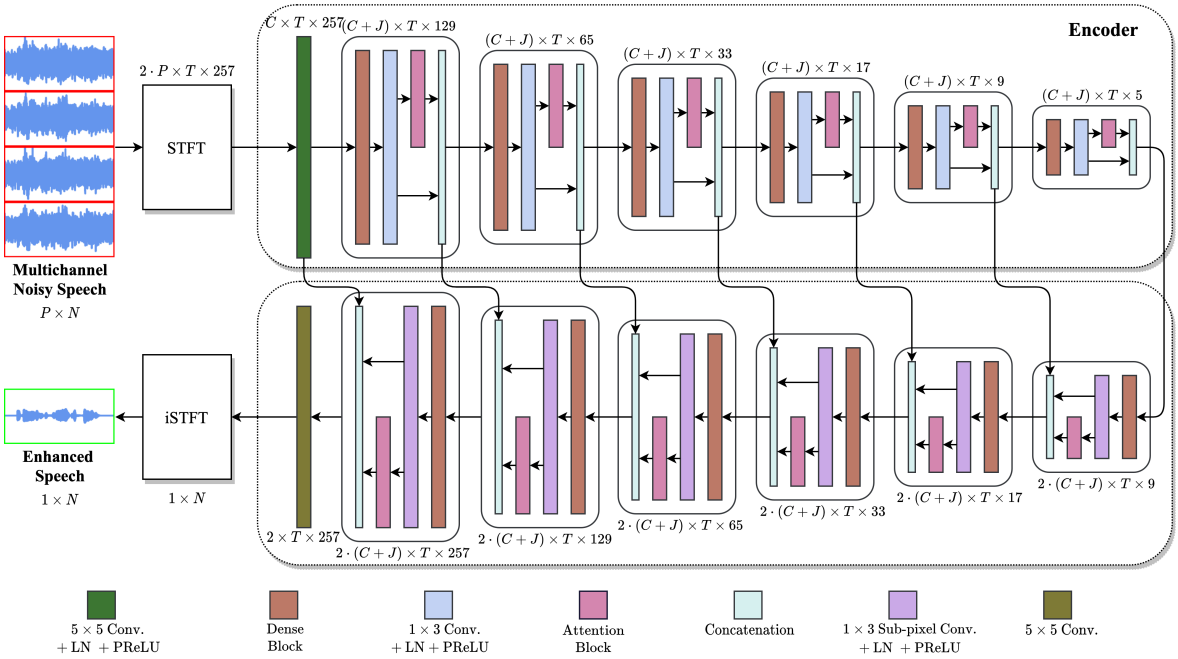


Fig. 1. The proposed ADCN for multi-channel complex spectral mapping.

Our experimental results indicate that an MVDR beamformer becomes redundant when two different approaches are used in two stages. Also, we obtaine significantly better results when a stronger model in the first stage is followed by a relatively weaker model in the second stage.

2. PROBLEM DEFINITION

A multi-channel noisy speech $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P] \in \mathbb{R}^{P \times N}$ with N samples and P microphones is modeled as

$$\begin{aligned} x_p(n) &= y_p(n) + z_p(n) \\ &= h_p(n) * s(n) + z_p(n) \\ &= (h_p^d(n) + h_p^r(n)) * s(n) + z_p(n)] \\ &= h_p^d(n) * s(n) + [h_p^r(n) * s(n) + z_p(n)] \\ &= d_p(n) + [r_p(n) + z_p(n)] \\ &= d_p(n) + u_p(n) \end{aligned} \tag{1}$$

where $p = 1, 2, \ldots, P$, $n = 0, 1, \ldots N-1$. $\boldsymbol{s}$ is the source speech, $\boldsymbol{y}_p$ and $\boldsymbol{z}_p$ are respectively the reverberated speech and noise received at microphone p. $*$ denotes convolution operator and $\boldsymbol{h}$ is the room impulse response (RIR) of source speech. $\boldsymbol{h}^d$ is the direct-path RIR and $\boldsymbol{h}^r$ is the reverberation-path RIR of the source speech. $\boldsymbol{u}$ is the overall interference including noise and room reverberation. A multi-channel speech enhancement algorithm aims at obtaining a good estimate $\hat{\boldsymbol{d}}_r$ of the direct-path speech at a reference microphone r from multi-channel noisy recording $\boldsymbol{x}$.

3. ATTENTIVE DENSE CONVOLUTIONAL NETWORK

The architecture of the proposed ADCN is shown in Fig. 1. It is a UNet architecture with an encoder and a decoder. The input to ADCN, $\boldsymbol{X} = \text{STFT}(\boldsymbol{x})$, is of shape $2 \cdot P \times T \times 257$ with T frames. It is transformed to shape $C \times T \times 257$ using a 5×5 convolutional layer with layer normalization (LN) and parametric ReLU (PReLU). Next, it is processed using a stack of 6 encoder blocks and 6 decoder blocks. The output of a decoder block is concatenated with the output from a corresponding symmetric block in the encoder. The final output is computed by a 5×5 convolutional layer with 2 output channels. The output waveform is obtained using an inverse STFT (iSTFT) layer at the output.

The encoder block comprises a stack of a dense block, a 1×3 convolutional block using a stride of 2 for downsampling with LN and PReLU, and an attention block. The output of the attention block is concatenated with its input to get the final output. The decoder block is similar to the encoder block except that it uses 1×3 sub-pixel convolution for upsampling [20] in the place of strided convolution for downsampling.

The architecture of the dense block is shown in Fig. 2. It comprises a stack of five 3×3 convolutional layer with C output channels, LN and PReLU. The input to a given convolutional layer in a dense block is a concatenation of the block input and outputs from preceding convolutional layers in the block.

The architecture of the attention block is shown in Fig. 3. An input of shape $C \times T \times L$ is first transformed using three separate

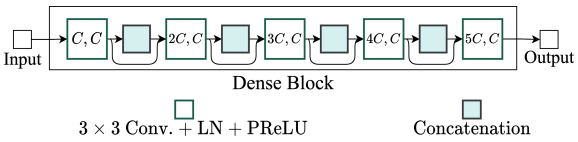


Fig. 2. Dense block in ADCN. a and b in a, b inside a box respectively denote the number of input and output channels.

1×1 convolutional layers to get query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ of shapes $E \times T \times L$, $E \times T \times L$, and $J \times T \times L$ respectively and then rearranged to 2d tensors of shapes $T \times E \cdot L$, $T \times E \cdot L$, and $T \times J \cdot L$. Next, the output from attention is computed as $\boldsymbol{A} = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V}$. Finally, $\boldsymbol{A}$ is rearranged to a 3d tensor of shape $J \times T \times L$.

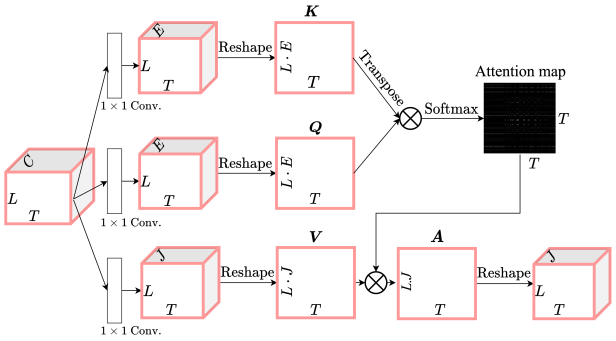


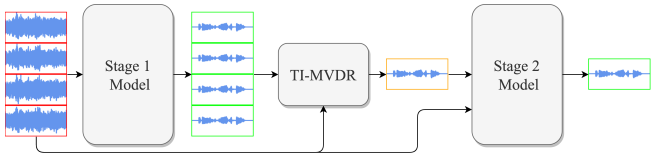
Fig. 3. Attention block in ADCN.

4. TWO-STAGE MULTICHANNEL SPEECH ENHANCEMENT

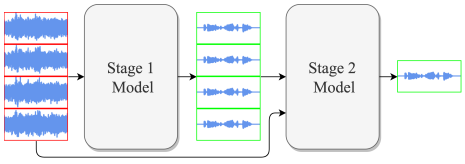
We evaluate three models: DCRN, ADCN and TPARN with the following approaches to two-stage processing.

4.1. Two-stage Approach with a Beamformer

The two-stage approach with an MVDR beamformer is shown in Fig. 4 (a). First, a DNN is trained to estimate enhanced speech at all channels. TPARN can output enhanced signals at all channels simultaneously as it is a multiple-input and multiple-output (MIMO) model. DCRN and ADCN, on the other hand, are multiple-input and single-output (MISO) model, and hence, require enhancing all channels independently by running the enhancement model P times for P channels. The output for the m^{th} microphone is computed by using a circularly shifted input $[\boldsymbol{x}_m, \boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_{m-2}, \boldsymbol{x}_{m-1}]$ [8]. This strategy works because we use a symmetric circular microphone array.



(a) The two-stage approach with a beamformer.



(b) The two-stage approach without a beamformer.

Fig. 4.

Table 1. Model comparisons for single-stage multi-channel speech enhancement.

Test Dataset	WSJCAM0			DNS		
Test Metric	SI-SDR	STOI	PESQ	SI-SDR	STOI	PESQ
Unprocessed	-3.8	70.9	1.63	-7.6	63.8	1.38
DCRN	9.4	96.5	3.31	4.6	90.1	2.57
TPARN	10.4	96.9	**3.43**	**8.4**	91.9	2.75
ADCN	**12.0**	**97.3**	3.42	7.8	**92.3**	**2.84**

Next, enhanced speech at all channels are used to estimate the coefficients of a time-invariant MVDR (TI-MVDR) beamformer using following equations.

$$\hat{\Phi}^{(d)}(f) = \frac{1}{T}\sum_{t=1}^{T} \hat{D}(t,f)\hat{D}(t,f)^H$$
$$\hat{\Phi}^{(u)}(f) = \frac{1}{T}\sum_{t=1}^{T} \hat{U}(t,f)\hat{U}(t,f)^H \quad (2)$$

where $\hat{\boldsymbol{D}} = \text{STFT}(\hat{\boldsymbol{d}}) \in \mathbb{C}^{P \times T \times F}$ has T frames and F frequency bins, $\hat{\boldsymbol{U}} = \boldsymbol{X} - \hat{\boldsymbol{U}}$, and $\hat{\boldsymbol{D}}(t,f)$ is the value at frame t and frequency bin f.

In our experiments, sound sources are assumed to be static within each utterance, therefore, time-invariant MVDR (TI-MVDR) is a better choice than time-varying beamformer [16]. The relative transfer function with respect to the reference microphone is computed as

$$\hat{\boldsymbol{c}}_r(f) = \mathbb{P}\{\hat{\Phi}^d(f)\}/\mathbb{P}\{\hat{\Phi}^d_r(f)\} \quad (3)$$

where $\mathbb{P}$ extracts the principal eigenvector and $\hat{\Phi}^d_r(f)$ is the r^{th} component of $\hat{\Phi}^d(f)$. The MVDR beamformer is computed as

$$\hat{\boldsymbol{w}}_r(f) = \frac{\hat{\boldsymbol{\Phi}}^u(f)^{-1}\hat{\boldsymbol{c}}_r(f)}{\hat{\boldsymbol{c}}_r(f)^H\hat{\boldsymbol{\Phi}}^u(f)^{-1}\hat{\boldsymbol{c}}_r(f)} \quad (4)$$

The beamformer output is computed as

$$\hat{BF}_r(t,f) = \hat{\boldsymbol{w}}_r(f)^H \boldsymbol{X}(t,f) \quad (5)$$

A waveform from the beamformer output is obtained as

$$\hat{\boldsymbol{b}}_r = \text{iSTFT}(\hat{\boldsymbol{B}}\hat{\boldsymbol{F}}_{\boldsymbol{r}}) \quad (6)$$

Finally, a second DNN model is trained to map the speech from beamformer and the noisy multi-channel speech to enhanced speech. The input to DCRN and ADCN is a concatenation of $\hat{\boldsymbol{b}}_r$ and $\boldsymbol{x}$ along the channel dimension in STFT. The input to TPARN is a concatenation of $\hat{\boldsymbol{b}}$ and $\boldsymbol{x}$ along the frame dimension as TPARN requires a sequential input across channels [12], [13].

4.2. Two-stage Approach without a Beamformer

The two-stage approach without a beamformer is shown in Fig. 4 (b). In this approach, a DNN is trained first to get an estimate of enhanced speech at all channels and then an another DNN is trained to map enhanced speech and noisy speech at all channels to the enhanced speech at the reference channel. The input to DCRN and ADCN is a concatenation of $\hat{\boldsymbol{d}}$ and $\boldsymbol{x}$ along the channel dimension in STFT. The input to TPARN is a concatenation of $\hat{\boldsymbol{d}}$ and $\boldsymbol{x}$ along the frame dimension.

Table 2. Model comparisons for two-stage multi-channel speech enhancement with and without beamforming.

Test Dataset			WSJCAM0			DNS		
Stage1↓	Stage2 ↓	Type ↓	SI-SDR	STOI	PESQ	SI-SDR	STOI	PESQ
Unprocessed		×	-3.8	70.9	1.63	-7.6	63.8	1.38
DCRN	DCRN	**(a)**	10.7	97.1	3.43	5.6	91.6	2.69
		(b)	9.9	96.8	3.46	6.9	91.0	2.64
	TPARN	(a)	11.4	97.2	3.51	7.3	90.9	2.63
		(b)	11.1	97.3	3.56	8.3	92.2	2.80
	ADCN	**(a)**	12.5	97.4	3.44	8.0	93.0	2.87
		(b)	11.2	97.1	3.47	7.5	91.5	2.73
TPARN	DCRN	(a)	11.2	97.2	3.45	6.7	92.0	2.73
		(b)	12.3	97.5	3.55	9.2	93.0	2.85
	TPARN	**(a)**	11.3	97.2	3.52	8.1	91.8	2.71
		(b)	12.1	96.9	3.47	8.5	92.0	2.76
	ADCN	**(a)**	12.9	97.4	3.42	8.9	93.5	2.95
		(b)	12.3	97.5	3.51	9.6	93.2	2.92
ADCN	DCRN	(a)	10.9	97.0	3.43	6.6	92.0	2.75
		(b)	12.7	97.5	3.47	8.6	92.9	2.85
	TPARN	(a)	11.8	97.2	3.50	7.9	91.4	2.67
		(b)	**13.8**	**98.0**	**3.64**	**10.0**	**93.7**	**2.99**
	ADCN	**(a)**	12.7	97.5	3.47	8.5	93.4	2.93
		(b)	12.4	97.5	3.48	8.9	92.9	2.89

5. EXPERIMENTS

5.1. Experimental Settings

We use a four-microphone circular array of radius of 10 cm with equal spacing between microphones. All the models are trained and evaluated on two different datasets. The first dataset is created using speakers from the WSJCAM0 [21] dataset and noises from the REVERB challenge [22]. A uniform T60 from $[0.2, 1.2]$ seconds is used for reverberation and a uniform SNR from $[5, 20]$ dB is used for noise. An algorithm for generating this dataset is given in [8]. The second dataset is created from the DNS 2020 corpus[1] [23]. For this dataset, T60 is used from $[0.2, 1.2]$ seconds and SNR is used from $[-10, 10]$ dB. The DNS dataset can be considered more challenging than the WSJ0CAM / REVERB dataset as it uses diverse and difficult non-stationary noises with low SNR values. The data generation algorithm for the DNS dataset is given in [12].

All the utterances are resampled to 16 kHz. We use a frame size of 32 ms, frame shift of 8 ms, $C = 64, E = 5$, and $J = 32$ for ADCN. TPARN and DCRN are trained using methods proposed in their original studies. ADCN is trained using a phase constrained magnitude (PCM) loss [20]. All the models are trained for 100 epochs with a batch size of 8 4-s long utterances randomly extracted at the training time. The initial learning rate is set to 0.0004 and is scaled by half if the validation score does not improve for five consecutive epochs.

The first microphone, $r = 1$, is used for objective evaluation. All the models are evaluated using short-time objective intelligibility (STOI) [24], perceptual evaluation of speech quality (PESQ) [25], and scale-invariant signal-to-distortion ratio (SI-SDR). STOI is reported in percentage.

5.2. Experimental Results

First, we compare DCRN, ADCN and TPARN in Table 1 for single-stage end-to-end training. ADCN obtains best results in all the cases except for SI-SDR at DNS where it is slightly worse that TPARN. A general performance order of these models is DCRN $<$ TPARN $<$ ADCN. Even though TPARN is worse than ADCN, it has computational advantages over ADCN and DCRN. For example, using ADCN in the first stage requires P forward passes for P channels, whereas TPARN can enhance signals at all channels in one pass.

Next, we compare two approaches: (a) two-stage with a beamformer, (b) two-stage without a beamformer. We consider an overall improvement over two datasets between (a), (b) and make bold the one with better improvements. We highlight both (a) and (b) if their performances are similar.

Firstly, we observe that with a complex spectral mapping model DCRN in the first stage, beamformer is better for complex spectral mapping models DCRN and ADCN, but worse for the waveform mapping model TPARN in the second stage. This suggest that for a weaker complex spectral mapping model, beamformer is helpful but only for models with the same approach, complex spectral mapping, in the second stage.

Next, we observe that with TPARN in the first stage, beamformer is worse for DCRN and and similar for ADCN and TPARN in the second stage. This suggest that for a relatively stronger model TPARN, beamformer does not provide any consistent improvement.

Further, we can see that with ADCN in the first stage, beamformer obtains worse results with DCRN and TPARN and comparable results with ADCN. These comparisons indicate that the beamformer is helpful only if the first stage model is weak and the second stage model uses same approach as the first stage model.

Finally, we note that the best results are reported with TPARN followed by DCRN for the pair (TPARN, DCRN) and with ADCN followed by TPARN for the pair (TPARN, ADCN). This suggests that a much better performance can be obtained without a beamformer by using different approaches, complex spectral mapping and waveform mapping in two stages and employing the stronger model in the first stage. Also, we find that ADCN followed by TPARN obtains significantly better results than the second best.

6. CONCLUSIONS

We have proposed a novel attentive dense convolutional network for multi-channel speech enhancement. We have also proposed a two-stage approach that obtains excellent results without a beamformer. The proposed approach uses ADCN for complex spectral mapping in the first stage and TPARN for waveform mapping in the second stage. Future research includes evaluating the effectiveness of the proposed approach for ASR improvements.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[2] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1702–1726, 2018.

[3] H. Erdogan *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," in *INTERSPEECH*, 2016, pp. 1981–1985.

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

[1] https://github.com/microsoft/DNS-Challenge/blob/master/LICENSE

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 692–730, 2017.

[6] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 457–468, 2018.

[7] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP*, 2018, pp. 1–5.

[8] Z.-Q. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *ICASSP*, 2020, pp. 486–490.

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

[10] B. Tolooshams *et al.*, "Channel-attention dense U-Net for multichannel speech enhancement," in *ICASSP*, 2020, pp. 836–840.

[11] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *ICASSP*, 2021, pp. 3415–3419.

[12] A. Pandey *et al.*, "TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," *arxiv:2110.10757*, 2021.

[13] ——, "TADRN: Triple-attentive dual-recurrent network for ad-hoc array multichannel speech enhancement," *arxiv:2110.11844*, 2022.

[14] Z.-Q. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 941–950, 2020.

[15] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

[16] ——, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[17] Z.-Q. Wang and D. Wang, "All-neural multi-channel speech enhancement.," in *INTERSPEECH*, 2018, pp. 3234–3238.

[18] Z. Zhang *et al.*, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *ICASSP*, IEEE, 2021, pp. 6089–6093.

[19] X. Mestre and M. A. Lagunas, "On diagonal loading for minimum variance beamformers," in *International Symposium on Signal Processing and Information Technology*, IEEE, 2003, pp. 459–462.

[20] A. Pandey and D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1270–1279, 2021.

[21] T. Robinson *et al.*, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 1, 1995, pp. 81–84.

[22] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, pp. 1–19, 2016.

[23] C. K. Reddy *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv:2005.13981*, 2020.

[24] 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*, pp. 2125–2136, 2011.

[25] 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 *ICASSP*, 2001, pp. 749–752.