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

Attention Wave-U-Net for Acoustic Echo Cancellation

Jung-Hee Kim and Joon-Hyuk Chang

Department of Electronics and Computer Engineering
Hanyang University, Seoul, Republic of Korea
901914s@daum.net, jchang@hanyang.ac.kr

Abstract

In this paper, a Wave-U-Net based acoustic echo cancellation (AEC) with an attention mechanism is proposed to jointly suppress acoustic echo and background noise. The proposed approach consists of the Wave-U-Net, an auxiliary encoder, and an attention network. In the proposed approach, the Wave-U-Net yields the estimated near-end speech from the mixture, the auxiliary encoder extracts the latent features of the far-end speech, among which the relevant features are provided to the Wave-U-Net by using the attention mechanism. With the attention network, the echo can be effectively suppressed from the mixture. Experimental results on TIMIT dataset show that the proposed approach outperforms the existing methods in terms of the echo return loss enhancement (ERLE) for the single-talk period and the perceptual evaluation of speech quality (PESQ) score for the double-talk period. Furthermore, the robustness of the proposed approach against unseen noise condition is also validated from the experimental results.

Index Terms: Acoustic echo cancellation, Wave-U-Net, auxiliary encoder, attention mechanism

1. Introduction

In many applications such as hands-free telephones, audio/video conferencing systems, and hearing aids, acoustic echo occurs due to coupling between a loudspeaker and a microphone in a communication system. That is, if the microphone picks up the far-end speech from the loudspeaker, the far-end user hears an echo of his/her own voice. In this case, it is desirable to eliminate the echo and to deliver the clean near-end speech only to the far-end user. Furthermore, suppressing the echo has been more challenging due to the nonlinear distortion produced by the miniaturized speakers in audio devices such as mobile phones [1].

The conventional approach to the acoustic echo cancellation (AEC) is to employ adaptive filter algorithms for estimating the acoustic echo path from the loudspeaker to the microphone. Various adaptive filter based AEC algorithms have been proposed to improve the performance when the double-talk occurs, the background noise coexists, or the non-linear echo arises. To resolve the double-talk issue, the adaptive filter can be associated with the double-talk detectors to stop the filter adaptation during the double-talk, or the adaptive filter itself can be devised to be robust against the double-talk by adopting the robust criterion such as ℓ_1-norm based minimization [2]. When the echo and background noise coexist, the noise suppression module is independently developed and simply combined with the echo suppression module in a serial fashion, which gives the sub-optimal performance only since the overall performance depends on its integrated structure [3]. Moreover, the non-linear distortion due to the loudspeakers can be also introduced in the AEC system. To overcome this difficulty, several non-linear models such as the Volterra model, the Hammerstein model, and neural networks have been utilized [4]. Despite such a lot of works, the adaptive filtering approach still has not shown satisfactory results in various real environments.

Recently, deep neural networks (DNN) have been received much attention due to their complicated non-linear modeling capacity, and they have been successfully applied to various speech signal processing tasks such as speech enhancement, source separation, and AEC. In the AEC applications, the DNN based residual echo suppression (RES) was introduced to estimate the optimal RES gain by using the residual echo and far-end speech [5]. In [3], a stacked DNN model in a sequential manner, one for noise suppression and another for acoustic echo suppression, was developed to simultaneously suppress the acoustic echo and background noise. Furthermore, a bidirectional long-short term memory (LSTM) based model was also presented to estimate an ideal ratio mask for resynthesizing the near-end speech from the magnitude spectrum of the mixture signal [6]. The convolutional recurrent network (CRN) with an LSTM based speech detector was recently proposed in [7], where the CRN estimates the complex spectrogram of the near-end speech from those of the far-end speech and mixture signal, and the LSTM based speech detector estimates the activity of the near-end speech to further suppress the residual echo and noise during the single-talk period. In [8], a deep gated recurrent unit (GRU) based network was introduced with the multitask learning of estimating both acoustic echo and near-end speech.

However, the aforementioned algorithms [3, 5–8] are mostly performed in the short time Fourier transform (STFT) domain, which means that their performance can be degraded due to several reasons such as the performance dependency on the frame size and no available correct phase information [9]. To tackle this problem, several time-domain based networks have been recently proposed and they have shown superior performance relative to the STFT domain counterparts in various fields [9–11].

Inspired by the success of the time-domain approach, we propose an attention Wave-U-Net for the AEC application. The Wave-U-Net [9], operated in the time-domain, was originally devised for the audio separation, and its variations have been successfully applied in other areas [11–13]. Compared with the Wave-U-Net, the proposed approach includes an auxiliary encoder to extract the features of the far-end speech. The extracted features by the auxiliary encoder are delivered into the Wave-U-Net by exploiting the attention mechanism [13], which effectively suppresses the echo in the latent space. The effectiveness of the attention mechanism is verified by the experimental results. Furthermore, it will be also shown that the proposed approach produces a similar performance on both seen and unseen noise conditions, which verifies its robustness against unseen noise condition.

http://dx.doi.org/10.21437/Interspeech.2020-3200

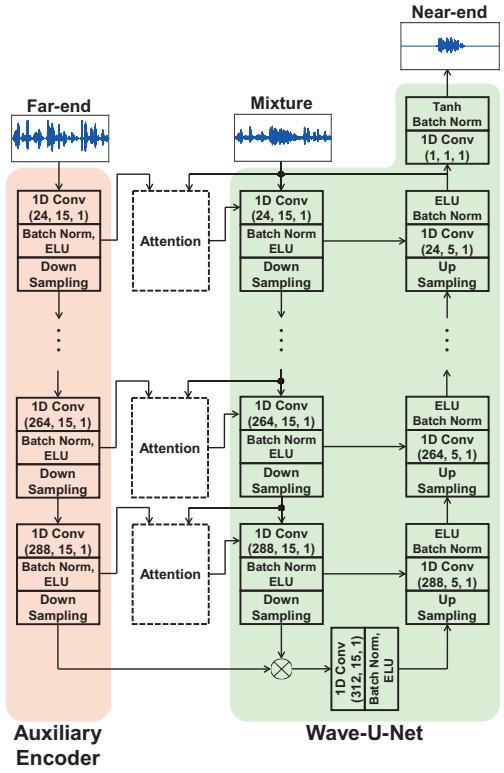

Figure 1: *Structure of the proposed attention Wave-U-Net. In 1D convolution block, three numbers in the parentheses represent number of channel, kernel size, and stride of 1D convolution, respectively.*

This paper is organized as follows: In Section 2, the problem of the AEC is briefly defined. Then, the Wave-U-Net based AEC is proposed along with the attention mechanism in Section 3. In Section 4, several experimental results are provided under various AEC environments to verify the performance of the proposed approach. Finally, conclusion remarks are given in Section 5.

2. Problem statement

In the AEC application, the mixture signal $y(n)$ is composed of acoustic echo $d(n)$, near-end speech $s(n)$, and background noise $v(n)$ as follows:

$$y(n) = d(n) + s(n) + v(n). \quad (1)$$

The acoustic echo $d(n)$, which can be also non-linearly distorted by the loudspeaker, is a modified version of the far-end speech by a room impulse response (RIR). The purpose of the AEC is to estimate the clean near-end speech $s(n)$ from the mixture $y(n)$ by jointly suppressing the acoustic echo $d(n)$ and background noise $v(n)$.

3. Proposed attention Wave-U-Net

3.1. Overall structure

In this section, the attention Wave-U-Net is proposed for the AEC application, as illustrated in Figure 1. The proposed structure consists of the Wave-U-Net, auxiliary encoder, and attention network. In the Wave-U-Net, the mixture signal is fed as an input, and the near-end speech is estimated as an output. In the auxiliary encoder, the far-end speech is encoded by using the same encoder structure as that of the Wave-U-Net and its meaningful features are given to the Wave-U-Net through the attention network, which efficiently suppresses the echo from the mixture signal.

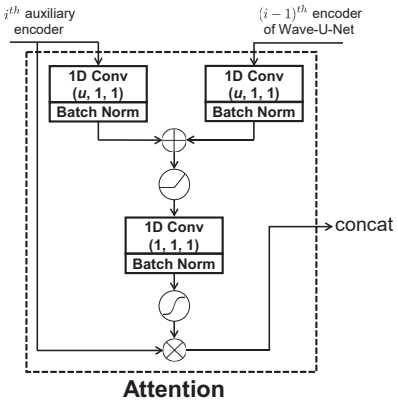

Figure 2: *Structure of attention network.*

There are several distinct characteristics between the original Wave-U-Net and the proposed attention Wave-U-Net. First, since the Wave-U-Net is originally designed for the audio source separation problem; thus, it takes a single input of the mixture signal and gives multiple outputs of separated signals. In the proposed approach, the attention Wave-U-Net accepts a mixture signal at the Wave-U-Net and a far-end speech at the auxiliary encoder, and estimates a near-end speech at the Wave-U-Net. Second, compared with the original Wave-U-Net, the proposed attention Wave-U-Net has the auxiliary encoder to yield the latent features of the far-end speech. Third, inspired by the attention mechanisms [13], the meaningful features of the far-end speech are accentuated by using the attention network. Then, the accented features of the far-end speech are concatenated with those of the mixture in the same i^{th} encoder layer of the Wave-U-Net, and the concatenated features are passed through the encoder of the Wave-U-Net to extract the relevant features in the latent space. Finally, with the extracted features from the encoder, the clean near-end speech is recovered through the decoder of the Wave-U-Net.

3.2. Attention network

As aforementioned above, the proposed architecture employs the attention mechanism [13] to identify the related features from the far-end speech in the latent space, which leads to improved performance. As shown in Figure 2, the latent features of the far-end speech in the i^{th} layer and those of the mixture in the $(i-1)^{th}$ layer are first mapped to an intermediate feature space with the same u-dimension by using a 1-D convolution with the kernel size 1 and a bias term. Here, u can be set to the minimum of the two input channel dimensions. After passing through exponential linear unit (ELU) activation function and adding them, they are additionally mapped to 1-dimensional feature space by using another 1-D convolution with the kernel size 1 and a bias term to yield the attention mask. Finally, the features of the far-end speech are element-wisely multiplied with the obtained attention mask, and then the masked features are concatenated with those of the i^{th} encoder layer of the Wave-U-Net.

3.3. Loss function

In the regression task, the signal-to-distortion ratio (SDR) has been widely used as a loss function. Therefore, the proposed algorithm minimizes the negative SDR function to train the attention Wave-U-Net. The SDR function is defined as follows:

$$SDR = 10\log_{10}\frac{\|s(n)\|^2}{\|s(n)-\hat{s}(n)\|^2}, \tag{2}$$

where $\|\cdot\|$ denotes ℓ_2-norm function, and $\hat{s}(n)$ is the estimated near-end speech.

4. Experimental results

4.1. Experiment settings

To verify the performance of the proposed attention Wave-U-Net, the similar settings as in [7, 8] were taken for the experiment. Specifically, the TIMIT dataset was utilized as the far-end and near-end speeches. Among 630 speakers in total, 100 pairs of farend-nearend speakers (i.e., 40 male-female, 30 male-male, 30 female-female) were randomly selected for training. To generate a far-end speech, three randomly chosen utterances of the same far-end speaker were concatenated. For each near-end speech, one utterance was randomly selected and extended to the same length as that of the far-end signal by padding zero. For each far-end speaker, five different far-end speeches were created, and seven different near-end speeches were generated for each near-end speaker, which results in 3500 train mixtures (about 9 hours) in total. For the validation and test, 30 pairs of far-end and near-end speakers were randomly selected from the remaining 430 speakers. In this time, five different near-end speeches were mixed with three different far-end speeches, which results in 450 mixtures for validation and test, respectively.

To model the non-linearity of the AEC system, the far-end signal $x(n)$ was further clipped and distorted as follows:

$$x_{\text{cl}}(n) = \begin{cases} -x_{\max}, & \text{if} \quad x(n) < -x_{\max}, \\ x(n), & \text{if} \quad |x(n)| \le x_{\max}, \\ x_{\max}, & \text{if} \quad x(n) > x_{\max} \end{cases} \tag{3}$$

$$x_{\text{nl}}(n) = 4\left(\frac{2}{1+\exp(-a\cdot b(n))} - 1\right) \tag{4}$$

where $x_{\max}$ was set to 0.8 times of the maximum magnitude of the original far-end signal $x(n)$, $b(n) = 1.5x_{\text{cl}}(n) - 0.3x_{\text{cl}}^2(n)$, and

$$a = \begin{cases} 4, & \text{if} \quad b(n) > 0, \\ 0.5, & \text{otherwise} \end{cases} \tag{5}$$

After that, the modified far-end signal is convolved with the randomly chosen room impulse response (RIR), which was generated by using the image method [14]. Specifically, the set of 200 RIRs was generated for training, and two other sets of 2 RIRs were created for validation and test. When generating the RIR, the specifications of Table 1 were randomly chosen. In addition, the distance of the microphone-loudspeaker was set to $1m$, and the length of the RIR was fixed to 512.

To create the train and validation mixtures, the near-end speech and the acoustic echo were mixed at five different signal-to-echo ratio (SER) levels (i.e., $\{-6, -3, 0, 3, 6\}$ dB). Furthermore, the noise was randomly cut and mixed with the near-end signal at four different signal-to-noise ratio (SNR) levels (i.e., $\{8, 10, 12, 14\}$ dB). For testing, three different SERs (i.e., $\{-1.5, 1, 5, 4.5\}$ dB) and SNRs (i.e., $\{11, 13, 15\}$ dB) were used to test the performance of the proposed approach under the mismatch conditions. Furthermore, 10 types of seen noises (i.e., bus, cafe, car, construction, kids, metro, office, railroad, restaurant, street noises) were used from the ITU-T recommendation P. 501 database [15] for train and validation, and 7 types of unseen noises (i.e., babble, bucaneer1, destroyer engine, f16, leopard, volvo, and white noises) were used from the NOISEX-92 database [16] for test.

Table 1: *Specifications for generating RIR.*

Specifications	Parameters
Room size ($m \times m \times m$)	$\{4, 6, 8, 10\} \times \{5, 7, 9, 11, 13\} \times \{3\}$
Reverberation time T_{60} (s)	{ 0.2, 0.3, 0.4 }

Finally, the performance of the proposed attention Wave-U-Net was evaluated in terms of echo return loss enhancement (ERLE) for the single talk period and perceptual evaluation of speech quality (PESQ) for the double-talk period.

4.2. Experimental results

To verify the performance of the proposed attention Wave-U-Net, we considered the following the STFT domain and time-domain based algorithms: i) stacked DNN [3] and CRN [7] for the STFT domain based model, and ii) Wave-U-Net [9], modified for the AEC problem, for the time-domain based model. In the CRN, the near-end speech detector can be employed as proposed in [7]; however, in this experiment, the CRN model was used only since it exhibited the better PESQ performance during the double-talk period than the CRN with the near-end detector as described in [7]. For the STFT domain based models, the frame size was set to 320 in the stacked DNN [3] as suggested, and 640 in the CRN [7] for the better performance. For the time-domain based models (i.e., the Wave-U-Net and proposed attention Wave-U-Net), the frame size was set to 16384 as recommended in [9]. All models were trained by using Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, but with a tuned learning rate for each algorithm. Training was performed until the validation loss stops improving for 20 epochs.

The PESQ and ERLE performances under various SER conditions are listed in Table 2 for both seen and unseen noise conditions. Note that each figure of Table 2 was obtained by averaging the results of various SNR conditions (i.e., $\{11, 13, 15\}$ dB). Compared with the STFT domain based methods [3, 7], the time-domain based methods (i.e., the Wave-U-Net [9] and proposed approach) performed better in terms of both ERLE and PSEQ. Furthermore, we can see from Table 2 that the time-domain approaches including the proposed method yielded a similar performance on both seen and unseen noise conditions while the STFT domain based algorithms, especially the stacked DNN [3], experienced a performance degradation on unseen noise condition, which reveals the robustness of the time-domain based algorithms against the unseen noise condition. When compared with the original Wave-U-Net with no attention mechanism, the proposed approach achieved better performance during both the single-talk and double-talk periods. Especially, the proposed approach effectively removes the echo from the mixture through the attention network, which leads to

Table 2: *PESQ and ERLE performance for seen and unseen noise.*

Noise Type		Seen			Unseen		
SER		-1.5dB	1.5dB	4.5dB	-1.5dB	1.5dB	4.5dB
unprocessed	PESQ	1.07	1.31	1.56	1.08	1.32	1.57
stacked DNN [3]	ERLE	32.40	34.13	35.62	25.81	25.13	24.12
	PESQ	2.24	2.46	2.65	2.18	2.38	2.56
CRN [7]	ERLE	29.93	28.31	26.38	29.04	27.00	24.64
	PESQ	2.44	2.57	2.67	2.39	2.51	2.60
Wave-U-Net [9]	ERLE	40.87	39.92	38.24	40.22	39.17	37.39
	PESQ	2.59	2.74	2.86	2.58	2.73	2.84
Proposed	ERLE	**42.20**	**41.74**	**40.36**	**41.73**	**41.10**	**39.54**
	PESQ	**2.65**	**2.81**	**2.92**	**2.64**	**2.78**	**2.88**

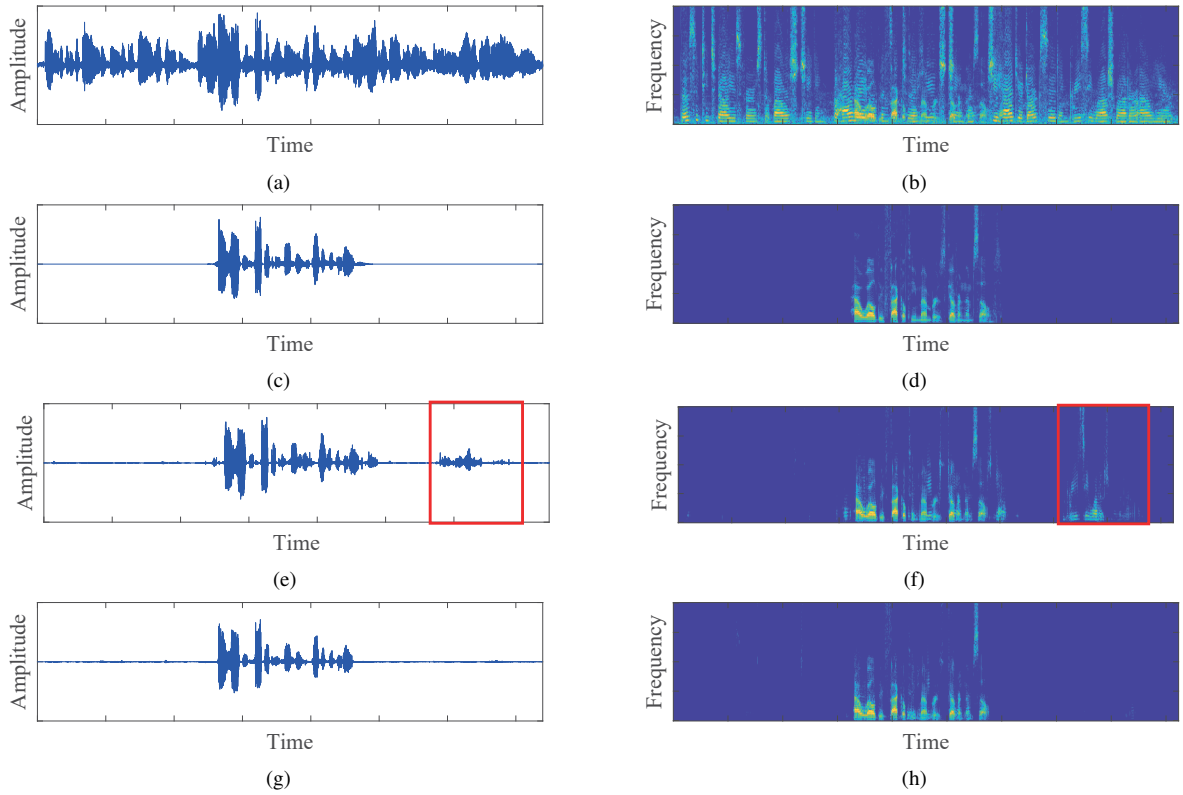

Figure 3: *Waveforms and spectrograms under -1.5dB SER, 11dB SNR, and babble noise conditions. (a) mixture, (c) clean near-end speech, (e) estimated near-end speech by the Wave-U-Net [9], and (g) estimated near-end speech by the proposed approach. (b), (d), (f), and (h) show their respective spectrogram.*

higher performance gain during the single-talk period than during the double-talk period. Figure 3 depicts the waveforms and spectrograms of the mixture, clean near-end speech, and estimated near-end speeches by the Wave-U-Net and the proposed approach on -1.5dB SER, 11dB SNR, and babble noise conditions. Figure 3 also verifies the usefulness of the attention mechanism, especially during the single-talk period (see a red rectangular box in (e) and (f) of Figure 3).

5. Conclution

In this paper, a novel attention Wave-U-Net was proposed for the AEC application. In the proposed approach, while the relevant features of the far-end speech are extracted by the auxiliary encoder and delivered by using the attention mechanism into the Wave-U-Net, the Wave-U-Net can effectively suppress the echo from the mixture with the properly extracted features. Compared with the existing algorithms including the original Wave-U-Net, the proposed attention Wave-U-Net achieved superior performance for both single-talk and double-talk periods under both seen and unseen noise conditions.

6. Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)).

7. References

[1] M. M. Halimeh, C. Huemmer, and W. Kellermann, "A neural network-based nonlinear acoustic echo canceller," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1827–1831, 2019.

[2] J.-H. Kim, J. Kim, J. H. Jeon, and S. W. Nam, "Delayless individual-weighting-factors sign subband adaptive filter with band-dependent variable step-size," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1526–1534, Jul. 2017.

[3] H. Seo, M. Lee, and J.-H. Chang, "Integrated acoustic echo and background noise suppression based on stacked deep neural networks," *Applied Acoustics*, vol. 133, pp. 194–201, Apr. 2018.

[4] J. Park and J.-H. Chang, "Frequency-domain Volterra filter based on data-driven soft decision for nonlinear acoustic echo suppression," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1088–1092, Sep. 2014.

[5] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in *Proc. INTERSPEECH*, Sep. 2015, pp. 1775–1779.

[6] H. Zhang and D. L. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. INTERSPEECH*, Sep. 2018, pp. 3239–3243.

[7] H. Zhang, K. Tan, and D. L. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. INTERSPEECH*, Sep. 2019, pp. 4255–4259.

[8] A. Fazel, M. El-Khamy, and J. Lee, "Deep multitask acoustic echo cancellation," in *Proc. INTERSPEECH*, Sep. 2019, pp. 4250–4254.

[9] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. International Society for Music Information Retrieval Conference* (ISMIR), Sep. 2018, pp. 334–340.

[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, Aug. 2019.

[11] X. Hao, X. Su, Z. Wang, H. Zhang, and Batushiren, "UNet-GAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," in *Proc. INTERSPEECH*, Sep. 2019, pp. 1786–1790.

[12] T. Nakamura and H. Saruwatari, "Time-domain audio source separation based on wave-u-net combined with discrete wavelet transform," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), May 2020, pp. 386–390.

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

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[15] ITU-T, "Test signals for use in telephonometry," *International Telecommunication Union*, 2007.

[16] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993.