Multichannel Speech Enhancement Using Complex-Valued Graph Convolutional Networks and Triple-Path Attentive Recurrent Networks

Xingyu Shen and Wei-Ping Zhu
Department of Electrical and Computer Engineering, Concordia University, Canada
E-mail: shenxingyu97@outlook.com; weiping@ece.concordia.ca

***Abstract*—Multichannel speech enhancement has gained significant attention for its capability of improving speech quality and intelligibility in noisy environments. This paper presents a novel approach to multichannel speech enhancement utilizing complex-valued graph-in-graph convolutional networks (GiGCN) and triple-path attentive recurrent networks (TPARN). The proposed model leverages complex-valued operations to capture spatial dependencies and decoupled LSTM blocks to model temporal correlations. Meanwhile, the TPARN can effectively fuse the frequency, time, and spatial features for the reconstruction of the enhanced speech. Our experimental results based on the CHiME-3 and L3DAS22 datasets show that the proposed integrated model outperforms the state-of-the-art methods in terms of the PESQ, STOI and WER performance metrics.**

I. Introduction

Traditional approaches to multichannel speech enhancement have often relied on beamforming techniques, which utilize spatial filtering to enhance speech signals captured by multiple microphones. Beamforming methods, such as minimum variance distortionless response (MVDR) [1][2] and generalized sidelobe canceler (GSC) [3], have demonstrated their effectiveness in many scenarios. However, these methods are limited by their reliance on accurate spatial information and may not fully exploit the temporal dynamics of speech signals.

With the advent of deep learning, there has been a significant shift towards using neural networks for speech enhancement. Convolutional neural networks (CNNs) and long short-term Memory (LSTM) networks have been widely adopted to model spatial and temporal dependencies in multichannel signals. For instance, recent works have employed CNNs to capture local spatial correlations [4] and LSTMs to model temporal dependencies in sequential data [5][6]. However, these methods often treat spatial and temporal features independently, potentially missing complex dependencies across these domains.

Graph convolutional networks (GCNs) have emerged as a powerful tool for modeling spatial relationships in various applications, including multichannel speech enhancement. GCNs effectively capture spatial dependencies across multiple channels by representing the microphones and their spatial relationships as a graph [7]. This approach allows for a more comprehensive understanding of spatial information, which can lead to more accurate enhancement of speech signals.

The use of complex-valued neural networks has shown great promise in directly processing the complex spectra obtained from short-time Fourier transform (STFT). Complex-valued operations can handle the phase information present in the STFT domain, which is crucial for accurate reconstruction of enhanced speech signal [8]. Moreover, attention mechanisms have been introduced to dynamically focus on important features across the time and frequency domains. These mechanisms can significantly improve the performance of speech enhancement models by selectively weighting those features that contribute most to the enhancement task [9].

Several state-of-the-art methods have advanced multichannel speech enhancement by combining various neural network architectures. For example, the use of spatial autoencoders to capture spatial dependencies while preserving the temporal structure has shown significant improvements in speech quality [10]. Additionally, decoupled spatial and temporal processing frameworks have demonstrated the potential to reduce computational complexity while maintaining high performance [11]. Dense frequency-time attentive networks have further enhanced the capability of speech enhancement models by integrating detailed frequency and temporal features [12].

Despite these advancements, there remains a need for a unified approach that can effectively leverage spatial, temporal, and frequency information. In this paper, we propose a novel method for multichannel speech enhancement that integrates complex-valued graph in graph convolutional networks (GiGCN) with a triple-path attentive recurrent network (TPARN). Our approach leverages the strengths of complex-valued operations to capture spatial dependencies directly from the STFT domain, while exploiting decoupled LSTM blocks to model both short-term and long-term temporal correlations. Moreover, a triple-path architecture is employed to extract the frequency, time, and spatial features effectively, allowing for a comprehensive feature fusion.

The proposed model is extensively evaluated on two benchmark datasets, i.e., CHiME-3 [13] and L3DAS22 [14]. Our experimental results show that the proposed approach significantly improves the quality and intelligibility of the enhanced speech, as compared to the state-of-the-art methods in various noisy environments. These findings corroborate the effectiveness and robustness of our integrated model in handling complex acoustic scenarios for real-world applications.

II. Model Description

A. Method Overview

Our proposed multichannel speech enhancement model integrates the complex-valued graph in graph convolutional networks (GiGCN) with a triple-path attentive recurrent network (TPARN) to effectively capture the spatial, temporal, and frequency information. The architecture of the proposed integrated model is depicted in the upper half of Figure 1. The input noisy speech signals from multiple channels are first transformed into the frequency domain using the short-time Fourier transform (STFT). These transformed signals are then processed through a series of complex-valued GiGCN layers to capture spatial dependencies, resulting in spatial features denoted as $\mathbf{H}_{space}$. The enhanced features from the GiGCN block are subsequently passed through decoupled LSTM blocks, which separately model short-term and long-term temporal dependencies, producing the temporal features $\mathbf{H}_{time}$. Finally, the TPARN integrates these spatial and temporal features, along with frequency features $\mathbf{H}_{freq}$ derived from the input STFT, to form a cohesive representation. These integrated features are then fused and used to reconstruct the into enhanced speech signals.

B. Complex-Valued GiGCN Block

The complex-valued GiGCN block leverages the properties of complex-valued neural networks for processing the STFT coefficients, which are naturally represented in the complex domain. This allows for the direct manipulation of both amplitude and phase information, crucial for accurately reconstructing enhanced speech signals. Given a noisy speech signal captured by N microphones, the input signal $\mathbf{x}(t)$ can be represented as:

$$\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)] \tag{1}$$

where $x_i(t)$ is the signal from the i-th microphone. Applying the STFT to each channel, we obtain the complex-valued spectrogram:

$$\mathbf{X}(k,l) = [X_1(k,l), X_2(k,l), \ldots, X_N(k,l)] \tag{2}$$

where $X_i(k,l)$ represents the STFT of the i-th microphone signal at frequency bin k and time frame l. The complex-valued convolution operation is defined as:

$$(\mathbf{W} \star \mathbf{X})(k,l) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} \mathbf{W}(m,n)\mathbf{X}(k-m, l-n) \tag{3}$$

where $\mathbf{W}(m,n)$ is the complex-valued filter weight and $\mathbf{X}(k,l)$ is the input complex spectrogram.

To capture spatial dependencies among microphones in a multichannel setup, we model the system as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ represent microphones and edges $\mathcal{E}$ represent the connections between them [15]. These connections are encoded in an adjacency matrix $\mathbf{A}$, whose element a_{ij} indicates the connection strength or similarity between microphones i and j.

In this setup, the graph convolution operation updates the feature vector $\mathbf{H}_i$ of each node by aggregating features from its neighboring nodes $j \in \mathcal{N}(i)$. This update at layer l is expressed as:

$$\mathbf{H}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \mathrm{a}_{ij} \mathbf{W}_{\text{GCN}}^{(l)} \mathbf{H}_j^{(l)}\right) \tag{4}$$

Here, $\mathbf{H}_i^{(l+1)}$ represents the updated feature vector for node i, $\mathbf{W}_{\text{GCN}}^{(l)}$ is the layer-specific learnable weight matrix, and σ is a non-linear activation function like ReLU.

GiGCNs extend this framework by introducing a nested graph structure, where each primary node (microphone) is further detailed as a secondary graph [16]. This nested structure is particularly beneficial for complex-valued data, such as STFT coefficients, where each coefficient has both real and imaginary parts. The GiGCN operation, which captures more intricate spatial relationships, is given by:

$$\mathbf{H}_{i,j}^{(l+1)} = \sigma\left(\sum_{k \in \mathcal{N}(i)} \mathrm{a}_{ik}^{\text{nested}} \mathbf{W}_{\text{GiGCN}}^{(l)} \mathbf{H}_{k,j}^{(l)}\right) \tag{5}$$

where $\mathbf{H}_{i,j}^{(l)}$ represents the feature of the j-th component (either real or imaginary) of the i-th node at layer l. The nested adjacency matrix $\mathbf{A}^{\text{nested}}$ and weight matrix $\mathbf{W}_{\text{GiGCN}}^{(l)}$ facilitate the aggregation of features across the graph.

This dual-level graph structure allows the model to capture detailed spatial dependencies within and across nodes, enhancing its ability to process and improve multichannel audio signals. The nested structure ensures that both intra-node and inter-node interactions are considered, providing a more comprehensive understanding of the spatial relationships in the data.

The complex STFT features are first decomposed into their real and imaginary components. Each component undergoes feature extraction using multi-scale GiGCN layers, which employ filters of varying sizes (3x3, 5x5, and 7x7). This multi-scale approach allows the capture of spatial features at different resolutions.

For both the real and imaginary parts, the features derived from the different filter sizes are concatenated, resulting in enhanced real and imaginary feature sets. These sets are then fused to produce a unified complex-valued feature representation, integrating information from both components. The detailed operation of the complex-valued GiGCN block is illustrated in the lower half of Figure 1.

This fused complex feature set effectively captures the intricate spatial details present in the STFT domain and serves as the input to the Decoupled LSTM blocks. These blocks are designed to further refine the features by modeling short-term

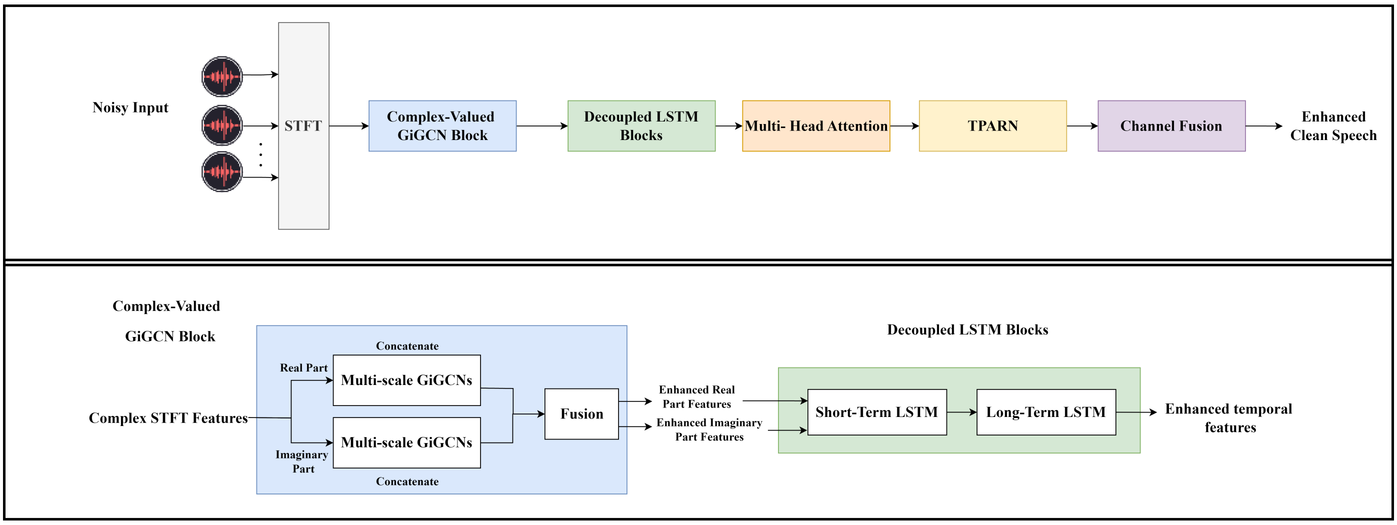

Fig. 1. Proposed multichannel speech enhancement framework

and long-term temporal dependencies, thereby enhancing the model's ability to improve speech quality in noisy environments.

C. Decoupled LSTM Blocks

The decoupled LSTM blocks are specifically designed to separately model the short-term and long-term temporal correlations in the enhanced features obtained from the complex-valued GiGCN block. By decoupling the temporal modeling into distinct short-term and long-term LSTM networks, we can more precisely capture temporal dynamics across different time scales, which is crucial for enhancing speech signals. The input features $\mathbf{H}_{enhanced}$ from the complex-valued GiGCN block are processed through two separate networks: short-term LSTM (ST-LSTM) and long-term LSTM (LT-LSTM). The ST-LSTM focuses on capturing rapid temporal variations, which are essential for preserving fine-grained temporal details in the speech signal. Conversely, the LT-LSTM is designed to capture slower, more sustained temporal dependencies, which are important for maintaining the overall structure and coherence of the speech signal over longer periods.

The outputs from the ST-LSTM and LT-LSTM networks are concatenated to form the final temporal features $\mathbf{H}_{temporal}$. This comprehensive temporal representation ensures that the enhanced speech signal retains both detailed and structural temporal characteristics.

D. Triple-Path Attentive Recurrent Network (TPARN)

The TPARN [17] further combines the frequency, time, and spatial features through a unified triple-path architecture. The temporal features $\mathbf{H}_{temporal}$, obtained from the decoupled LSTM blocks, are combined with the spatial features $\mathbf{H}_{space}$ extracted from the GiGCN block and frequency features $\mathbf{H}_{freq}$ from the STFT domain. These three types of features are processed in parallel paths within the TPARN: the frequency path captures frequency-domain correlations, the time path handles temporal correlations, and the spatial path processes spatial correlations across different channels. These paths employ convolutional, recurrent, and graph-based layers, respectively. The outputs from these paths are integrated using a multi-head attention mechanism (MHA), which combines the distinct features into a unified representation. This combined output, denoted as $\mathbf{H}_{TPARN}$, is formulated as:

$$\mathbf{H}_{TPARN} = \text{MHA}(\mathbf{H}_{freq}, \mathbf{H}_{time}, \mathbf{H}_{space}) \tag{6}$$

E. Channel Fusion

The channel fusion module combines the enhanced features from multiple channels into a single, unified representation. This process is essential for consolidating the spatial information captured from different locations. The input to the channel fusion module, $\mathbf{H}_{TPARN}$, is processed through a fully connected layer:

$$\mathbf{H}_{fused} = \text{ReLU}(\mathbf{W}_{fusion}\mathbf{H}_{TPARN} + \mathbf{b}_{fusion}) \tag{7}$$

where $\mathbf{W}_{fusion}$ and $\mathbf{b}_{fusion}$ are the weights and biases of the fully connected layer, respectively. The final output is transformed back to the time domain using ISTFT to obtain the enhanced clean speech.

F. Loss Function

The model is trained using a combination of time-domain and frequency-domain loss functions. The time-domain loss is calculated as the mean squared error (MSE) between the enhanced time-domain signal $\hat{\mathbf{s}}(t)$ and the ground truth clean signal $\mathbf{s}(t)$:

$$\mathcal{L}_{time} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{s}}(t) - \mathbf{s}(t))^2 \tag{8}$$

The frequency-domain loss is calculated as the MSE between the STFTs of the enhanced and ground truth signals:

$$\mathcal{L}_{freq} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} |\hat{\mathbf{S}}(k,l) - \mathbf{S}(k,l)|^2 \tag{9}$$

The combined loss function is a weighted sum of the time-domain and frequency-domain losses:

$$\mathcal{L} = \alpha \mathcal{L}_{time} + \beta \mathcal{L}_{freq} \tag{10}$$

where α and β are the weighting coefficients that balance the contributions of the two losses. In our experiments, we set $\alpha = 0.5$ and $\beta = 0.5$ to equally weigh the contributions of both losses.

The above-proposed comprehensive approach leverages the strengths of each component to enhance speech signals effectively and its detailed architecture is shown in the lower portion of Figure 1.

III. Experiments

A. Datasets

To evaluate the performance of our proposed multichannel speech enhancement model, called integrated model below for comparison, we utilized two widely recognized datasets: L3DAS22 and CHiME-3. The L3DAS22 dataset, designed for 3D speech enhancement tasks, includes over 40,000 virtual 3D audio environments recorded using a 1st order Ambisonics microphone, providing four-channel recordings [14]. Clean utterances are sourced from the Librispeech [18] corpus, and noise signals are from the FSD50K dataset [19]. The CHiME-3 dataset comprises noisy recordings from six channels in real-world environments, including buses, cafes, pedestrian areas, and street junctions [13], making it suitable for robust automatic speech recognition and multichannel speech enhancement tasks.

B. Experimental Setup

The noisy speech signals are transformed into the frequency domain using the STFT, with a window size of 1024 samples, a hop size of 512 samples, and an FFT size of 1024 points. The resulting complex spectrograms are processed by our integrated model, comprising complex-valued GiGCN, decoupled LSTM blocks, TPARN, and channel fusion modules. The model is trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 16 for 100 epochs. Data augmentation involves adding various types of noise at different SNRs, ranging from 0 dB to 20 dB, ensuring robustness across a wide range of noisy conditions. The model is evaluated on separate test sets of unseen noisy speech recordings to ensure a fair performance assessment.

C. Experimental Results and Analysis

The evaluation metrics used in this study include perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI) [26], and Word Error Rate (WER). We have also used an overall metric to evaluate our model based on dataset L3DAS22, which is a combination of STOI and WER [14]:

$$\text{Metric} = \frac{\text{STOI} + (1 - \text{WER})}{2} \tag{11}$$

The WER is computed based on the transcription of the estimated target signal and that of the reference signal, both decoded by a pre-trained Wav2Vec2.0-based ASR model [27].

The experimental results on the CHIME-3 dataset demonstrate that our integrated model yields a superior performance compared to other models, with higher average PESQ and STOI scores across all scenarios (BUS, CAF, PED, STR), as shown in Table I. Our model achieves an average PESQ of 1.905 and STOI of 0.922, outperforming other state-of-the-art methods. The ablation studies show that each component of our model contributes to the overall performance, with the integrated model achieving the highest scores.

Based on the L3DAS22 dataset (Table II), our proposed model significantly outperforms other state-of-the-art models. Our integrated model achieves the highest STOI score of 0.941 and the lowest WER of 0.065, resulting in a superior overall metric of 0.938. The ablation studies show that each component of our model contributes to the overall performance. Specifically, using a single 3x3 or 5x5 GiGCN block slightly decreases performance, indicating the importance of the multi-scale approach. The removal of either short-term or long-term LSTM also leads to lower performance, highlighting the necessity of capturing both short-term and long-term dependencies. The simplified channel fusion method results in lower STOI and higher WER, demonstrating the effectiveness of our sophisticated fusion technique.

Overall, our proposed model demonstrates significant improvements in PESQ, STOI, and WER metrics over state-of-the-art methods on both datasets. The comprehensive ablation studies further validate the effectiveness of our design choices, highlighting the importance of each component in achieving superior speech enhancement performance.

IV. Conclusions

In this work, we have presented a novel multichannel speech enhancement model integrating complex-valued GiGCNs, decoupled LSTM blocks, TPARN, and channel fusion. Our simulations based on L3DAS22 and CHiME-3 datasets demonstrated significant improvements over state-of-the-art methods in terms of PESQ and STOI metrics. Our ablation studies validated the contributions of each component, underscoring the importance of multi-scale approaches, temporal modeling, and feature integration. It is also shown that the proposed approach has a robust performance across various noisy environments, indicating a strong potential for practical applications in real-world scenarios.

References

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

TABLE I

ABLATION EXPERIMENTS AND PERFORMANCE COMPARISON WITH DIFFERENT MODELS ON THE CHIME-3 DATASET.

Model	**PESQ**					**STOI**				
	BUS	CAF	PED	STR	Avg.	BUS	CAF	PED	STR	Avg.
Our Integrated Model	1.942	**1.863**	**1.899**	1.925	**1.905**	**0.938**	**0.910**	**0.927**	**0.935**	**0.922**
Single 3x3 GiGCN Block	1.781	1.757	1.787	1.803	1.782	0.919	0.902	0.911	0.931	0.916
Single 5x5 GiGCN Block	1.797	1.768	1.792	1.808	1.792	0.921	0.905	0.913	0.933	0.918
Only Short-Term LSTM	1.784	1.760	1.788	1.805	1.784	0.923	0.907	0.914	0.934	0.920
Only Long-Term LSTM	1.777	1.754	1.782	1.798	1.778	0.917	0.900	0.909	0.929	0.913
Only Frequency and Time Paths in TPARN	1.773	1.748	1.775	1.793	1.772	0.916	0.899	0.908	0.928	0.912
Simplified Channel Fusion (Averaging)	1.786	1.763	1.791	1.807	1.787	0.918	0.903	0.912	0.932	0.916
Noisy (channel 1)	1.272	1.153	1.176	1.263	1.216	0.903	0.808	0.807	0.877	0.849
CNN [20]	1.775	1.540	1.618	1.751	1.671	0.920	0.885	0.882	0.913	0.900
LSTM-IPD [21]	1.908	1.496	1.512	1.730	1.661	0.921	0.847	0.835	0.890	0.873
UNet-GCN [7]	1.655	1.448	1.495	1.645	1.560	0.918	0.875	0.871	0.909	0.893
STGCSEN [15]	**2.002**	1.743	1.747	**1.941**	1.858	0.923	0.891	0.876	0.914	0.901

TABLE II

ABLATION EXPERIMENTS AND PERFORMANCE COMPARISON WITH DIFFERENT MODELS ON THE L3DAS22 DATASET.

Model	**WER**	**STOI**	**Metric**
Our Integrated Model	**0.065**	**0.941**	**0.938**
Single 3x3 GiGCN Block	0.069	0.936	0.934
Single 5x5 GiGCN Block	0.070	0.935	0.932
Only Short-Term LSTM	0.068	0.937	0.935
Only long-Term LSTM	0.071	0.934	0.931
Only Frequency and Time Paths in TPARN	0.072	0.932	0.930
Simplified Channel Fusion (Averaging)	0.070	0.933	0.932
Spatial-DCCRN [22]	0.071	0.931	0.931
SEU Speech [23]	0.101	0.902	0.900
JLESS [24]	0.174	0.836	0.831
CCA SPEECH [25]	0.240	0.831	0.796

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[3] M. R. Bai and F.-J. Kung, "Speech enhancement by denoising and dereverberation using a generalized sidelobe canceller-based multichannel wiener filter," *Journal of the Audio Engineering Society*, vol. 70, no. 3, pp. 140–155, 2022.

[4] S. Chakrabarty and E. A. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[5] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2019, pp. 298–302.

[6] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 271–275.

[7] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 3415–3419.

[8] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[9] H. Chen, X. Peng, Q. Jiang, and Y. Guo, "Residual unet with attention mechanism for time-frequency domain speech enhancement," in *2022 41st Chinese Control Conference (CCC)*, IEEE, 2022, pp. 7007–7011.

[10] M. M. Halimeh and W. Kellermann, "Complex-valued spatial autoencoders for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 261–265.

[11] A. Pandey and B. Xu, "Decoupled spatial and temporal processing for resource efficient multichannel speech enhancement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 206–12 210.

[12] D. Lee and J.-W. Choi, "Deft-an: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, 2023.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2015, pp. 504–511.

[14] E. Guizzo, C. Marinoni, M. Pennese, *et al.*, "L3das22 challenge: Learning 3d audio sources in a real office environment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9186–9190.

[15] M. Hao, J. Yu, and L. Zhang, "Spatial-temporal graph convolution network for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Con-*

ference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6512–6516.

[16] S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 1157–1171, 2022.

[17] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6497–6501.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[19] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

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

[21] W. Rao, Y. Fu, Y. Hu, *et al.*, "Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 679–686.

[22] L. Shubo, Y. Fu, J. Yukai, *et al.*, "Spatial-dccrn: Dccrn equipped with frame-level angle feature and hybrid filtering for multi-channel speech enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 436–443.

[23] J. Cheng, C. Pang, R. Liang, J. Fan, and L. Zhao, "Dual-path dilated convolutional recurrent network with group attention for multi-channel speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[24] J. Bai, S. Huang, H. Yin, Y. Jia, M. Wang, and J. Chen, "3d audio signal processing systems for speech enhancement and sound localization and detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[25] H. Wang, Y. Fu, J. Li, M. Ge, L. Wang, and X. Qian, "Stream attention based u-net for l3das23 challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2010, pp. 4214–4217.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.