Interspeech 2024
1-5 September 2024, Kos, Greece

DeFTAN-AA: Array Geometry Agnostic Multichannel Speech Enhancement

Dongheon Lee[1], *Jung-Woo Choi*[1]

[1]Korea Advanced Institute of Science and Technology, Republic of Korea

donghen0115@kaist.ac.kr, jwoo@kaist.ac.kr

Abstract

We propose an array geometry agnostic multichannel speech enhancement model, which is trained on a single microphone array but can enhance speech in various arrays with different shapes and numbers of microphones. To enable array-agnostic processing, the model employs a gated split dense block (GSDB) that separates foreground speech and background noise regardless of array geometry. Furthermore, to design an array-agnostic encoder compatible with different numbers of microphones, we introduce the spatial transformer (ST) that aggregates spatial information by channel-wise self-attention. The proposed space-object cross-attention (SOCA) block alleviates overfitting to a specific array configuration through cross-attention between spatial features and object features. Experimental results demonstrate the efficacy of the proposed model across various array geometries in both simulated and real-world datasets.

Index Terms: array-agnostic speech enhancement, gated split dense block, spatial transformer, space-object cross-attention

1. Introduction

Over the past decade, significant progress in deep learning approaches has brought remarkable advancements in speech enhancement and automatic speech recognition (ASR) technologies [1]. When speech signals include noises and reverberations, ASR performance can deteriorate significantly, necessitating preprocessing through speech enhancement [1, 2].

Various approaches to speech enhancement, including single-channel [3], multichannel [4, 5], and audio-visual methods [6], have been proposed. This study deals with multichannel speech enhancement, which offers the advantage of leveraging spatial information over single-channel enhancement [7, 8]. However, there are several challenges to be addressed for using multichannel speech enhancement models in real-world applications [9]. One such challenge is to consistently enhance multichannel noisy reverberant speech irrespective of microphone array geometries, such that the time and effort for re-training of a huge deep neural network (DNN) model can be avoided.

Recently, many array geometry agnostic speech enhancement and separation models have been proposed [9, 10, 11, 12]. To realize array geometry agnostic speech enhancement, most of the previous studies have utilized diverse training data simulated for various array geometries [9, 12, 13]. However, creating a dataset encompassing all possible array geometries is impractical. Therefore, it is necessary to develop models that consistently enhance noisy and reverberant speeches for various array configurations without fine-tuning or re-training.

In this study, we introduce a model capable of array geometry agnostic speech enhancement through training with only a single array, named array-agnostic dense frequency-time attentive network (DeFTAN-AA). We first demonstrate that the recent model (DeFTAN-II) [14] based on the spectro-temporal transformer can achieve high performance for various shaped arrays, through the utilization of a split dense block (SDB). Then, we proposed a gated split dense block (GSDB) for array geometry agnostic speech enhancement. The GSDB implements information gating using the gated linear unit (GLU) [15] to selectively aggregate the foreground speech-related features in various multichannel data.

To secure independence over the number of microphones, we propose sequential channel attention between spatial features and individual microphone data. To this end, we introduce a spatial transformer (ST) that extracts spatial features using channel-wise self-attention and combines them with the original multichannel data, enabling efficient spatial feature encoding regardless of the number of microphones. Additionally, we introduce a space-object cross-attention (SOCA) module that implements channel-wise cross-attention between spatial features encoded by ST and object features extracted from a single-channel encoder. This facilitates the information exchange between spectro-temporal and spatial features. Moreover, the cross-attention includes no training parameter relevant to the number of microphones. We validate these information processing modules by comparing speech enhancement performance on the test dataset collected from various unseen arrays. Furthermore, we demonstrate the superiority of the proposed model by comparing it with baseline models using simulated and real-world datasets.

2. Proposed method

2.1. Model architecture

The model architecture, depicted in Figure 1, utilizes a complex spectral mapping method to directly estimate the complex spectrogram of clean speech from multichannel noisy and reverberant speech. The model employs a spatial transformer (ST) for encoding and extracting spatial relations between each microphone pair, wherein the frequency and microphone channel dimensions are designated as the embedding and sequence dimensions, respectively. After encoding, up-convolution (UpConv) expands the channel dimension and feeds individual microphone data into weight-shared DeFTAN-II blocks.

The input spectrogram is also processed by a single-channel convolutional encoder, which extracts spectro-temporal features through a Conv2d layer. These spectro-temporal features are fused with spatial features processed by ST through the space-object cross-attention module (SOCA). The cross-attended features are then summarized by mean-pooling applied along the microphone channel dimension. Subsequently, the features pass

10.21437/Interspeech.2024-669

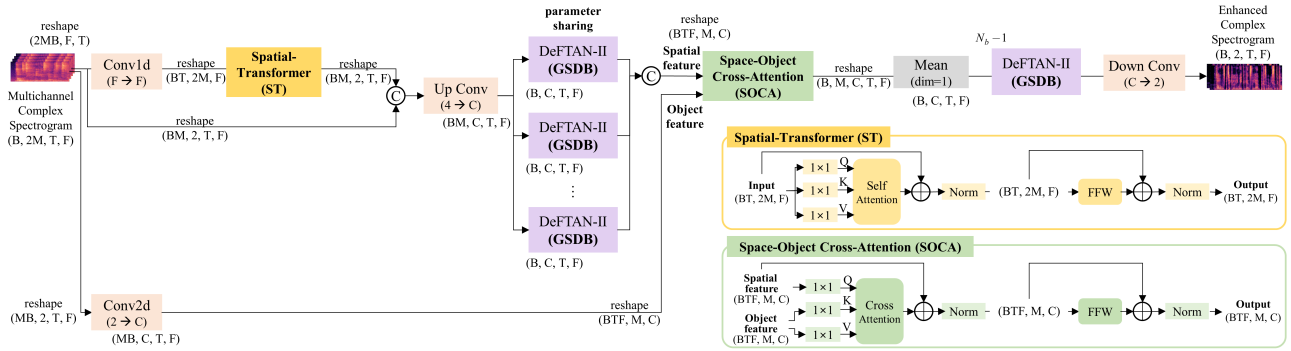

Figure 1: *Overall architecture of the proposed DeFTAN-AA.*

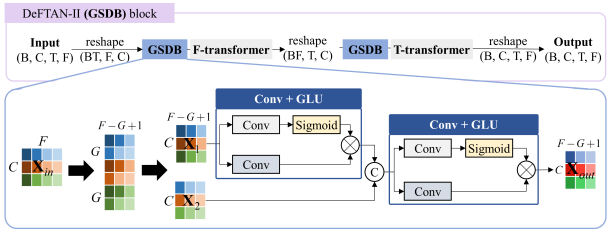

Figure 2: *Architecture of DeFTAN-II block and gated split dense blocks (GSDB).*

through a series of DeFTAN-II blocks, followed by Down-Conv, resulting in the complex spectrogram of the estimated speech. In this study, the frequency (F) and time (T) transformer in the DeFTAN-II blocks are identical to vanilla transformers [16] except for efficient attention [17] utilized to reduce their computational complexities.

2.2. Gated Split Dense Block (GSDB)

The existing model, DeFTAN-II [14], consists of a series of DeFTAN-II blocks, each of which includes the one-dimensional split dense block (1D SDB) followed by a F-transformer, another 1D SDB, and a T-transformer. The key factor that differentiates DeFTAN-II from its predecessor (DeFTAN) [5] is SDB. Unlike the dense block, SDB can construct embeddings by considering adjacent frequency or time information with channel information simultaneously through the unfold-transpose convolution structure [18] and channel shuffle [19]. In detail, a total of G sub-features are created by the unfold operation and stacked along the channel dimension with channel shuffle. Afterward, the features are passed through a convolution layer to extract information. This design reduces computational complexity and allows for capturing comprehensive relationships between channels, time, and frequency.

The use of SDB has brought significant performance leaps across various array geometries, which became the motivation to develop an array-agnostic model in this study. However, to reduce the geometry dependency further, we employ a gating mechanism to SDB in this work. The gated linear unit (GLU) applies gating to retain the foreground features possessing speech characteristics among the extracted spatial features, and operates independently of array geometry. The proposed gated split dense block (GSDB) shown in Figure 2 allows for controlled modulation of feature outputs from each convolutional layer. Although the increased output channel dimensions result in higher computational complexity, the Conv-GLU structure is expected to facilitate more accurate mappings.

2.3. Spatial Transformer (ST)

An array-agnostic model should be able to enhance speeches even when the number of microphones in the array changes. However, the conventional convolutional encoders [14, 18, 20, 21, 22] processing microphone signals utilize convolution kernels whose channel dimension depends on the number of microphones (M), making pre-trained models completely unusable when the number of microphones changes. To address this limitation, the proposed ST employs channel attention for encoding. In the $(2M, T, F)$-sized input feature, the time (T) dimension is treated as the batch dimension, frequency (F) as the embedding dimension, and microphone $(2M)$ as the sequence dimension. The self-attention generates an attention map of size $2M \times 2M$ to capture spatial information, followed by a feed-forward network. Subsequently, the real and imaginary terms are separated from the microphone dimension and the feature is rearranged into a $(M, 2, T, F)$-size. The spatial feature encoded through ST is concatenated with the reshaped complex spectrogram, followed by Up-Conv.

Generating attention maps for microphone channels enables us to exploit the relationships between each pair of microphones irrespective of sensor counts. This capability facilitates array geometry agnostic speech enhancement.

2.4. Space-Object Cross-Attention (SOCA)

Space-object cross-attention (SOCA) prevents the model from overfitting to the trained number of microphones. To resolve the overfitting issue, previous studies have utilized the parameter sharing and transform-average-concatenate (TAC) structure [9, 11], or performed single-channel speech enhancement followed by multichannel beamforming [10, 8]. However, to capture more complex spatial relations, we need more comprehensive processes other than the simple channel mixing of TAC or multichannel beamforming. To this end, we first refine the spatial features obtained from ST using parameter-shared DeFTAN-II blocks. Through the parameter-shared processing for each microphone, the model can expect balanced learning of spectro-temporal patterns for each microphone. Next, we introduce the cross-attention between spatial and object features. Here, the object features refer to features separately encoded by the Conv2d layer configured to extract only spectro-temporal patterns. The cross-attention is performed by taking the microphone dimension as the sequence dimension. It considers the spatial feature as the query, and the object feature as the key and value. The linear projection is applied to all queries, keys, and values before calculating the attention map. The SOCA promotes information exchange between spatial and spectro-temporal features without training parameters

depending on the number of microphones, thus lending itself to number-of-sensors agnostic processing. The attended output is then added to the spatial feature and processed by the feedforward network. Lastly, the features undergo mean-pooling along the microphone dimension, yielding the output of dimension (B, C, T, F). This number-of-sensors-independent data is then fed into a series of DeFTAN blocks for further refinement and then merged into the complex spectrogram output.

3. Experiments

3.1. Dataset

We designed training and validation datasets using a reference array, given by a four-channel circular microphone array with a 10 cm radius, while using various array geometries for the test dataset. Initially, we generated noisy-reverberant training and validation datasets using the DNS challenge 2020 corpus [23], which included both speech and noise components. To simulate spatialized speech and noise signals, rooms were generated by randomly populating width, depth, and height from the prescribed intervals of [5, 10] m, [5, 10] m, and [3, 4] m, respectively. The room impulse responses (RIRs) were computed using the image source method through the *pyroomacoustics* library [24]. This dataset encompassed signal-to-noise ratios (SNRs) ranging from [-10, 10] dB and reverberation times (T_{60}) spanning [0.2, 1.2] s. The high-level noise data in this dataset exhibited non-stationary characteristics, posing a challenge for noise removal. We spatialized both speech and noise using the procedure outlined in [25]. All models were trained only with the reference array.

For the test dataset, we utilized speech from Librispeech [26] and noise from the DNS challenge dataset. The dataset generation method was similar to that of the training set, except for the array geometry. We conducted experiments with six distinct array configurations: the reference (circular) array with a 10 cm radius, circular arrays of variable radius, rectangular arrays, linear arrays, tetrahedral arrays, and random arrays. The aperture size of arrays, except the reference array, randomly varied from 2 cm to 20 cm. Here, the aperture size denotes the length of a larger edge for the rectangular array designed with a 2:1 aspect ratio, the total length for the linear array, and the distance between each microphone pair for the tetrahedral array. For the random arrays, we placed the microphones at random points within a sphere whose radius varied within the range of aperture size. The test dataset includes 1,000 data for each microphone array, with the room, source, and array center configurations being identical except for the array geometry.

In addition, we conducted real-world experiments to explore the applicability of real-world scenarios. To capture noisy reverberant speech, we used a ReSpeaker array v2.0 4-channel circular array with a radius of 2.7 cm and a rectangular array of size 8.2 cm $\times$ 4.2 cm constructed by selecting four microphones on the miniDSP UMA-16 array. The test was done in an office room of size ($W = 4.1, D = 6, H = 2.5$) m and $T_{60} = 0.2$ s. For the test, the model trained by the simulation data described in Sec. 3.1 was directly utilized without fine-tuning. Speech and noise signals fed into two loudspeakers positioned at 1 m in front of the wall were scaled such that the SNR becomes 5 dB in a signal level, and the distances between speakers and microphone were within the range of 1 - 2 m. For the real recordings, we used 100 samples with durations [3, 12] seconds for each utterance. The speech and noise were obtained from both the unseen LibriSpeech and DNS challenge test datasets.

3.2. Experimental procedure

We extracted the complex spectra from the four microphone signals using a short-time Fourier transform (STFT) with a rectangular window of 32 ms length (512 samples) and a 50% overlap. We utilized the real and imaginary (RI) components of the complex spectra with frequency dimensions of $F = 257$. For our model configurations, we set the total number of the cascaded DeFTAN-II blocks as $N_b = 4$, the number of microphones $M = 4$, the channel dimension length of the Up-Conv block as $C = 64$, and both the kernel size of Up and Down Conv as $k = 3$. All models underwent training for 100 epochs, utilizing the phase-constrained magnitude (PCM) loss [27]. When the validation loss did not decrease after five epochs, we applied early stopping. We employed the Adam optimizer with a batch size of 4, starting with a learning rate of 0.001.

All models discussed in this study are non-causal systems. These models are evaluated with both the speech enhancement metric, scale-invariant signal-to-distortion ratio (SI-SDR), perceptual evaluation of speech quality (PESQ), and the speech recognition metric, word error rate (WER). We evaluated WER by feeding the enhanced speech from the trained model to wav2vec 2.0 [28] pre-trained on the Librispeech 960 h dataset sampled at 16 kHz. The experimental results of the reference array are expected to yield the highest performance. We can examine the influence of radius differences from the results of the circular array. Comparing the results of the other arrays with those of the reference and circular arrays allows us to examine how well the model performs array geometry agnostic speech enhancement. Furthermore, to investigate the effect of the number of microphones change, we constructed test sets using 3, 4, and 8 microphones for each array geometry. We also compared parameter sizes and computational complexities.

4. Results

The results of the ablation studies are presented in Table 1. The model trained solely with the F- and T-transformers, without the SDB, exhibits the worst performance. The model that only learned the frequency or time relationships proves ineffective for speech enhancement. The model trained with the SDB denotes the baseline (DeFTAN-II) model, which shows significantly improved performance compared to those without SDB, underscoring the importance of utilizing spatial information in multichannel speech enhancement. The SDB allows for more precise mapping and reduces performance variations across array geometries. The SDB and GSDB enable array geometry agnostic speech enhancement by grouping features into groups along the channel dimension and concatenating with a different group at each convolution layer. The iterative process of exchanging information over the channel dimension gradually reduces the dependency on specific array geometry. In the case of GSDB, the maximum difference in enhancement performance between the reference array and other arrays is 0.7 dB in terms of SI-SDR, 0.12 in terms of PESQ, and 4.3% in terms of WER. However, for SDB, these values are 1.1 dB, 0.19, and 9.4%, respectively. The results indicate that using GSDB leads to better array generalization compared to using SDB, through GLU extracting signal-dependent information rather than array-dependent information selectively.

To check whether this improved performance originated from enhanced temporal or spectral feature aggregation, we compared SDB-MC and GSDB-MC with their single-channel version, SDB-SC and GSDB-SC, respectively, which take

Table 1: *The results of ablation study for a fixed number of microphones ($M = 4$). The 'wo' and 'w' denote 'without' and 'with', respectively. SC and MC denote single- and multi-channel, respectively.*

Methods	Ch.	SI-SDR (dB) ↑						PESQ ↑						WER (%) ↓						Param.	MAC/s
		Ref.	Cir.	Rec.	Lin.	Tet.	Ran.	Ref.	Cir.	Rec.	Lin.	Tet.	Ran.	Ref.	Cir.	Rec.	Lin.	Tet.	Ran.		
Noisy	-	-7.8	-7.8	-7.9	-7.9	-7.8	-7.9	1.38	1.37	1.36	1.34	1.35	1.35	97.0	97.3	97.8	98.5	97.8	98.0	-	-
wo SDB	MC	2.8	2.6	1.0	1.2	1.8	1.7	1.80	1.76	1.56	1.60	1.67	1.65	73.1	75.3	85.5	85.1	79.3	79.4	0.4 M	6.6 G
w SDB	MC	9.8	9.7	8.7	8.8	9.1	9.2	2.99	2.97	2.82	2.80	2.88	2.90	11.2	12.3	18.5	20.6	15.2	14.5	0.8 M	12.4 G
w GSDB	MC	9.6	9.4	9.2	8.9	9.6	10.0	2.96	2.93	2.87	2.84	2.92	2.94	13.2	14.3	15.5	17.5	13.9	13.2	1.1 M	17.2 G
w SDB	SC	4.4	4.2	4.3	4.4	4.1	4.1	2.10	2.08	2.09	2.05	2.01	2.02	66.8	68.3	67.7	68.4	69.0	69.7	0.8 M	12.4 G
w GSDB	SC	4.3	4.1	4.1	4.5	4.2	4.3	2.05	2.04	2.00	2.07	2.03	2.04	69.1	69.3	70.3	66.7	68.1	68.1	1.1 M	17.2 G
w GSDB+ST	MC	**10.3**	**10.2**	**9.9**	**9.9**	**10.0**	**10.1**	**3.02**	**3.01**	**2.98**	**2.98**	**3.00**	**3.01**	**10.3**	**11.1**	**12.0**	**12.0**	**11.4**	**11.3**	2.0 M	19.3 G
w GSDB+SOCA	MC	8.6	8.5	7.9	8.0	8.1	8.2	2.88	2.85	2.78	2.80	2.80	2.81	19.5	20.1	23.8	24.5	21.7	21.3	1.3 M	33.6 G
Proposed	MC	9.4	9.3	9.0	9.0	9.2	9.3	2.95	2.94	2.85	2.83	2.88	2.89	14.7	15.0	16.3	16.1	15.3	15.2	2.2 M	35.7 G
Clean	-	-	-	-	-	-	-	-	-	-	-	-	-	4.3	4.2	4.3	4.2	4.2	4.2	-	-

Table 2: *Performance comparisons with state-of-the-art models for the changing number of microphones. (evaluation metric: SI-SDR)*

Methods	$M = 3$						$M = 4$						$M = 8$						Param.	MAC/s
	Ref.	Cir.	Rec.	Lin.	Tet.	Ran.	Ref.	Cir.	Rec.	Lin.	Tet.	Ran.	Ref.	Cir.	Rec.	Lin.	Tet.	Ran.		
Noisy	-7.8	-7.8	-7.9	-7.9	-7.8	-7.9	-7.8	-7.8	-7.9	-7.9	-7.8	-7.9	-7.8	-7.8	-7.9	-7.9	-7.8	-7.9	-	-
Mask-beamforming [10]	2.1	2.0	2.2	1.7	2.3	1.9	2.9	2.9	2.7	2.2	2.4	2.5	3.5	3.5	3.7	3.4	3.2	3.3	3.6 M	0.9 G
FasNet+TAC [11]	3.4	3.3	3.2	3.1	3.2	3.3	4.5	4.4	3.8	3.8	3.6	3.7	4.7	4.7	4.1	4.1	4.0	4.1	2.9 M	14.8 G
NBDF [12]	4.1	4.0	3.8	4.0	3.8	4.1	4.6	4.6	4.7	4.2	4.5	4.8	5.1	5.2	4.9	4.5	4.7	4.8	1.2 M	37.3 G
DRC-Net [4]	-	-	-	-	-	-	6.4	6.0	4.4	4.2	4.8	4.7	-	-	-	-	-	-	2.6 M	13.9 G
w GSDB	-	-	-	-	-	-	9.6	9.4	8.9	9.2	9.6	10.0	-	-	-	-	-	-	1.1 M	17.2 G
w GSDB+ST	7.4	7.2	7.3	7.3	7.0	7.4	**10.3**	**10.2**	**9.9**	**9.9**	**10.0**	**10.1**	8.3	8.2	8.1	8.0	7.9	8.0	2.0 M	19.3 G
w GSDB+SOCA	7.8	7.8	7.7	7.6	7.5	7.9	8.6	8.5	7.9	8.0	8.1	8.2	8.9	8.8	8.5	8.4	8.9	9.0	1.3 M	33.6 G
Proposed	**8.5**	**8.4**	**8.2**	**8.1**	**8.1**	**8.2**	9.4	9.3	9.0	9.0	9.2	9.3	**9.6**	**9.5**	**9.4**	**9.4**	**9.7**	**9.6**	2.2 M	35.7 G

only single-channel input. The huge enhancement differences between single-channel and multi-channel models stress that SDB-MC and GSDB-MC effectively harness the advantages of multichannel and spatial information processing to realize array-agnostic speech enhancement ability, rather than solely relying on the temporal and spectral features.

To ensure speech enhancement remains possible even with changes in the number of microphones, we conduct ablation studies with ST and SOCA. In Tables 1 and 2, the model only using GSDB cannot be utilized for different numbers of microphones because of its fixed convolution kernel size. We also demonstrate ST can improve enhancement performance and reduce performance variations across arrays. This is possible because ST exploits the relationship between every pair of microphones individually by the channel attention. Despite the highest performance of the GSDB+ST model, there is a drawback of being overfitted to the number of microphones used during training, as shown in Table 2. Accordingly, the GSDB+ST model may not generalize well when the number of microphones changes. SOCA complements the overfitting and achieves generalization over the number of microphones, yielding gradually increasing performance with increasing numbers of microphones. Although the increase in computational complexity results from parameter sharing, this result stresses that the sequential process involving parameter sharing, cross-attention, and mean-pooling helps generalization for the number of microphones. While using only ST yields optimal performance when the number of microphones is fixed and array geometry changes, employing both ST and SOCA becomes necessary when considering variations in the number of microphones.

In Table 2, we compared the performance of the proposed models with the baseline model: Mask-beamforming [10], FasNet+TAC [11], narrow-band deep filtering (NBDF) [12], and DRC-Net [4]. Figure 3 presents the results of experiments conducted in a real-world environment. Mask-beamforming performs single-channel speech enhancement for every channel, followed by a minimum-variance distortionless response (MVDR). FasNet+TAC and NBDF are representative array geometry agnostic models, proven to be effective when trained

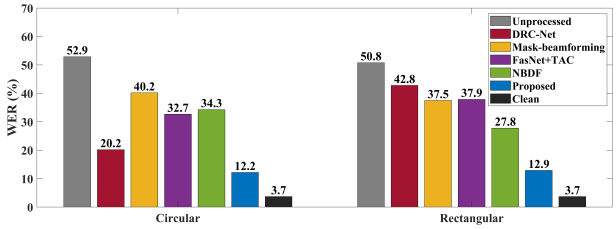

Figure 3: *Real-world experimental results. WERs on a circular array (left) and rectangular array (right).*

with various arrays. Although DRC-Net is not an array geometry agnostic model, it was used as a baseline for comparing speech enhancement performance. Experimental results demonstrate that the proposed model outperforms all others in terms of performance across different numbers of microphones and array geometries, with high performance in real-world experiments as well. The proposed model also achieves the generalized performance with different shapes and numbers of microphones. Especially noteworthy with the proposed model is its remarkably low WER difference of 0.7% between the circular and rectangular array geometry observed in real-world experiments, indicating its high potential for real-world applications.

5. Conclusion

This study proposes array geometry agnostic multichannel speech enhancement models through comprehensive spatial information processing using gated split dense block, spatial transformer, and space-object cross-attention. We compared the proposed models to the conventional models to validate the advantage of suggested modules in enhancing speeches acquired from various arrays unseen during training. Furthermore, we conducted a performance comparison along with the number of microphones. This comparison shows that the proposed method can analyze spatial relations between microphone channels of unseen arrays to realize array-agnostic speech enhancement and recognition. The results of real-world experiments have demonstrated the practical applicability of the proposed model.

6. References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] R. Haeb-Umbach *et al.*, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, 2019.

[3] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Virtual, 2021, pp. 6648–6652.

[4] J. Liu and X. Zhang, "DRC-NET: Densely connected recurrent convolutional neural network for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Singapore, 2022, pp. 166–170.

[5] D. Lee and J.-W. Choi, "DeFT-AN: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Process. Lett.*, vol. 30, pp. 155–159, 2023.

[6] D. Michelsanti *et al.*, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[7] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automat. Speech Recognit. Understanding*, Scottsdale, AZ, 2015, pp. 436–443.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Shanghai, China, 2016, pp. 196–200.

[9] T. Yoshioka *et al.*, "VarArray: Array-geometry-agnostic continuous speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Singapore, 2022, pp. 6027–6031.

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

[11] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Barcelona, Spain, 2020, pp. 6394–6398.

[12] S. Zhang and X. Li, "Microphone array generalization for multichannel narrowband deep speech enhancement," in *Proc. Interspeech*, Brno, Czech, 2021, pp. 666–670.

[13] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, Singapore, 2022, pp. 271–275.

[14] D. Lee and J.-W. Choi, "DeFTAN-II: Efficient multichannel speech enhancement with subgroup processing," *arXiv preprint arXiv:2308.15777*, 2023.

[15] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Lear.*, Sydney, Australia, 2017, pp. 933–941.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Info. Process. Sys.*, vol. 30, Long beach, CA, USA, 2017.

[17] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proc. IEEE/CVF Winter Conf. Appl. Comp. Vis.*, Waikoloa, HI, 2021, pp. 3531–3539.

[18] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023.

[19] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comp. Vis. Patt. Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6848–6856.

[20] C. Quan and X. Li, "Multichannel Speech Separation with Narrow-band Conformer," in *Proc. Interspeech 2022*, 2022, pp. 5378–5382.

[21] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Multichannel speech enhancement without beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Singapore, 2022, pp. 6502–6506.

[22] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.

[23] C. K. Reddy *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2492–2496.

[24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, 2018, pp. 351–355.

[25] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Singapore, 2022, pp. 6497–6501.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, South Brisbane, Australia, 2015, pp. 5206–5210.

[27] A. Pandey and D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1270–1279, 2021.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Info. Process. Sys.*, vol. 33, pp. 12 449–12 460, Virtual, 2020.