INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Cross-channel Attention-based Wave-U-Net for Multi-channel Speech Enhancement

Minh Tri Ho[1], *Jinyoung Lee*[1], *Bong-Ki Lee*[2], *Dong Hoon Yi*[2], *Hong-Goo Kang*[1]

[1]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
[2]Artificial Intelligence Lab, LG Electronics Co., Seoul, South Korea

mtho96@dsp.yonsei.ac.kr

Abstract

In this paper, we present a novel architecture for multi-channel speech enhancement using a cross-channel attention-based Wave-U-Net structure. Despite the advantages of utilizing spatial information as well as spectral information, it is challenging to effectively train a multi-channel deep learning system in an end-to-end framework. With a channel-independent encoding architecture for spectral estimation and a strategy to extract spatial information through an inter-channel attention mechanism, we implement a multi-channel speech enhancement system that has high performance even in reverberant and extremely noisy environments. Experimental results show that the proposed architecture has superior performance in terms of signal-to-distortion ratio improvement (SDRi), short-time objective intelligence (STOI), and phoneme error rate (PER) for speech recognition.

Index Terms: Multi-channel Speech Enhancement, Wave-U-Net, Cross-Channel Attention

1. Introduction

Speech enhancement aiming to enhance the quality of the target speech signal is crucial to improving the robustness to noise for speech recognition [1, 2]. With the development of deep learning, data-driven speech enhancement approaches have shown breakthroughs when using a single microphone. In most of single channel approaches [3, 4, 5] the speech signal is first transformed into the frequency domain, after which time-frequency (TF) masks are estimated to determine the amount of noise reduction in each TF bin. However, their performance improvements are not significant in low signal-to-noise environments because of their limitations on estimating the phase spectrum. Williamson et al. [6] estimated the TF masks in the complex domain, but it was not easy to train the network. To overcome this limitation, Wave-U-Net [7] was proposed as a time domain approach. Since the Wave-U-Net model directly estimates the clean target signal waveform, it does not need to consider the phase problem and spectral efficiency caused by the time-frequency transformation with a fixed-length analysis frame.

When multiple microphones are available, the performance of speech enhancement algorithms can be further improved because of spatially related information between the microphones [8]. Statistical approaches such as beamforming [9] and multi-channel Wiener filtering [10] first estimate the direction of arrival (DOA) between microphones, then enhance the incoming signal from the estimated source direction but attenuate interference from other directions using a linear filter [8]. Although the methods are fast and lightweight, their performance and robustness are not reliable in harsh environments. Recently, deep-learning based approaches have been introduced to the first stage of enhancement [11, 12]. In addition, end-to-end models have been designed to extract both spectral and spatial features. Examples of this approach include the Wave-U-Net model [7] and time-domain convolutional autoencoder model [13]. Since end-to-end structures only mimic the behavior of beamforming at the first layer of the encoder, their capabilities for spatial filtering should be limited by the signal-to-noise ratio (SNR) of the input signal and the number of microphones. In our preliminary experiments, we found out that this simple approach would not work in extremely low SNR conditions.

To overcome the problem, we modify the Wave-U-Net structure to process each channel separately, but utilize inter-channel information using a nonlinear spatial filter designed by a lightweight attention block. Motivated by the attention block proposed in [14], we exploit inter-channel information in a more straightforward manner which directly compares information in one channel with another. Moreover, since encoding each input channel independently helps to preserve spatial information, the output of each encoding layer can be repeatedly utilized to estimate inter-channel information

Our contributions in this paper are three-fold. First, we propose a novel modification of the well-known Wave-U-Net architecture for multi-channel speech enhancement. It separately encodes each channel, then interchanges information between encoded channels after each downsampling block is efficiently combined by a convolution size of one before passing to the skip connection. Secondly, we introduce a cross-channel attention block to boost network performance by effectively exploiting spatial information of multi-channel data. To the best of our knowledge, our paper is the first work proving a model's performance in an artificial multi-channel acoustic scenario with all of the following four intricate challenges: minimum number of microphones (only two microphones with a small distance between them), varying positions of both speech and noise sources, reverberation and extremely low SNR conditions (-10 dB, -5 dB cases).

The rest of this paper is organized as follows. Section 2 gives a review of related works while section 3 describes our baseline model. Section 4 represents our proposed cross-channel attention Wave-U-Net architecture. Section 5 describes our experiment and analyses the experiment results, followed by the conclusion in the final section.

2. Related Works

With the success of deep-learning based single-channel speech enhancement approaches, much research has been done on combining models with traditional statistic-based beamforming

This work is supported and funded by LG Electronics Co., Ltd.

http://dx.doi.org/10.21437/Interspeech.2020-2548

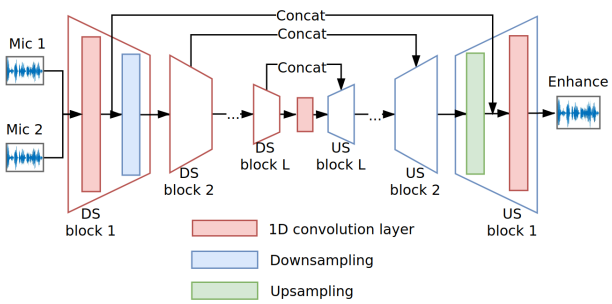

Figure 1: *Wave-U-Net structure for multi-channel data. DS stands for a downsampling block and US does for an upsampling block. L denotes the number of downsampling/upsampling blocks.*

algorithms. Typical works that manifest this idea include [11], [12] and [15]. The common strategy of these works is that the neural network tries to reduce single channel noise using TF masks at the first stage, after which beamforming is applied to linearly integrate the multi-channel signals. This approach not only gives better results than the pure statistical methods but also shows robustness in terms of various noisy types and SNR ranges. However, in this method, the neural network only learns the spectral characteristics of a single channel, not spatial information. To address this limitation, recent deep learning approaches have used inter-channel features as additional inputs to the network, such as generalized cross-correlation (GCC), interaural phase or level difference (IDP, ILD) [16, 17]. Although learning spectral information together with spatial features results in performance improvements, adding spatial features as the separated input makes it difficult to learn the mutual relationship between spatial and spectral information.

To handle multi-channel data, the author of Wave-U-Net proposed that the first layer of the network takes into account all input channels. A similar solution could be found in [13], in which the authors used a dilated convolution autoencoder instead of the U-Net structure. However, when handling all the input channels together, the first convolution layer only played the role of performing nonlinear channel fusion.

3. Wave-U-Net Baseline for Speech Enhancement

Wave-U-Net was originally developed for a singing voice source separation task by reconfiguring the U-Net structure [5] in the time domain. Based on an encoder-decoder structure, it introduces skip connections between the same layer in downsampling and upsampling blocks so that the high-level features of the decoding layer also include local features from the encoding later [7].

The single channel Wave-U-Net structure can be extended to a multi-channel structure if the number of channels of the input waveform signal is increased to correspond to the number of microphones. Therefore, the input shape of the multi-channel model is $T \times C$, where T and C are the number of audio samples and channels, respectively. The second dimension is treated as a feature map of the first convolution layer. We use this simple form of multi-channel Wave-U-Net illustrated in Fig. 1 as the baseline for the proposed model.

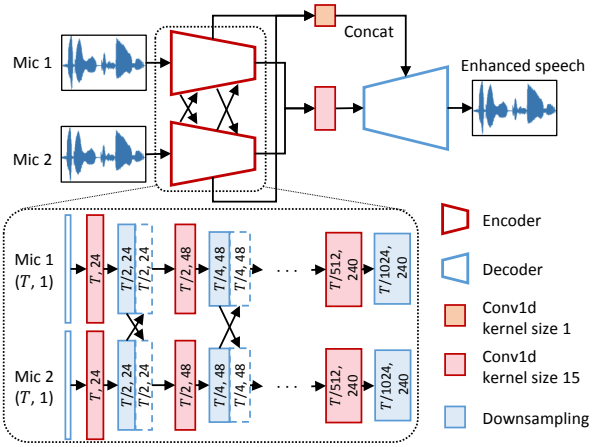

Figure 2: *Proposed Cross-channel attention-based Wave-U-Net structure for multi-channel data. Feature maps obtained at the encoding layer are interchanged between channels after processing each downsampling block.*

4. Proposed Model

In this section, we propose a cross-channel attention-based multi-channel Wave-U-Net model by modifying the baseline architecture described in Section 3. Fig. 2 illustrates the architecture of our proposed model. Although the proposed algorithm can be generalized to an arbitrary number of channels, we fix the number of channels to two for simplicity in this paper.

4.1. Encoder Structure

To provide flexibility for the processing of each channel and to explicitly utilize cross-channel relationships, the encoder processes each channel independently. Feature maps from the encoder of each channel are used as inputs to the cross-channel attention block, then are interchanged between channels. The main objective of the cross-channel attention block is to derive the relationship between two channels. Details of this block are described in subsection 4.3. At the bottleneck of the network, feature maps obtained by the encoder of each channel are concatenated, after which they are projected into one feature map using a 1-D convolution layer. When the number of channels is greater than two, one channel is chosen as a reference channel and feature maps are interchanged between the reference and other channels.

4.2. Decoder Structure

In each decoding layer, feature maps extracted from the encoding layer are fused by a 1-D convolution layer with size 1. The processing is different from the baseline structure that uses a direct skip connection between the same level layers of the encoder and decoder. The size 1 convolution has two main roles. Firstly, features from the encoder are effectively combined in a manner such that the network itself can learn to drop or bring any features. Secondly, this convolution helps to reduce the number of network parameters by half of the total feature map size. Details on the network parameters are summarized in Table 1 in Section 5.

4.3. Cross-channel Attention Block

In real-life situations, the target source position does not change much compared to interference ones; therefore, the time delay

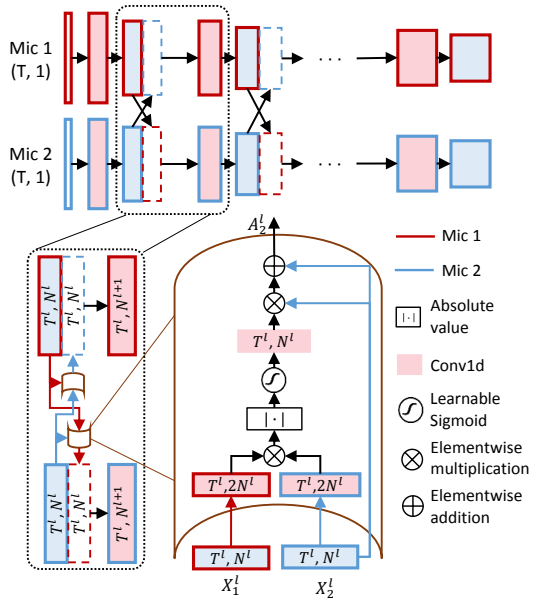

Figure 3: *Cross-channel attention block (brown box) between two encoders. The attention block is applied after every downsampling block in the encoder.* (T^l, N^l) *represents the shape of feature map tensor at downsampling block* l.

between channels in the voice active region is shorter than the one in the interference region. In addition, the power of the voice active region is likely to be higher than those of noise ones even in low SNR cases. Our proposed cross-channel attention block utilizes this characteristic to emphasize voice active regions while attenuating directive interference regions. Fig. 3 illustrates a block diagram of the proposed cross-channel attention block.

X_i^l is the feature map corresponding to the encoder of channel i. The notation $(.)^l$ is the indicator of the attention block located at the l^{th} downsampling block. X_i^l has shape $T^l \times N^l$, where T^l and N^l are the number of samples in the time domain and the number of feature maps at the downsampling block l^{th}, respectively. At first, two feature maps are put into a 1-D convolution layer with kernel size 1, followed by a hyperbolic tangent (tanh) activation function. Inspired by works in [18] and [19], this convolution layer plays a role as a linear transformation of the input tensor into an intermediate space. On the other hand, the tanh activation bounds the input tensor to the range of $[-1, 1]$, which is useful when it is used as an input to the learnable sigmoid activation function later. Afterward, the two signals are element-wise multiplied with each other. Intuitively, the multiplication operation emphasizes the regions which are slowly varying in time and have high power; thus, we expect that they would be highly related to voice active regions. Next, we feed the absolute value of the multiplication result to the input of the learnable sigmoid function:

$$\sigma_{\alpha,\beta}(x) = \frac{1}{1 + e^{-\alpha(x-\beta)}}, \tag{1}$$

where $\sigma_{\alpha,\beta}$ is the sigmoid function parameterized by two parameters α and β. The sigmoid function works as a filter to reduce the noise components. The parameter β controls the threshold highlighting signal whose value greater than it while attenuating signal with a smaller value. On the other hand, the parameter α controls the softness of the mask; a large value of α pushes the signal close to saturated values 0 and 1.

Table 1: *Network setup.* $l \in 1, 2, ..., L$ *is layer index, for baseline:* $L = 10$*, for propose model:* $L = 12$

		Baseline model	**Proposed model**
Encoder	Number of encoder	1	2
	Encoder input	(16384,2)	(16384,1)
	Number of DS block	12	10
	1Dconv kernel size	15	15
	Number of kernel of l^{th}-1Dconv layer	$24l$	$24l$
Bottleneck	1Dconv kernel size	15	15
	Number of kernel	312	264
Decoder	Number of decoder	1	1
	Number of DS block	12	10
	1Dconv kernel size	5	5
	Number of kernel of l^{th}-1Dconv layer	$24l$	$24l$

Afterward, the mask is transformed back to the original signal's space by 1-D convolution with kernel size 1 and sigmoid activation. This process can be modeled as:

$$M^l = \sigma(f(\sigma_{\alpha,\beta}[|tanh(f(X_1^l, \Phi_1^l)) \odot tanh(f(X_2^l, \Phi_2^l))|], \Phi_3^l)), \tag{2}$$

where $\odot$ denotes element-wise multiplication between two matrices. $M^l \in \mathbb{R}^{T^l \times N^l}$ represents the mask and $f(X, \Phi_i^l)$ represents the 1-D convolution with kernel size 1 on the input X of and kernel Φ_i. Signals from one layer after multiplying with the mask are added again to obtain the final output A_i^l:

$$A_i^l = M^l \odot X_i^l + X_i^l. \tag{3}$$

The residual connection has two advantages. Firstly, it helps to avoid the gradient vanishing problem frequently observed when multiple layers are used. Note that multiplying feature maps with the masking values in the range $[0, 1]$ continuously reduces the value over layers. Secondly, in case the clean signal is mis-filtered out in a certain layer, this operation keeps the information so that it can still be processed in the next layers. The output A_i^l of attention block corresponding to feature map X_i^l of channel i is "cross" concatenated to the feature map of another channel; in short, A_1^l is concatenated to X_2^l and vice versa.

5. Experiment

5.1. Database Setup

5.1.1. Room Simulation and Noise Setup

Multi-channel data used for this experiment is generated by a spatialization of single-channel data in artificially defined room conditions. We used the image source method (ISM) [20] to calculate the room impulse response to each microphone. Pyroomacoustics library [21] was used to implement the ISM for acoustic simulation with a room geometry of 8-meters length, 8-meters width, and 3-meters height. Reverberation was included with the first reflection order. Two omni-directional microphones were fixed at the middle of the wall with 8 centimeters horizontally apart. With these setups, we established a polar co-ordinate with the pole at the middle of two microphones and the polar axis was perpendicular to the wall containing 2 microphones. The clean source was located in front of two microphones with the radial distance of 1 meter and the angle varied

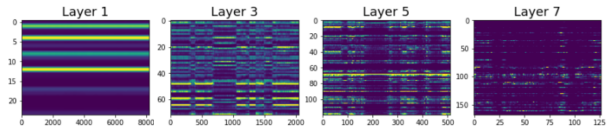

Figure 4: *From left to right: Visualization of attention mask at layer 1, layer 3, layer 5 and layer 7. Darker color indicates lower value. Noise type: NOISEX-92/destroyerops, SNR: 0 dB*

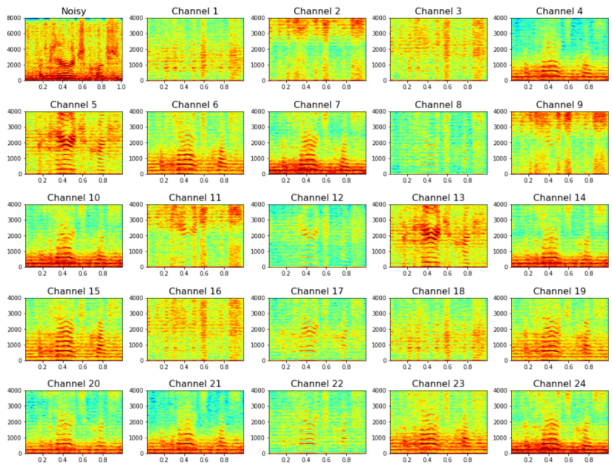

Figure 5: *Activation output's spectrograms of 24 feature maps of the first convolution layer of channel 1. Noise type: NOISEX-92/destroyerops, SNR: 0 dB*

randomly from -30 to 30 degree. The noise source was located in the room with the radial distance changing from 2 to 4 meters while the angle varying from -90 to 90 degree. The clean and noise sources were placed so as the angle difference between them was at least 15 degree.

We used the well-known TIMIT database [22] for the clean source signal. For the noise source, we selected 15 noise types from 3 noise databases: NOISEX-92, AURORA-4 [23] and DEMAND database [24]. 15 noise types were separated into 11 types for seen data and 4 types for unseen data. We covered a wide range of SNR cases, including an extremely low SNR scenario, from -10 dB to 10 dB.

5.2. Network and Training Setup

Details of the network setup for our experiment are summarized in Table 1. For the baseline Wave-U-Net structure, we kept the same architecture in [7]. The proposed model contained two separate encoders for two channels as described in Section 3. To reduce the network size, we decreased the number of downsampling blocks and upsampling blocks to 10 instead of 12.

For the training setup, we trained both the baseline and proposed model using weighted signal-to-distortion (wSDR) loss [25]. Our experiments showed that the wSDR loss performed better than the mean square error (MSE) loss because it helped to compensate for the time delay of the noisy signal. We divided our data set into a training set, a validation set and a test set with approximately 30,000 utterances (30 hours), 10,000 utterances (10 hours) and 15,000 utterances, respectively. We used ADAM optimizer with learning rate 0.0001, decay rates $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 32. Early stopping was performed if there has been no improvement on the validation set for 20 epochs.

Table 2: *Average SDR improvement (dB), STOI and PER (%)*

Method	SDRi	STOI	PER
Noisy	-	0.686	76.440
Ideal MVDR	10.134	0.834	55.125
Single channel Wave-U-Net	11.082	0.857	54.410
Multichannel WaveUNet, MSE loss	16.469	0.948	41.411
Multichannel WaveUNet, wSDR loss	17.034	0.952	40.880
Proposed	**18.032**	**0.961**	**39.323**

5.3. Experiment Result

We evaluated our proposed model via three objective measurements: signal-to-distortion ratio improvement (SDRi), short time objective intelligibility (STOI) and phoneme error rate (PER). The PER score was evaluated via the listen, attend and spell (LAS) model [26] by inputting 40-dimensional log-mel features of enhanced speech and comparing the model outputs with the ground truth labels in the TIMIT data set. The LAS model was pre-trained with two pyramidal BLSTM layers at the Listener and one attention-based LSTM layer for the AttendAndSpell module. We compared our model with the ideal MVDR beamformer, single channel Wave-U-Net which was trained with only one channel of the multi-channel data, the baseline multi-channel Wave-U-Net with MSE loss and the baseline multi-channel Wave-U-Net with wSDR loss. Experimental results were summarized in Table 2. The proposed method shows the best performance for all the three metrics.

The visualization of attention masks in Fig. 4 illustrates different mask's behaviors at different downsampling blocks. At early layers, the masks focus on certain feature map channels. For example, at layer 1 the mask highlights the $2^{nd}, 5^{th}, 9^{th}, 13^{th}$ channels of the feature map. Among 24 feature map channels of the first downsampling block shown in Fig. 5, these channels highlighted voice active region where it has high energy and small amount of noisy component. In addition, the 2^{nd} and 9^{th} channels mainly present information of the high frequency (from 3-4kHz) part of speech, which relates to the plosive sound. The 5^{th} and 13^{th} channels contain harmonic information of speech, but disregard the low-frequency noise region. At deeper layers, the shape of masks gradually changes to handle time delay related information, proving that the masks try to synchronize time delay between two channels. Masks at the 3^{rd} and 5^{th} layers are more selective in voice active region where only a few channels are highlighted. Moreover, masks of the layers containing low-frequency information such as the 7^{th} layer have very low value, which matches with the fact that noise power is dominated at the low-frequency region.

6. Conclusion

In this paper, we proposed a cross-channel attention-based Wave-U-Net for multi-channel speech enhancement, aiming to straightforwardly and efficiently exploit spatial information. Given the minimum number of microphones, our experimental results showed considerable improvements in terms of SDR, STOI and PER in an artificial room scenario with reverberation and extremely low SNR conditions. For future developments, we will be exploiting intra-channel temporal information by flexibly introducing recurrent layers, and re-designing input layers to address the latency issue of the model.

7. References

[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.

[2] B. T. Atmaja, M. N. Farid, and D. Arifianto, "Speech enhancement on smartphone voice recording," 2016.

[3] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7092–7096.

[4] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International conference on latent variable analysis and signal separation*. Springer, 2017, pp. 258–266.

[5] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *ISMIR*, 2017.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[7] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," *arXiv e-prints*, p. arXiv:1806.03185, Jun 2018.

[8] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[9] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[10] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[11] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6697–6701.

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

[13] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," in *Interspeech*, 2019, pp. 86–90.

[14] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," *arXiv preprint arXiv:2001.11542*, 2020.

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

[16] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[17] Z.-Q. 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, 2018.

[18] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[19] 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)*. IEEE, 2019, pp. 249–253.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," 1976.

[21] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2018.8461310

[22] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[23] D. Pearce and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[25] 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.

[26] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.