# DMANET: DEEP LEARNING-BASED DIFFERENTIAL MICROPHONE ARRAYS FOR MULTI-CHANNEL SPEECH SEPARATION

*Xiaokang Yang* [1], *Jianguo Wei* [1]*

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China

## ABSTRACT

In this paper, we develop a novel differential microphone arrays network (DMANet) for solving the multi-channel speech separation problem. In DMANet we explore a neural network combined to differential microphone arrays (DMAs) beamforming technique. Specifically, a sequence of differential operation is introduced alternately into network. Based on the filter-and-sum network (FaSNet), we show how DMANet significantly improves the separation performance. Numerical experiments demonstrate that the proposed network has a clearly advantageous improvement on SI-SNR with a smaller model.

***Index Terms***— Differential microphone arrays, speech separation, deep learning, beamforming

## 1. INTRODUCTION

The single-channel or multi-channel speech separation is a basic task with a wide range of applications, including automatic speech recognition (ASR) and speech communication conference system, mobile communication, etc. With the advent of deep learning, the performance of single speech separation [1–3] under near-field with clear-talking scenarios is significantly improved. However, in smart speaker or public surveillance scenarios, the speech is recorded in a far-field setup where the microphone distance from signal sources can be greater than 1m. In this case, the task is much more challenging [4].

Microphone array (MA) beamforming is widely used to recover a speech signal of interest from noisy observations in various human-machine interfaces and voice communications [5, 6]. Compared to conventional beamformer such as filter-and-sum filter, differential beamformer with MA can form frequency-invariant beampatterns and has the potential to attain high directional gains, and consequently it has been widely investigated [7–9].

On the other hand, the result of DMAs beamformers can be degraded by several factors [10, 11]. 1) white noise amplification, as a result of large norm of filter coefficients. 2) limited steering ability. 3) irregularity of the beampatterns and the directivity factor at some frequencies. To this regard, robust DMA beamformers have been designed to solve these issues [10, 12, 13], and the realization of manufacture of microphone arrays remains to be explored.

Deep learning-based beamforming systems, sometimes called neural network beamformers, have been an active research topic in recent years [14]. Most neural network beamformers can be broadly categorized into two main categories [15]. The first category aims at learning a set of beamforming filters to replace conventional filter-and-sum (FaS) or minimum variance distortionless response (MVDR) in either the time domain or frequency domain [16–19]. The second category, which refers to as the regression-based approach, incorporates beamforming without explicitly generating the beamforming filters, but just with neural network [20].

Previous studies have shown that DMAs beamformers and neural network beamformers have significantly success. In practical applications, it is still difficult to control the white noise gain for DMAs beamformers and neural network beamformers also have the disadvantage of low-latency. FasNet-TAC proposed in [14] is suitable for realtime and low-latency applications, however the performance of FasNet-TAC may be limited by FaS compared to DMAs beamformers. To address the above limitation of previous neural network beamformers and DMAs beamformers, we propose time-domain adaptive differential microphone arrays network (DMANet) with taking full advantage of DMAs and FaS beamformers. The analysis of mechanism of DMANet is also provided and how DMANet achieves a state-of-the-art performance is investigate.

The remainder of this paper is organized as follows. The signal model and the classical DMAs beamformers are established in Section 2. In Section 3, we present the proposed DMANet model. Subsequently, the experimental setup is evaluated in Section 4 and the corresponding results are discussed in Section 5. Finally, we conclude the paper in Section 6.

## 2. DIFFERENTIAL MICROPHONE ARRAYS

### 2.1. Signal Model

Generally speaking, a microphone array refers to a sound collection system that uses multiple microphones to sample a spatially diverse sound field [21]. As these microphones

Thanks to NSFC of China (No. U1936102, No. 61876131), Key RD Program of Tianjin (No. 19ZXZNGX00030) for funding.

are designed into a particular geometry, we need a reference point or microphone to set up the coordinate system in which a back-end processor can regularly process microphone signals. This kind of arrays, when combined with proper signal processing algorithms which can take advantage of the spatial information, can be used to solve many important acoustic problems such as speech enhancement, dereverberation, speech separation, etc. Here, we consider the uniform linear array(ULA) of $M$ microphones signal model shown in Figure 1, and without loss of generality, we show how to design conventional DMA beamformer.

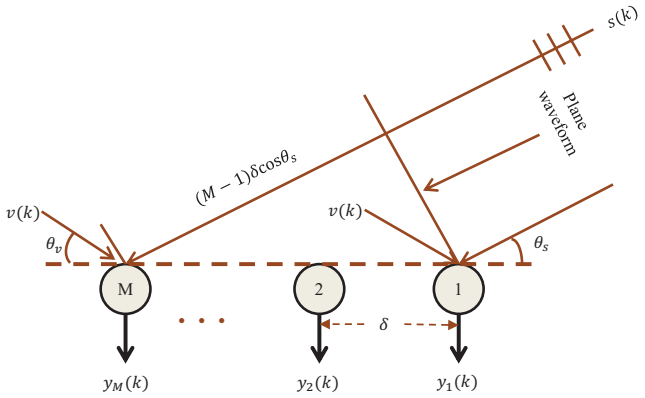



**Fig. 1**. Illustration of a uniform linear microphone array. $s(k)$ and $v(k)$ are , respectively, the desired source signals and noise signals.

### 2.2. Conventional DMA Beamformer

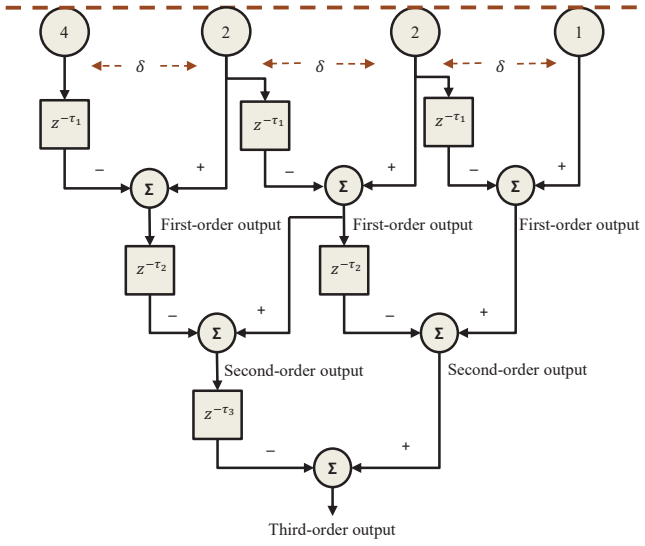



**Fig. 2**. Conventional structure of first-, second-, and third-order DMAs.

Differential microphone arrays (DMAs) can be understood as some processors and the response based on arrays to the acoustic pressure field with the spatial derivatives. To form a directional pattern, one would need to measure the differentials of the acoustic pressure field, which can be achieved by combining the outputs of a number of omnidirectional sensors [22].

Figure 2 illustrates how first-, second-, and third-order DMAs are constructed with a linear geometry. As the same as filter-and-sum beamformers, DMA beamformers can realize the function of adaptive filtering, which just needs to change the delay $\tau_m$ dynamically. However, unlike the additive arrays in which the delay $\tau_m$ is related to desired signals, the delay $\tau_m$ is related to noise signals in differential arrays. Considering the general case where the signal of interest comes from the direction $\theta_s$, we can express the observation signal vector as

$$\begin{aligned}\mathbf{y}(w, cos\theta_s) &\triangleq [Y_1(w)Y_2(w)...Y_M(w)]^T \\ &= \mathbf{d}(w, cos\theta_s)\mathbf{X}(w) + \mathbf{v}(w)\end{aligned} \tag{1}$$

where $\mathbf{d}(w, cos\theta)$ is the steering vector. Applying a spatial filter $\mathbf{h}(w)$ to $\mathbf{y}(w)$, we can obtain the corresponding beamformer defined by

$$\mathbf{Z}(w) = \mathbf{h}^H(w)\mathbf{y}(w) \tag{2}$$

where the superscript $^H$ denotes the conjugate-transpose operator. It should be noted that the spatial filter $\mathbf{h}(w)$ satisfies the following constraints

$$\mathbf{d}^H(w, cos\theta_s)\mathbf{h}(w) = 1 \tag{3}$$

$$\mathbf{d}^H(w, \theta_\ell)\mathbf{h}(w) = 0, \quad \ell \neq s. \tag{4}$$

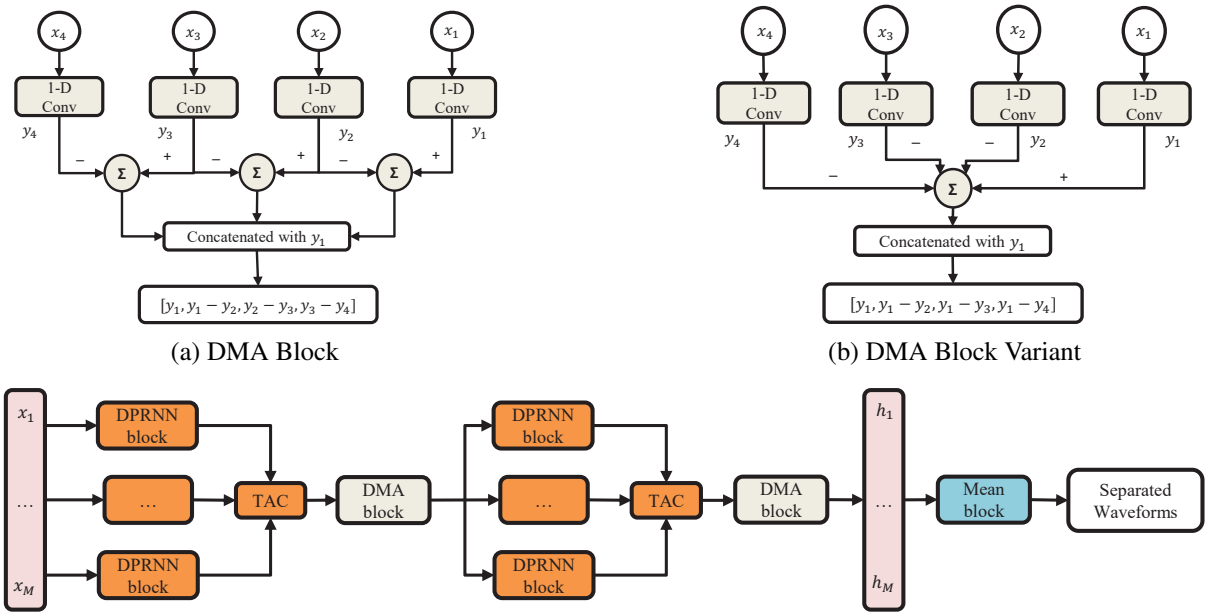
So the beampattern of first-order DMAs has one nulls at $\theta_v$ and $N$-order DMAs have $N$ distinct.

## 3. DMANET MODEL

The objective of this paper is to introduce a novel differential microphone arrays network (DMANet) to estimate a set of spatial filter coefficients. A major advantage of DMANet is that it can take advantage of DMAs' high directional gain, avoid the the unpredictability and improve the adaptability of the system. In this section, an first-order differential network is analyzed for illustrating our motivation and then an Nth-order differential network is discussed. Finally, the DMANet model is proposed.

### 3.1. First-order differential network

The proposed differential block is depicted in Figure 3 (a). Here, we also use four microphones in consistency with the Figure 2. Note that it keeps the reference channel in the model output. The main innovation is that the first-order differential




(a) DMA Block

(b) DMA Block Variant

(c) DMANet Block Diagram

**Fig. 3**. System flowchart for the proposed DMANet system.

network employ the 1-D convolution network instead of the delay operation in the first-order DMAs. Considering the diversity of DMAs, we show a DMA block variant in Figure 3 (b), in which the differentials can be approximated by estimating the difference between the reference microphone and the other microphones.

### 3.2. $N$th-order differential network

Motivated by the Multi-RNN and $N$th-order DMA, we propose $N$th-order differential network, in which DMA block is repeated $N$ times and the output of every layer can be formulated as follows:

$$h_m^i = \begin{cases} y_m^i & m = 0 \\ y_{m-1}^i - y_m^i & 0 < m < M, \text{for DMA block} \\ y_0^i - y_m^i & 0 < m < M, \text{for DMA block variant} \end{cases} \quad (5)$$

Here, $y_m^i$ is the output of $i$th layer 1-D convolution network for $m$th microphone. For simplicity, we call the $N$th-order differential network DMANet. It should be pointed out that the 1-D convolution network mainly depend on local information and therefore it may fail to get the long-term dependence included in the filter estimation.

### 3.3. DMANet

DPRNN has achieved excellent performance in single channel speech separation [3]. Based on this model and the transform-average-concatenate(TAC) with the normalized crosscorrelation feature (NCC), FaSNet-TAC can perform best in some multi-channel datasets [14]. Considering the same multi-channel output of DMA blocks, we apply FaSNet-TAC to DMA blocks by corresponding the blocks behind FaSNet-TAC. Figure 3 (c) shows our proposed DMANet. In our DMANet, segmentation and overlap-add stage [3] are necessary and encoder-decoder based on 1-D Conv are incorporated into DMANet, but the number of kernels can be a little less. All $h_m$ are averaged and convolved with their corresponding channels $x_m$ and it results in following DMNet output:

$$Y_c = \sum_{m=1}^{M} x_m \circledast \overline{h}_m, c = 1, ..., C \quad (6)$$

where C is the number of source and $\overline{h}_m$ is the average of $h_m$.

## 4. EXPERIMENTAL PROCEDURES

### 4.1. Dataset

We evaluate our approach by conducting the task of multi-channel two-speaker noisy speech separation with fixed geometry 6 microphone arrays. Two speakers and nonspeech noise are randomly selected from the 100-hour Librispeech dataset and 100 Nonspeech Corpus, respectively. An overlap ratio between the two speakers is uniformly sampled between 0% and 100% such that the average overlap ratio across the dataset is 50%. Other information about dataset can be found in [14].

**Table 1**. Experment results on 6-mic circular array. SI-SNRi is reported on decibel scale.

| Model | # of param. | Speaker angle | | | | Overlap ratio | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $<15°$ | 15-45° | 45-90° | $>90°$ | $<25\%$ | 25-55% | 55-75% | $>75\%$ | |
| FaSNet-TAC | 2.9 | 9.1 | 11.1 | 12.6 | 13.4 | 15.6 | 12.4 | 10.1 | 8.0 | 11.5 |
| DMANet-Only | 1.27 | 2.1 | 2.6 | 3.1 | 3.2 | 4.8 | 3.4 | 2.0 | 1.9 | 2.6 |
| DMANet-Variant | 1.27 | 2.1 | 2.7 | 3.2 | 3.6 | 4.9 | 3.3 | 2.3 | 2.0 | 2.9 |
| DMANet-Pure | 2.76 | 9.4 | 11.2 | 12.6 | 13.6 | 15.8 | 12.4 | 10.1 | 8.1 | 11.7 |
| **DMANet** | **2.76** | **9.6** | **11.7** | **12.8** | **14.2** | **16.5** | **13.5** | **11.1** | **8.5** | **12.1** |

### 4.2. Model configurations

We evaluate the proposed DMANet on various types of DMA blocks as shown in 3.3. To be specific, we design many different variants of DMANet:

1. **DMANet-Only**: Only DMA Block is repeated in DMANet system and kernel size of 1-D convolutional block is 3.
2. **DMANet-Variant**: Only DMA Block Variant is reapted in DMANet system and the kernel size is also fixed to 3.
3. **DMANet-Pure**: DMANet-Pure consists of DPRNN and DMA Block with kernel size 1.
4. **DMANet**: DMANet is the proposed model in which DPRNN and DMA Block Variant with kernel size 1 are combined in DMANet system.

In the training stage, the networks are trained for 100 epochs on 4-second long segments. The initial learning rate is set to $1e^{-3}$. The learning rate is decayed by 0.88 for every two epochs. If the accuracy of validation set is not improved for 10 consecutive epoch, early stopping is applied. Adam is used as the optimizer. Gradient clipping with maximum L2-norm of 5 is applied for all experiments. All models are trained with utterance-level permutation invariant training (uPIT) [23] to maximize scale-invariant SNR (SI-SNR). SI-SNR is defined as:

$$s_{target} = \frac{\langle \hat{x}, \tilde{x} \rangle \, x}{\|x\|^2} \tag{7}$$

$$e_{noise} = \tilde{x} - s_{target} \tag{8}$$

$$SI - SNR := 10 log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \tag{9}$$

where $x$, $\tilde{x}$ are clean and estimated source respectively. The training target is always the reverberant clean speech signals. SI-SNR improvement (SI-SNRi) as the separation performance metric is reported. And the setting parameters of FaSNet-TAC are adopted in this paper.

## 5. RESULTS AND DISCUSSIONS

We compare the performance of the DMANet with different DMA blocks and FaSNet-TAC on multi-channel time-domain method. Table 1 shows the experiment results with the 6-mic circular array described in [14]. As we mentioned in Section 3.3, 1-D Conv lacks the ability in dealing with long-term dependence, and only DMA block and DMA block variant can not achieve the desired performance, although their model size is smaller than other models. Comparing the performance of DMANet-Only with DMANet-Variant's, we can see that the result of DMANet-Variant trained with SI-SNRi metric is a little better than DMANet-Only's. One explanation for this finding may be due to that DMANet-Variant can filter more directional nonsense signals than DMANet-Only. There are more different differential operations which can be viewed as more distinct nulls in DMA block variant.

The DMANet model proposed in this work significantly outperforms the performance of FaSNet-TAC in SI-SNRi metrics with a smaller model size, which is highlighted in Table 1. The result of DMANet-Pure also shows that DMA block works. As discussed in Section 2, some undesired signals are subtracted or filtered so that the desired signals are enhanced. Due to the limitation of experimental conditions, we believe that we haven't found the DMANet's best performance and it deserves more attention in our future research.

## 6. CONCLUSION

In this paper, we introduce an effective and efficient DMANet with DMA block variants to separate speech in multi-channel time domain. By combining DMA block with FaSNet-TAC, we establish an end-to-end multi-channel speech separation model. With taking advantage of DMAs' high directional gain, DMANet can achieve the best performance in SI-SNRi metric with smaller model size compared with other state-of-the-art strategies. Our numerical results tend to show that the performance by DMANet may be a really rough setting. This will lead our future research which would also imply some improved designs.

## 7. REFERENCES

[1] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans-*

*actions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[2] Yi Luo and Nima 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.

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

[4] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan 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.

[5] Michael Brandstein, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2001.

[6] Gary W Elko, "Differential microphone arrays," in *Audio signal processing for next-generation multimedia communication systems*, pp. 11–65. Springer, 2004.

[7] Gary W Elko and Anh-Tho Nguyen Pong, "A simple adaptive first-order differential microphone," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. IEEE, 1995, pp. 169–172.

[8] Shmulik Markovich, Sharon Gannot, and Israel Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[9] Jacob Benesty, Jingdong Chen, Chao Pan, et al., *Fundamentals of differential beamforming*, Springer, 2016.

[10] Xi Chen, Jacob Benesty, Gongping Huang, and Jingdong Chen, "On the robustness of the superdirective beamformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 838–849, 2021.

[11] Gongping Huang, Jingdong Chen, and Jacob Benesty, "Insights into frequency-invariant beamforming with concentric circular microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2305–2318, 2018.

[12] Gongping Huang, Jacob Benesty, and Jingdong Chen, "On the design of frequency-invariant beampatterns with uniform circular microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1140–1153, 2017.

[13] Liheng Zhao, Jacob Benesty, and Jingdong Chen, "Design of robust differential microphone arrays with the jacobi–anger expansion," *Applied Acoustics*, vol. 110, pp. 194–206, 2016.

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

[15] Yi Luo, Cong Han, Nima Mesgarani, Enea Ceolini, and Shih-Chii 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.

[16] Zhuohuang Zhang, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, and Dong Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6089–6093.

[17] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[18] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al., "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.

[19] Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, and Michiel Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech 2016*, 2016.

[20] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[21] Jacob Benesty, Jingdong Chen, and Israel Cohen, *Design of circular differential microphone arrays*, vol. 12, Springer, 2015.

[22] Jacob Benesty and Chen Jingdong, *Study and design of differential microphone arrays*, vol. 6, Springer Science & Business Media, 2012.

[23] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper 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.