INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A Causal U-net based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement

Xinlei Ren, Xu Zhang, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, Bing Yu

Kuai Shou Technology Co. Beijing, China

renxinlei@kuaishou.com

Abstract

People are meeting through video conferencing more often. While single channel speech enhancement techniques are useful for the individual participants, the speech quality will be significantly degraded in large meeting rooms where the far-field and reverberate conditions are introduced. Approaches based on microphone array signal processing are proposed to explore the inter-channel correlation among the individual microphone channels. In this work, a new causal U-net based multiple-in-multiple-out structure is proposed for real-time multi-channel speech enhancement. The proposed method incorporates the traditional beamforming structure with the multi-channel causal U-net by explicitly adding a beamforming operation at the end of the neural beamformer. The proposed method has entered the INTERSPEECH Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing. With 1.97M model parameters and 0.25 real-time factor on Intel Core i7 (2.6GHz) CPU, the proposed method has outperforms the baseline system of this challenge on PESQ, Si-SNR and STOI metrics.

Index Terms: multi-channel speech enhancement, U-NET, encoder-decoder, deep learning

1. Introduction

Video conferencing has become a new normal as people collaborate from geographically disjoint offices more often than before. For large conference rooms, circular and linear microphone arrays are usually employed to achieve better speech capture and enhancement. Traditional signal processing based beamforming techniques [1][2] have been proposed, where the multi-channel optimal filters are estimated to only boost the signals coming from the desired target direction while attenuating the interferences from the other directions.

In recent years, deep learning based speech enhancement (SE) approaches have achieved significant improvement over the signal processing based methods, especially for single-channel SE [3][4][5]. Motivated by this success, the multi-channel deep learning based speech enhancement methods are proposed. Many of these methods incorporate the deep neural network (DNN) with the traditional beamforming broadly known as the neural beamformers. In [6][7], a single channel mask is produced by the single channel deep noise suppression network for each channel. The multi-channel spatial covariances of the noise signals are calculated for MVDR beamforming using these single-channel masks. [8] employs deep learning to train the first network to exploit the inter-channel phase and level pattern features for the two-channel inputs. These features are served as the additional directional features to the second multi-channel neural source separation network. [9] proposes an all deep learning MVDR (ADL-MVDR) beamformer that trains the Conv-Tasnet[10] variant with the interaural phase difference (IPD) alongside with the log-power spectra (LPS) as the input features. These methods either require additional spatial features such as the inter-channel phase and level patterns or only employ the DNN to estimate a single channel mask.

More recently, the U-net structure, previously achieved state-of-the-art music separation [11] performance, is employed for single and multi-channel speech enhancement. Methods described in [12] employ the Wave-U-net [3] structure to estimate a single channel speech from the input multi-channel noisy speech signals. Time-frequency domain U-net is also proposed in [13] where the Channel-Attention is placed between the encoder and the decoder to produce the estimated single channel speech and noise. Comparing to the neural beamformers, these U-net approaches generally do not require explicit spatial features, where the output single channel clean speech estimation is created directly from the input multi-channel (MISO).

In this paper, a causal multiple-in-multiple-out (MIMO) U-net neural beamformer is proposed to combine the MISO U-net with the beamforming structure. The contribution of this paper can be summarized as follows. First, the proposed method extends the causal U-net[14][15] structure previously proposed for single channel speech enhancement to produce multi-channel time-frequency domain complex beamforming filter. Second, the proposed method incorporates the traditional beamforming structure with the multi-channel causal U-net by explicitly adding a beamforming operation at the end of the neural beamformer. Comparing to the existing neural beamformers, the proposed method does not require explicit spatial feature such as the IPDs. In comparison to the existing MISO U-nets, the proposed method directly outputs the complex beamforming filters to work with the beamforming layer. The evaluation results show improved PESQ, Si-SNR and STOI scores over the MISO U-net using the same dataset and U-net configurations. The proposed method has also entered the INTERSPEECH Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing(ConferencingSpeech 2021 Challenge). With 1.97M model parameters and 0.25 real-time factor on Intel Core i7 (2.6GHz) CPU, the proposed method has outperformed the baseline system of the ConferencingSpeech 2021 Challenge on PESQ, Si-SNR and STOI metrics.

2. Formulation of the problem

The signal recorded by the m^{th} microphone array can be represented by:

$$y_m(t) = x(t) * h_m(t) + n_m(t), m = 0, 1..., M-2, M-1 \quad (1)$$

where m represents microphone index, M represents the number of microphones, $y_m(t)$ represents the signal recorded by the m^{th} microphone, $x(t)$ represents the clean speech signal, $h_m(t)$ represents the transfer function from the clean speech to the m^{th} microphone, $n_m(t)$ represents the noise signal recorded by the m^{th} microphone, and $*$ denotes the convolution

http://dx.doi.org/10.21437/Interspeech.2021-1457

operator. In the time-frequency domain transformed by Short-Time Fourier Transform (STFT), (1) can be expressed as:

$$\mathbf{Y}_m(l,f) = \mathbf{X}_m^{early}(l,f) + \mathbf{X}_m^{late}(l,f) + \mathbf{N}_m(l,f) \quad (2)$$

where

$$\mathbf{X}_m^{early}(l,f) = \mathbf{X}(l,f) \cdot \mathbf{H}_m^{early}(l,f) \quad (3)$$

$$\mathbf{X}_m^{late}(l,f) = \mathbf{X}(l,f) \cdot \mathbf{H}_m^{late}(l,f) \quad (4)$$

$\mathbf{X}_m^{early}(l,f)$ represents the direct and early responses of the room RIR, $\mathbf{X}_m^{late}(l,f)$ represents the late reverberation of the room RIR. l represents the frame index, f represents the frequency index and $\cdot$ represents the multiplication operation. The proposed approach estimates the clean speech as follow:

$$\hat{\mathbf{X}}^{early}(l,f) = \sum_{m=0}^{M-1} \{\mathbf{Y}_m(l,f) \cdot \mathbf{W}_m(l,f)\} \quad (5)$$

where $\mathbf{W}_m(l,f)$ is the estimated complex beamforming filters by our system. After transforming $\hat{\mathbf{X}}^{early}(l,f)$ by inverse Short-Time Fourier Transform(iSTFT), the enhanced time domain signal is estimated.

3. Proposed system

3.1. System overview

Figure 1 presents the architecture of the proposed system, which combines a beamforming structure and a causal U-net to achieve real-time multi-channel speech enhancement.

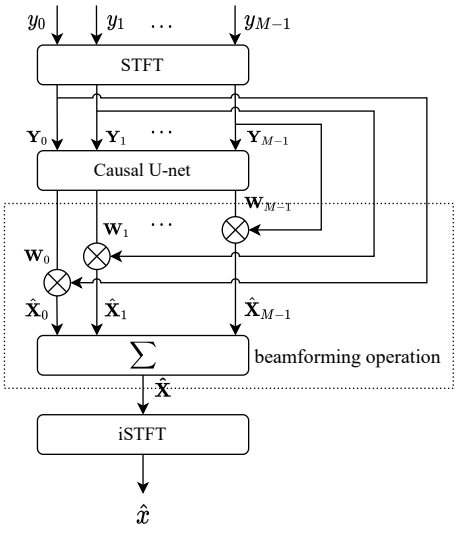

Figure 1: *The proposed system based on causal U-Net.*

The input to the proposed system is a $T = 4$ seconds multi-channel time domain audio signal $y \in \mathbb{R}^{M \times (T \times S)}$, where $S = 16000$ is the sample rate of the audio we used. After taking a STFT operation (512 points frame length and 256 points hop size), a time-frequency domain representation $\mathbf{Y} \in \mathbb{C}^{M \times L \times F}$ is obtained, where $L = 249$ is the frame number, and $F = 256$ is the number of frequency bins which means that only the first 256 points of STFT complex spectrograms are used. The spectrograms are then transposed to $\mathbf{Y} \in \mathbb{R}^{2F \times L \times M}$ by concatenating the real and imaginary parts over frequency axis. The output of the network is $\mathbf{W} \in \mathbb{R}^{2F \times L \times M}$, transposing and reshaping it, a complex mask matrix $\mathbf{W} \in \mathbb{C}^{M \times L \times F}$ is obtained.

Due to the success of the traditional signal beamforming for multi-channel speech enhancement, the proposed approach performs enhancement just like that of the beamforming.

3.2. Model structure

The causal U-net[15] is used as our neural network, which is a well-known encoder-decoder architecture and use no future information. The encoder consists of 8 Conv2d blocks for extracting high-level features gradually, and the decoder consists of 8 Conv2DTranspose blocks for reconstructing the size of input features from the output of the encoder. Between the encoder and decoder, skip connections are used to concatenate each layer in the encoder with its corresponding layer in the decoder. Figure 2 shows an example of causal convolutions U-net[15] with kernel size 2 and stride 1, which only utilizes the current and history information.

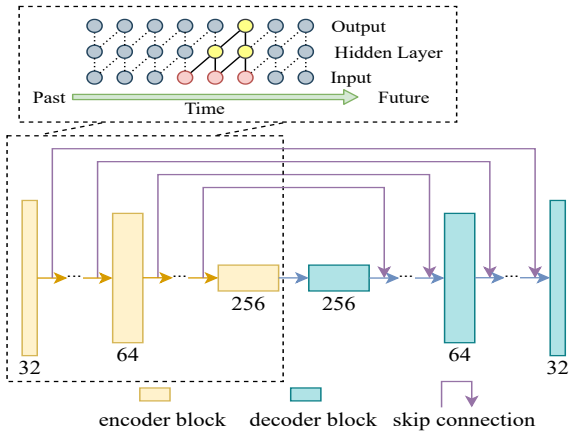

Figure 2: *An example of causal convolutions U-Net.*

Table 1: *The configuration of each Conv2d layer in the encoder*

Layer	Filter number	Kernel	Stride
Conv2d^{1st}	32	$(6,2)$	$(2,1)$
Conv2d^{2nd}	32	$(6,2)$	$(2,1)$
Conv2d^{3rd}	64	$(7,2)$	$(2,1)$
Conv2d^{4th}	64	$(6,2)$	$(2,1)$
Conv2d^{5th}	96	$(6,2)$	$(2,1)$
Conv2d^{6th}	96	$(6,2)$	$(2,1)$
Conv2d^{7th}	128	$(2,2)$	$(2,1)$
Conv2d^{8th}	256	$(2,2)$	$(1,1)$

The configuration of each Conv2d layer in the encoder are presented in Table 1. The encoder has 8 Conv2d blocks. Each block is consisted by a Conv2d layer followed by a batch normalization layer, a dropout layer and a LeakyReLU activation function. The dropout rate is set to 0.5. The corresponding decoder also has 8 blocks. Each of the decoder block is identical to the encoder block except for replacing the Conv2d layer with the Conv2DTranspose layer. The input shape of each layer in the encoder-decoder architecture is specified in [BatchSize, Frequency, Frame, Channel] format. The causal U-net is followed

by a dense layer with 512 units to estimate the complex mask.

In the implementation of causal U-net, K zeros frames are padded in front of input feature, and the last K frames are discarded for the output mask, the frame number K is decided by the kernel size over the time dimension, which is set to 8 in the proposed model.

3.3. Beamforming

After estimating $\mathbf{W} \in \mathbb{C}^{M \times L \times F}$, The enhance process is performed as that of the traditional beamforming. The complex spectrogram of noisy signal $\mathbf{Y} \in \mathbb{C}^{M \times L \times F}$ is multiplied by $\mathbf{W} \in \mathbb{C}^{M \times L \times F}$ with element-wise multiplication, then the result is summed over the channel axis to estimate a single-channel enhanced complex spectrogram $\hat{\mathbf{X}} \in \mathbb{C}^{L \times F}$. A time domain enhanced signal $\hat{x}$ is obtained after a iSTFT operation on $\hat{\mathbf{X}} \in \mathbb{C}^{L \times F}$. From the perspective of this process, $\mathbf{W} \in \mathbb{C}^{M \times L \times F}$ can be understood as the weights of a filter-and-sum beamformer.

3.4. Loss function

The mean absolute error(MAE) defined in time domain is used as loss function in the proposed system. Comparing to the time domain mean square error(MSE) used in [16], the MAE pays more attention to small signals and have better noise suppression performance. In addition to the speech component, the noise component is also added to the loss function, which is beneficial for improving the noise robustness of proposed model. And these two items have the same weight:

$$loss_{mae} = |x - \hat{x}| + |n - \hat{n}| \tag{6}$$

where x and $\hat{x}$ are clean speech and estimated clean speech respectively, n and $\hat{n}$ are noise and estimated noise respectively. They satisfy the following relationship:

$$x + n = \hat{x} + \hat{n} = y \tag{7}$$

where y is noisy speech.

3.5. Post-filter

To further suppress the residual noise, we apply a Wiener filter[17] with the noise estimation algorithm based on minimum tracking as the post-filtering strategy, which is commonly employed in the traditional speech enhancement system. The window length of the minimum tracking algorithm is set to 4 seconds, and the coefficients of the Wiener filter is calculated by:

$$G(l, f) = \frac{\lambda_y(l, f) - \lambda_{\hat{n}}(l, f)}{\lambda_y(l, f)} \tag{8}$$

where $\lambda_y(l, f)$ is power spectral density of the noisy signal and $\lambda_n(l, f)$is the estimated noise power spectral density. l and f denote the frame index and the frequency index respectively.

4. Experiments and Results

4.1. Datasets

The clean speech dataset includes four open source speech databases: AISHELL-1[18], AISHELL-3[19], VCTK[20] and Librispeech(train-clean-360)[21]. The speech utterances with SNR larger than 15dB are selected for training. The total duration of clean training speech is around 550 hours. The noise dataset is composed of MUSAN[22] and Audioset[23], the total duration is around 120 hours. Besides these two open source databases, 98 real meeting room noise files recorded by high fidelity devices are also used. The speech and noise files selected by ConferencingSpeech 2021 challenge [24] from these databases are used for data augmentation.

We generate 20000 multi-channel room impulse responses (RIRs) based on the configuration of the microphone array with the image method[25], and the range of RT60 is between 0.1 and 1.2 s. The room size ranges from $3 * 3 * 3\ m^3$ to $8 * 8 * 3\ m^3$. The microphone array is randomly placed in the room with height ranges from 1.0 to 1.5 m. The sound source, including speech and noise, comes from any position in the room with height ranges from 1.2 to 1.9 m. The angle between two sources are wider than 20°. The distance between sound source and microphone array ranges from 0.5 to 5.0 m.

Based on the speech, noise and RIR datasets, a total of 1000 hours of multi-channel noisy data are generated, and 70% for training while 30% for validating. 2000 extra clips are generated for testing. These data are generated with following augmentations:

- **REVERBERATION:** The single-channel speech and noise are convolved with the RIRs to generate multi-channel data. The target speech preserves 50ms early reverberation, this is because the reflections arriving within 50 ms after the direct sound is actually beneficial for intelligibility. The reflections, which arrives 50 ms after the direct sound, degrades the speech intelligibility[26][27]
- **SCALING:** The amplitude of the training data is randomly selected within the range of [-50, -0.87]dB.
- **EQ:** We filter the data with various filters for simulating EQ, these filters include low-pass filter, de-emphasis filter and so on.
- **SNR:** We mix the speech and noise with the random SNR between -3 and 25 dB.

4.2. Experiments setup

Figure 3 shows the configuration of the microphone array(MA) provided by ConferencingSpeech 2021 challenge. It is a linear microphone array with non-uniformly distributed 8 microphones(marked as MA No.1).

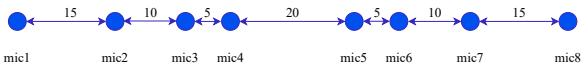

Figure 3: *The information of MA No.1. The unit in this figure is centimeter*

There are two tasks for ConferencingSpeech 2021 challenge:

- **Task1: Multi-channel speech enhancement with single microphone array.** This task is focusing on processing speech from the MA No.1. It needs to meet real-time requirements: no future information can be used; the processing time of the test clip should be less and equal than the time of the test clip and frame length should be less than or equal to 40ms.
- **Task2: Multi-channel speech enhancement with multiple distributed microphone arrays.** Five microphone arrays are provided for this task, and there is no any constraints, any algorithms can be explored with these microphone arrays.

For task1 and task2, we only develop our system on the data of MA No.1, and the system aims to estimate the clean speech of the 3rd microphone. We use the Adam optimizer to train the model. The initial learning rate is set to 0.001, and it will decay 0.5 when the validation loss does not decrease for 5 epochs.

4.3. Results

First, several causal U-net models with different inputs and outputs are evaluated. The details model configuration are shown as followed:

- **SISO-U-net:** it inputs single-channel noisy spectrogram and outputs single-channel complex mask. The single-channel mask is multiplied with noisy spectrogram to generate enhanced speech spectrogram.
- **SISO+IPD-U-net:** it inputs single-channel noisy spectrogram along with the interchannel phase difference(IPD) features[8], which are extracted from four microphone-pairs((0,4),(1,5),(2,6),(3,7)). The output is single-channel complex mask.
- **MISO-U-net:** it inputs multi-channel noisy spectrogram and outputs single-channel complex mask.
- **MIMO-U-net+BF:** it inputs multi-channel noisy spectrogram and outputs multi-channel complex mask, and a beamforming operation is applied based on the output mask. And this is our proposed system without post-filter.
- **MIMO-U-net+BF+PF:** this is the proposed MIMO-U-net network followed by post-filter.

We evaluate these models on the 2000 testing clips via four objective measurements: PESQ[28], STOI[29], Extended-STOI[30] and Si-SNR[10]. The result is presented in Table 2. By utilizing multichannel information, models with IPD information or multichannel inputs outperform the SISO model. With multichannel input, MISO model achieves higher PESQ comparing to SISO+IPD model, which means the network can automatically learn multichannel information better than IPD. By estimating multichannel masks and applying multiply and sum operations similar to beamforming, MIMO model can further improve the PESQ to 1.95, compared to that of 1.89 for MISO model. With post-filter, although the PESQ slightly decreases from 1.95 to 1.919, our internal subject listening test shows that MIMO-PF model achieves best listening experiences.

Secondly, we compare our system with the baseline on the development set provided by ConferencingSpeech 2021 challenge. The baseline is mainly a single-channel approach based on the deep learning. It includes 3 LSTM layers and a DENSE layer, and it inputs the STFT complex spectrogram of the first channel with IPD features. We use the proposed system for both task1 and task2. Table 3 represents the objective scores of both tasks on the data of MA No.1. We can see our proposed system performs better than baseline for all objective measurements.

Table 4 shows the Mean Opinion Score(MOS) evaluated by ConferencingSpeech 2021 organizer. Besides of global MOS, speech MOS (S-MOS) and noise MOS (N-MOS) are also used to evaluate speech distortion and noise reduction respectively. Confidence Interval (CI) of MOS score is provided in the table. 435 utterances are evaluated and each of which is heard by more than 20 listeners. Two rounds of subjective evaluation have been performed, the first round includes submissions from all teams and the second round includes only several top-scoring teams. The final results for each task are based on the combined results from both rounds. It's obvious that the proposed system achieves significant improvement on MOS for both task1 and task2.

The total number of parameters of the proposed neural network is $\sim$1.97M, and the processing time of one frame(16ms) is on average $\sim$4ms tested on an Intel Core i7 (2.6 GHz) CPU.

Table 2: *The objective scores on the 2000 testing clips*

	PESQ	**STOI**	**E-STOI**	**Si-SNR**
Noisy	1.278	0.728	0.587	1.893
SISO-U-net	1.841	0.844	0.740	7.475
SISO+IPD-U-net	1.855	0.847	0.746	7.501
MISO-U-net	1.890	0.852	0.758	7.959
MIMO-U-net+BF	1.950	0.861	0.764	8.008
MIMO-U-net+BF+PF (proposed)	1.919	0.857	0.759	7.935

Table 3: *The objective scores on the data of MA No.1*

	PESQ	**STOI**	**E-STOI**	**Si-SNR**
Task1				
Noisy	1.515	0.823	0.690	4.474
baseline	1.999	0.888	0.780	9.159
proposed	2.125	0.908	0.817	9.287
Task2				
Noisy	1.506	0.824	0.693	4.504
baseline	1.983	0.887	0.780	9.228
proposed	2.125	0.909	0.818	9.343

Table 4: *MOS on evaluation test set of the challenge*

	MOS	**S-MOS**	**N-MOS**	**CI**
Task1				
Noisy	2.56	2.93	3.03	
proposed	4.02	3.87	3.87	0.02
improvement	1.46	0.94	0.84	
Task2				
Noisy	2.51	2.88	2.99	
proposed	4.14	3.93	3.92	0.02
improvement	1.63	1.05	0.93	

5. Conclusions

In this paper, a causal U-net based neural beamformer is proposed for real-time multi-channel speech enhancement. With combining MIMO structure and a beamforming operation, the proposed system outperforms SISO and MISO U-net structure on PESQ, Si-SNR and STOI metrics. Besides, the proposed system has achieved better performance than the baseline system of ConferencingSpeech 2021 Challenge, and also significantly improves the MOS of noisy speech.

6. References

[1] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[2] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[3] R. Giri, U. Isik, and A. Krishnaswamy, "Attention wave-u-net for speech enhancement," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 249–253.

[4] Y. Hu, Y. Liu, S. Lv, M. Xing, S. min Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *INTERSPEECH*, 2020.

[5] J. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," in *INTERSPEECH*, 2020.

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

[7] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[8] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[9] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," 08 2021.

[10] Y. Luo and N. 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.

[11] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020, deezer Research.

[12] H. Lee, H. Y. Kim, W. H. Kang, J. Kim, and N. S. Kim, "End-to-End multi-channel speech enhancement using inter-channel time-restricted attention on raw waveform," in *Proc. Interspeech 2019*, 2019, pp. 4285–4289.

[13] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnasavamy, "Channel-Attention dense u-net for multichannel speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 836–840.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," vol. 9351, 10 2015, pp. 234–241.

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

[16] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4390–4394.

[17] P. Loizou, *Speech Enhancement: Theory and Practice*, 01 2007.

[18] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[19] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker Mandarin TTS corpus and the baselines," 2020.

[20] C. Veaux, J. Yamagishi, and K. Macdonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[21] 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)*, 2015, pp. 5206–5210.

[22] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[24] W. Rao, L. Xie, Y. Wang, T. Yu, S. Watanabe, and Z. Tan, "Far-field multi-channel speech enhancement challenge for video conferencing (ConferencingSpeech 2021)," in *ConferencingSpeech 2021*, 2021.

[25] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1976.

[26] A. Nabelek, T. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86 4, pp. 1259–65, 1989.

[27] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5434–5438.

[28] ITU, "ITU-R Rec. P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2007.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010.

[30] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.