SPATIAL-TEMPORAL GRAPH CONVOLUTION NETWORK FOR MULTICHANNEL SPEECH ENHANCEMENT

Minghui Hao, *Jingjing Yu**, *Luyao Zhang*

Electronic and Information Engineering, Beijing Jiaotong University

ABSTRACT

Spatial dependency related to distributed microphone positions is essential for multichannel speech enhancement task. It is still challenging due to lack of accurate array positions and complex spatial-temporal relations of multichannel noisy signals. This paper proposes a spatial-temporal graph convolutional network composed of cascaded spatial-temporal (ST) modules with channel fusion. Without any prior information of array and acoustic scene, a graph convolution block is designed with learnable adjacency matrix to capture the spatial dependency of pairwise channels. Then, it is embedded with time-frequency convolution block as the ST module to fuse the multi-dimensional correlation features for target speech estimation. Furthermore, a novel weighted loss function based on speech intelligibility index (SII) is proposed to assign more attention for the important bands of human understanding during network training. Our framework is demonstrated to achieve over 11% performance improvement on PESQ and intelligibility against prior state-of-the-art approaches in multi-scene speech enhancement experiments.

Index Terms— Graph convolution network, spatial dependency extraction, spatial-temporal convolution module, SII-weighted loss function, speech enhancement

1. INTRODUCTION

Multichannel speech enhancement uses the spatial diversity of array to extract clean speech from noisy mixtures received by microphones. It can be applied in many applications, such as audio surveillance system, immersive multimedia system, human-machine interface, automatic speech recognition, and remote conference [1-4]. However, the task is still challenging in real cases due to (1) lack of accurate array position makes the spatial dependency fail to be fully exploited; (2) complex spatial-temporal relations of multichannel signals cannot be effectively modeled and utilized to decorrelate speech.

The traditional beamforming algorithms, such as minimum variance distortionless response (MVDR) beamformer [5-6], can get optimal maximum directivity in the target direction, but usually suffer from serious performance degradation in realistic challenging scenarios due to the unavoidable measurement errors of mic positions, target localization errors, and mismatching propagation models. In recent years, deep neural networks (DNN) have achieved significant success in speech enhancement tasks [7]. For example, the long short-term memory network (LSTM) [8, 9] is naturally adapted to the speech sequences with dynamic length and achieves better noise suppression ability in severe scenes. Paper [10, 11] use convolutional neural network (CNN) to process signal spectrograms according to 2D image convolution and usually need more layers to extract high-dimensional features. However, all these methods only concentrate on the temporal information of acoustic signal, regardless of spatial relations between microphones, which is crucial for the decorrelation of noise and target speech. To further utilize the spatial dependencies of array, the graph convolutional network (GCN) is recently applied to model the inter-channel dependencies by aggregating information from neighbor channels [12]. Paper [13] embeds GCN with UNet[14, 15] structure to extract both spatial and temporal information to achieve better speech enhancement performance. However, it directly uses the original GCN network structure without consideration about the nature of acoustic signals and the intercorrelation of spatial-temporal features.

To overcome these problems and fully explore the dynamic spatial dependency of multichannel signals with time domain features, we propose a spatial-temporal graph convolutional network (STGCSEN) aiming at improving the quality of speech against noisy background without any prior information of array and acoustic scene. Spatial-temporal (ST) module including spatial GCN block and time-frequency domain CNN block is designed to extract the spatial, temporal and spectral correlations of multi-source mixed signals, such as background noise, burst interferences and speakers' phonemes. Through channel fusion, multi-dimensional correlation features are combined adaptively to reconstruct speech components. During the network training process, a novel loss function weighted by normalized speech intelligibility index (SII) is proposed with the purpose to

* Corresponding to jjyu@bjtu.edu.cn. The support of the Fundamental Research Funds for Central Universities (Grant No. 2021JBM004) is gratefully acknowledged.

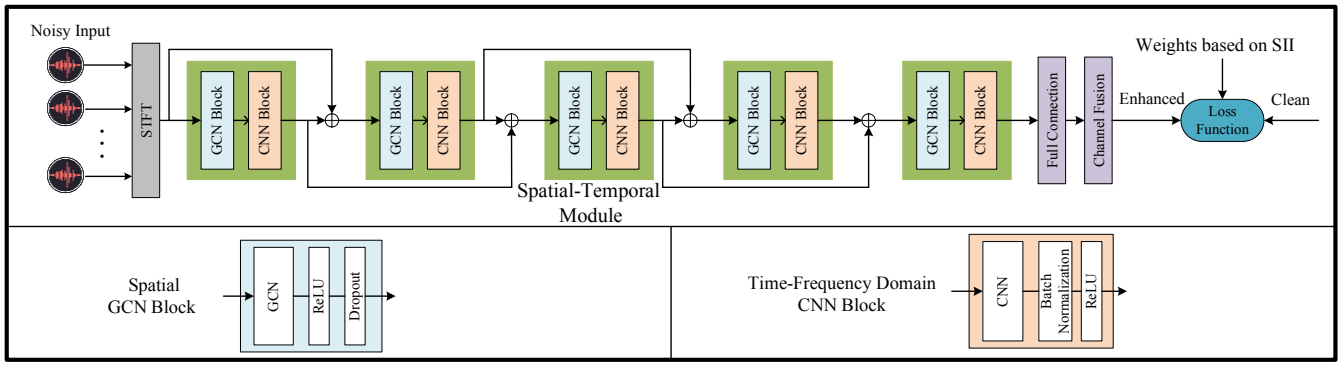

Fig. 1 Framework of proposed spatial-temporal graph convolution network. ⊕ represents the concatenating operation.

assign more attention to speech frequency bands that are important for human understanding. Experiments in different realistic noisy scenes are performed to demonstrate the superior speech enhancement performance of proposed framework, when comparing with commonly used LSTM, CNN and UNet-GCN methods.

The structure of this paper is as follows: Section 2 presents proposed framework for extracting spatial-temporal relations and enhance speech. Section 3 discusses the experimental results of recovered speech and important SII bands. Section 4 concludes this paper.

2. PROPOSED METHOD

2.1. Framework Overview

The goal of multichannel speech enhancement is to estimate the target speech from noisy mixture recordings received by multiple microphone channels. Let $\boldsymbol{X} \in \mathbb{R}^{\mathrm{C}\times\mathrm{F}\times\mathrm{T}}$ be the amplitude spectrum of received noisy speech after short-time Fourier transform (STFT), where F is the number of frequency bins, T is the number of time frames, and C is the number of microphone channels in array. The problem of speech enhancement can be formalized as learning a mapping function $h(\cdot)$: $\boldsymbol{X} \in \mathbb{R}^{\mathrm{C}\times\mathrm{F}\times\mathrm{T}} \xrightarrow{h(\cdot)} \widehat{\boldsymbol{Y}} \in \mathbb{R}^{\mathrm{F}\times\mathrm{T}}$, where $\widehat{\boldsymbol{Y}}$ is the estimated spectrum of clean speech $\boldsymbol{Y}$.

Other than the time-frequency features of acoustic signals, the spatial dependency of multichannel noisy signals received by distributed microphones is the crucial factor to decorrelate sounds transmitted from different source positions. To effectively model and utilize these dependencies for speech enhancement, a spatial-temporal graph convolutional framework is proposed, aiming at improving the quality of speech against noisy background without any prior information of array position and acoustic scene.

As shown in Fig. 1, proposed framework mainly consists of multiple cascaded ST modules and a channel fusion module, where the ST modules extract the spatial and time-frequency information of $\boldsymbol{X}$. The GCN block in ST module is used to extract the spatial dependency among multichannel signals related to different microphone positions. Then the CNN Block is used to capture the time-frequency feature of signal in each microphone channel to further decorrelate speech from noise. The channel fusion module generates an estimated target speech $\widehat{\boldsymbol{Y}}$ according to the multi-dimensional correlation feature of $\boldsymbol{X}$ derived from previous blocks. Finally, the network is trained based on a novel-defined weighted loss function that assigns more attention for the important frequency bands of human understanding.

2.2. Spatial Dependency Extraction

The spatial dependency of the target speech components related to distributed microphone positions can be estimated according to the amplitude and phase differences of received multichannel signals. When considering each mic as a node distributed in non-Euclidean space and the microphone array as a graph, the problem of extracting spatial dependency of multichannel noisy speech can be realized by GCN to aggregate neighbor node information through graph theory [16].

In our work, an undirected graph is defined as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ represents the set of microphones in array and $\boldsymbol{E}$ is the set of graph edges representing correlation relations between signals received by pairwise microphones. These relations can be characterized by the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$, where $\boldsymbol{A}_{i,j}$ represents the correlation weight from the jth to ith microphone. The degree matrix $\boldsymbol{D} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$ is a diagonal matrix $\boldsymbol{D}_{i,i} = \sum_j \boldsymbol{A}_{i,j}$. Different with the binary adjacency matrix commonly used in graphic neural network, our adjacency matrix consists of dynamic values obtained by real-time training to represent the actual phase shift and attenuation relations between frames of pairwise-mic signals. This definition of adaptive adjacency matrix can also prevent the performance degradation of traditional speech enhancement algorithms, brought by dynamic change and measurement errors of array position.

By renormalizing the adjacency matrix and the degree matrix, graph convolution of $\boldsymbol{X}_{:,:,t}$ can be computed by:

$$g_{\boldsymbol{\theta}} *_{\boldsymbol{G}} \{\boldsymbol{X}_{:,:,t}\} = \widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}_{:,:,t}\boldsymbol{\theta} \quad (1)$$

where $t \in [1, \dots, \mathrm{T}]$, $\boldsymbol{\Theta} \in \mathbb{R}^{\mathrm{C}\times\mathrm{M}}$ is the parameter matrix of M graph convolution kernels, $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}_{\mathrm{C}}$ and $\widetilde{\boldsymbol{D}}_{i,i} = \sum_j \widetilde{\boldsymbol{A}}_{i,j}$ are the renormalization operations, $\boldsymbol{I}_{\mathrm{C}}$ is the identity matrix with order C. In order to assign more flexibility for the nonlinear aggregating function of neighbor node information through GCN, a ReLU activation layer and a Dropout layer are added after GCN layer to form the spatial GCN block. As shown in Fig. 1, the output of GCN block with input $\boldsymbol{X}$ can be expressed as:

$$\boldsymbol{\Gamma}\{\boldsymbol{X}\} = Concat_{\mathrm{T}}\left(Dropout\left(ReLU\left(g_{\boldsymbol{\Theta}} *_{\boldsymbol{G}} \{\boldsymbol{X}_{:,:,t}\}\right)\right)\right) \quad (2)$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{\mathrm{C}\times\mathrm{M}\times\mathrm{T}}$ and $Concat_{\mathrm{T}}$ is the concatenating operation along frames for $t = 1, \dots, \mathrm{T}$.

2.3. Spatial-Temporal Convolution Module

Other than the spatial dependency of received multichannel signals related to distributed microphone positions, there are multiple correlations existing in the time-frequency domain of signal in each channel, derived from the stationary background noise, burst interferences and phoneme characteristics of speakers in acoustic scenes. In addition, the spatial dependency and time-frequency correlation features of multichannel noisy signals interact with each other. Therefore, we have designed cascaded ST modules to extract those time-varying spatial-temporal characteristics of received multichannel noisy signals for blind enhancement of speech components.

Considering stronger and more stable spatial dependency of array signals, proposed ST module consists a front-end GCN block (discussed in section 2.2) and a CNN block to extract time-frequency correlations of signals in each microphone channel. The mathematical expression of ST module is given as:

$$\boldsymbol{Z} = ReLU(BN(\boldsymbol{\Phi} * \boldsymbol{\Gamma}\{\boldsymbol{X}\})) \quad (3)$$

where $*$ indicates the standard convolution operation of CNN block, $\boldsymbol{\Phi}$ is the 2D convolution kernels in time-frequency domain. As shown in Fig. 1, multiple ST modules are cascaded with skip connections. The output of the lth ST module $\boldsymbol{Z}^{(l)}$ is given as:

$$\boldsymbol{Z}^{(l)} = ReLU(BN(\boldsymbol{\Phi} * \boldsymbol{\Gamma}\{\boldsymbol{Z}^{(l-1)} \oplus \boldsymbol{Z}^{(l-2)}\})) \quad (4)$$

where $\boldsymbol{Z}^{(0)} = \boldsymbol{X}$ and $l \in \{2, \dots, \mathrm{L}\}$. L represents the number of cascaded ST modules in network. Through experimental study, the optimal value of L is equal to C-1, which is consistent with the number of aggregating neighbor nodes and can be well explained by the graph convolution theory.

2.4 Channel Fusion

After the cascaded ST modules, features of speech components derived from spatial dependency and time-frequency correlation are captured. Then, a fully connected block and a channel fusion block are applied to fuse these multi-dimensional dependency features for target speech estimation as:

$$\widehat{\boldsymbol{Y}} = \sum_{c=1}^{\mathrm{C}} \boldsymbol{W}^c \odot (\mathcal{F}\{\boldsymbol{Z}^{(\mathrm{L})}_{c,:,:}\}) \quad (5)$$

where $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{\mathrm{F}\times\mathrm{T}}$, $\odot$ is the Hadamard product, and $\mathcal{F}\{\cdot\}$ represents the fully connected block. $\{\boldsymbol{W}^1, \boldsymbol{W}^c, \dots . \boldsymbol{W}^{\mathrm{C}}\}$ are the characteristic weight matrix of recovered speech spectrum in each microphone channel.

2.5. Weighted Loss Function based on Speech Intelligibility Index

With the purpose to effectively improve the speech intelligibility of recovered target speech, during network training, a novel loss function weighted by normalized SII [17] are proposed to assign more attention to the important frequency bands of human understanding. The loss function is defined as:

$$Loss_{SII} = \frac{1}{\mathrm{F}\times\mathrm{T}} \sum_{f=1}^{\mathrm{F}} \sum_{t=1}^{\mathrm{T}} w_f^{norm} \left(\boldsymbol{Y}_{f,t} - \widehat{\boldsymbol{Y}}_{f,t}\right)^2 \quad (6)$$

where $w_f^{norm} = \frac{\mathrm{F}-1}{\sum_{f=1}^{F} w_f} w_f$ is the normalized intelligibility weights for the fth frequency bin of speech, $w_f = \xi\left(f * (\frac{\mathrm{F_s}}{2} * (\mathrm{F}-1))\right)$, $\xi(\cdot)$ is the intelligibility polynomial function fitted from the SII indices and $\mathrm{F_s}$ is the sample rate of microphones.

3. EXPERIMENTS

3.1. Experimental Setup

The data used in experiments is from the public dataset CHIME-3 [18], which includes 6-channel microphone recordings from talkers speaking in multiple noisy scenes, such as bus, cafeteria and street. It consists of 7138, 1640, 1320 utterances of training, developing, and testing sets. In our experiments, we cut the data in 3s slots, then use 1455 slots for training and 364 slots for testing (91 slots per scenario).

Proposed STGCSEN network is compared with three baselines: (1) CNN [11, 19], 2D convolutional neural network with redundant convolutional encoder-decoder structure; (2) LSTM-IPD [20], a network composed of 3-layer real-valued LSTMs with the complex STFT spectrum to estimate inter-channel phase differences; (3) UNet-GCN[13], a model incorporating GCN in the embedding space of UNet structure. The speech enhancement performance of each model is assessed by PESQ and STOI [21-22]. For our model, 5 cascaded ST modules are applied. The size of convolution kernels is 5×2 and the stride is 2×1, except for the 3rd module is set to 3×3 and 1×1. The learning rate is set to 0.006, and the batch size is 12.

Table 1 Results (PESQ and STOI) on different models

	PESQ					STOI				
	BUS	CAF	PED	STR	Avg.	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	1.275	1.150	1.178	1.260	1.216	0.905	0.810	0.805	0.879	0.850
CNN	1.778	1.538	1.621	1.753	1.673	0.923	0.888	**0.885**	0.914	**0.903**
LSTM-IPD	1.912	1.491	1.509	1.734	1.662	**0.925**	0.850	0.833	0.888	0.874
UNet-GCN	1.659	1.444	1.492	1.643	1.560	0.922	0.877	0.869	0.911	0.895
STGCSEN (ours)	**2.004**	**1.745**	**1.743**	**1.944**	**1.859**	0.924	**0.893**	0.879	**0.917**	**0.903**

Table 2 Results of subjective listening test

	BUS	CAF	PED	STR
CNN	79±6	80±5	82±5	82±6
LSTM-IPD	84±8	80±8	74±7	84±6
UNet-GCN	78±5	77±6	79±7	82±6
STGCSEN (ours)	**84±4**	**85±4**	**83±6**	**87±6**

3.2. Experimental Results and Discussion

3.2.1. Result comparisons of different models

The speech enhancement results in four real scenarios (BUS: bus, CAF: cafe, PED: pedestrian, STR: street junction) are given in Table 1. By comparing the PESQ of four models, proposed STGCSEN outperforms the baselines in all noisy scenarios, which shows 11%-19% average improvement over all scenes. In addition, our framework provides comparable STOI values with the optimums in four scenarios. From the auditory perception test, the results of our framework in the scenes with moving noise sources perform better than the baselines by showing more similarity with clean speech and sounding more natural in the silent segments. Therefore, by effectively extracting the spatial and temporal information of multichannel signals, proposed STGCSEN achieves comparable or optimal performance for speech enhancement and shows stronger adaptability to different noisy scenarios.

Results from a subjective listening experiment based on Multi-Stimulus Test with Hidden Reference and Anchor (MUSHRA) are given in Table 2. 8 untrained listeners gave their subjective evaluation scores (between 0 and 100) for the enhanced speeches in different scenarios. The enhanced speeches obtained by our STGCSEN show better quality and intelligibility based on human audibility.

3.2.2. Analysis of loss functions

Speech enhancement results with SII-weighted loss function and traditional MSE loss function are compared on the important frequency bands for human understanding. According to the fitting function of SII, 51% importance of speech intelligibility is related to the bands between 1k-3.15kHz, and 26% intelligibility related to 1.6k-2.5kHz with the peak in 2kHz. As shown in Fig. 2, the enhanced speech from models trained by SII-weighted loss function shows better performance in those important bands. To evaluate the effectiveness of SII-weighted loss function for the speech intelligibility recovery, we calculated the mean square errors of enhanced speech on the most important band for human hearing perception in 1.6k-2.5kHz. The errors in four scenarios of BUS, CAF, PED, and STR are reduced by 1.8%, 7.3%, 5.9%, and 5.0%, respectively (under 100 training epochs).

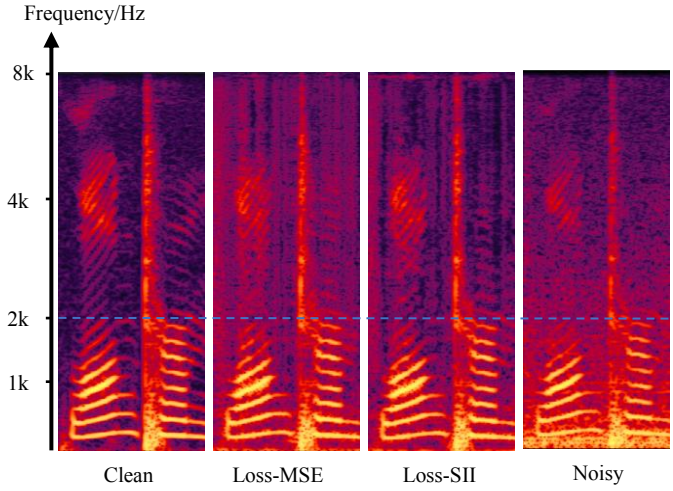

Fig. 2 Comparison of speech enhancement spectrogram with proposed SII-weighted loss function.

4. CONCLUSIONS

This paper proposes a spatial-temporal graph convolutional network for multichannel speech enhancement, aiming at improving the quality of speech against noisy background without any prior information of array position and acoustic scene. To better extract spatial, temporal, and spectral correlations of multichannel noisy signals, a graph convolution block is designed with learnable adjacency matrix to capture the inter-channel spatial dependency. And it is embedded with time-frequency convolution block as the ST module to fuse the multi-dimensional correlation features for target speech estimation. In addition, a SII-weighted loss function is used to assign different attentions to the speech frequency bands related to speech intelligibility. Experimental results have demonstrated the superiority of our model in multiple noisy scenarios. Some of the enhanced audio clips can be found from https://ahuei.github.io/stgcsen.

5. REFERENCES

[1] H. Lim, C. Kim, E. Ekmekcioglu, S. Dogan, A. P. Hill, A. M. Kondoz and X. Shi, "An approach to immersive audio rendering with wave field synthesis for 3D multimedia content," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 76-80.

[2] M. R Bai, J. Ih, and J. Benesty, Acoustic array systems: theory, implementation, and application, John Wiley & Sons, 2013.

[3] Petr Cerva, Jan Silovsky, Jindrich Zdansky, Jan Nouza and Ladislav Seps, "Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives," *Speech Communication*, vol. 55, no. 10, pp. 1033–1046, 2013.

[4] M. I. Saryuddin Assaqty, Y. Gao, A. Musyafa, W. Wen, Q. Wen and N. Juliasari, "Independent Public Video Conference Network," in *Proc. IEEE iSemantic*, 2020, pp. 142-148.

[5] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language processing*, pp. 260–276, 2009.

[6] K. Buckley, "Spatial/Spectral filtering with linearly constrained minimum variance beamformers," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 249-266, 1987.

[7] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[9] L. Sun, J. Du, L. Dai and C. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136-140.

[10] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, 2017, pp. 1-5.

[11] S. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Proc. Interspeech*, 2017, pp. 1993-1997.

[12] S. Guo, Y. Lin, N. Feng, C. Song and H. Wan, "Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 922-929.

[13] P. Tzirakis, A. Humar and J. Donly, "Multi-Channel Speech Enhancement Using Graph Neural Networks," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3415-3419.

[14] 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," *arXiv preprint arXiv:1903.03107*, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[16] T. N. Kipf and M. welling, "Semi-Supervised Classification with Graph Convolutional Networks," in I*nternational Conference on Learning Representations*, 2017, pp. 1-14.

[17] B. W. Y. Hornsby, "The Speech Intelligibility Index: What is it and what's it good for?" *The Hearing Journal*, vol. 57, no. 10, pp. 10-17, 2004.

[18] 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)*, 2015, pp. 504–511.

[19] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech*, 2018, pp. 3229-3233.

[20] W. Rao, Y. Fu, et al., "INTERSPEECH 2021 ConferencingSpeech Challenge: Towards Far-field Multi-Channel Speech Enhance-ment for Video Conferencing," *arXiv preprint arXiv:2104.00960v1*, 2021.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[22] 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.