ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414753

# U-CONVOLUTION BASED RESIDUAL ECHO SUPPRESSION WITH MULTIPLE ENCODERS

*Eesung Kim, Jae-Jin Jeon, Hyeji Seo*

AI R&D Team, Kakao Enterprise
235, Pangyoyeok-ro, Bundang-gu, Seongnam-si, Gyeonggi-do 13494, Korea

## ABSTRACT

In this paper, we propose an efficient end-to-end neural network that can estimate near-end speech using a U-convolution block by exploiting various signals to achieve residual echo suppression (RES). Specifically, the proposed model employs multiple encoders and an integration block to utilize complete signal information in an acoustic echo cancellation system and also applies the U-convolution blocks to separate near-end speech efficiently. The proposed network affords an improvement in the perceptual evaluation of speech quality (PESQ) and the short-time objective intelligibility (STOI), as compared to baselines, in scenarios involving smart audio devices. The experimental results show that the proposed method outperforms the baselines for various types of mismatched background noise and environmental reverberation, while requiring low computational resources.

***Index Terms***— Acoustic Echo Cancellation, Residual Echo Suppression, End-To-End Neural Network

## 1. INTRODUCTION

The recent increase in the use of personal smart speakers and telecommunication systems has necessitated the development of improved acoustic echo cancellation (AEC) or acoustic echo suppression (AES) algorithms for eliminating acoustic echo caused by acoustic coupling between the loudspeaker and near-field microphone. Although algorithms based on linear adaptive filters have been proposed [1], acoustic echo cannot be completely eliminated via the linear filtering method alone; this is due to the nonlinear echo resulting from factors such as the nonlinear response of the power amplifier and misalignment of the nonlinear acoustic transfer function. To suppress such nonlinear echo, residual echo suppression (RES), which is performed after linear adaptive filtering, is proposed [2–7].

RES algorithms based on deep learning models that can handle complex nonlinear relationships have exhibited promising performances [4–7]. In [4], the authors estimated the gain of RES using a deep neural network (DNN). Several studies [5, 7] have considered RES as a source separation problem, separating near-end speech from various signals in the AEC system. In [5], the authors used bidirectional long-term memory (BLSTM) to model context dependence, predict masks using features extracted from mixed signals and the near-end speech, and then reconstruct short-range speech signals using them. In [6], a CRN is trained in spectral mapping for estimating the spectrogram of the speech.

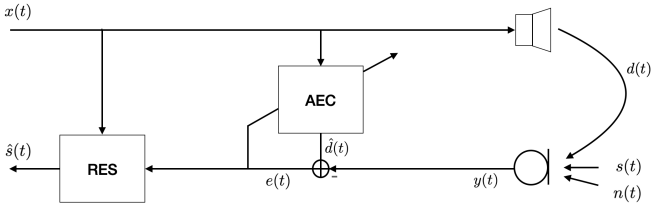


**Fig. 1**. Diagram of acoustic echo scenario.

Recently, time-domain separation systems with an adaptive front-end have gained popularity for source separation tasks [8]. Adaptive front-end approaches (e.g. convtasnet [8]) employ a convolutional layer to extract latent representations, which can be learned jointly with the mask estimation network. In [7], a modified convtasnet-based end-to-end neural network demonstrated promising results, as compared to previous baselines for RES. However, the system used limited signals as inputs for the network. In existing literature of end-to-end neural network, there is a lack of studies focusing on the various signals in an AEC system.

To address this, we propose an efficient end-to-end neural network for RES. Specifically, we use multiple encoders and depthwise separable convolution (DWS) [9] and the U-Convolution block (U-Convblock) [10]. The novelty of this study lies in the use of multiple encoders and an effective integration block using DWS convolution to incorporate output and echo signals of the AEC as well as microphone and reference signals in the latent space. Additionally, we apply U-Convblock, which is a core component for effectively estimating masks for RES. We evaluate the proposed method in terms of the short-time objective intelligibility (STOI) [11], perceptual evaluation of speech quality (PESQ) [12], and computational resources required. Results of the experiment indicate that the proposed method consistently achieved higher performances than the baselines under various conditions involving mismatched noise and reverberation, all while utilizing relatively limited computational resources.

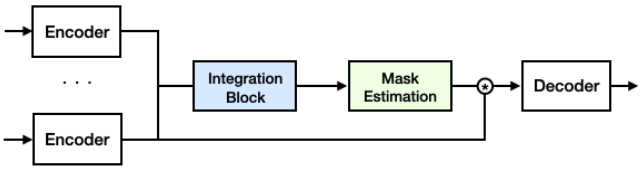


**Fig. 2**. Overview of proposed method with multiple encoders for suppressing residual echo.

## 2. PROBLEM DESCRIPTION

Let $x(t)$ denote the far-end signal and $s(t)$ and $n(t)$ denote the near-end speech and background noise, respectively, for a time index $t$. The AEC output signal, $e(t)$, and the microphone signal, $y(t)$, are denoted as follows:

$$y(t) = d(t) + s(t) + n(t) \tag{1}$$

$$e(t) = (d(t) - \hat{d}(t)) + s(t) + n(t), \tag{2}$$

where $d(t)$ is the echo signal, and $\hat{d}(t)$ is the linear echo estimate produced via a linear AEC system. Specifically, we define echo cancellation as a problem of eliminating echo in scenarios involving music played via a smart speaker. In this case, far-end speech, $x(t)$, and near-end speech, $s(t)$, represent the music and the speech of the user, respectively. A diagram of echo scenario is described in Fig. 1.

## 3. METHOD

The proposed model consists of multiple encoders, an integration block, a mask estimation network, and a decoder. The encoders perform 1-D convolution on multiple sources to obtain the corresponding latent representations. These representations are then incorporated by applying the integration block to model information regarding the sources signals. The mask estimation network computes near-end speech masks, which are multiplied to one of the source representations to obtain the latent representation estimate of near-end speech. The decoder reconstructs the estimated latent representation of near-end speech to a waveform, $s(t)$, by performing deconvolution. A schematic of the proposed model is shown in Fig. 2. For the training objective, instead of the Si-SDR loss function, commonly used for source separation, we use the time-domain logarithmic mean square error [13]:

$$L^{(T-LMSE)}(s(t), \hat{s}(t)) = 10 \log_{10} \sum_t |s(t) - \hat{s}(t)|^2 \tag{3}$$

### 3.1. Multiple 1-D convolution encoders

Recent studies [14] show that methods using multiple signals as the input are more effective in estimating near-end speech masks for improving RES performance. In this study, we introduce multiple encoders for utilizing several source signals related to echo suppression; these include the echo estimate, $\hat{d}(t)$; AEC output, $e(t)$; microphone, $y(t)$; and reference $x(t)$ signals. Each source signal transforms the latent representation independently via the corresponding 1-D convolution encoder.

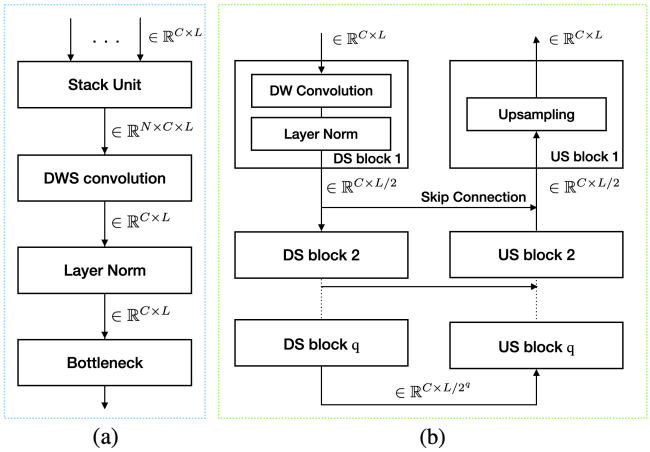


**Fig. 3**. Block diagram of (a) proposed integration block and (b) U-convolution block as mask estimation network.

### 3.2. Integration block for multiple latent representation

The integration block is designed to combine information pertaining to multiple sources in a latent space. We utilize DWS convolution to efficiently capture the spatial and input-wise correlation among several latent representation [9]. We first stack the latent representation derived by encoders along new dimensions and then perform DWS convolution using the channel followed by the bottleneck layer and layer normalization, as described in Fig. 3(a). For the layer normalization of all blocks, we use cumulative layer normalization (cLN) to satisfy causal processing [8].

### 3.3. Mask estimation network with U-convolution block

U-Convblock was recently introduced in [10]. In a U-Convblock, temporal information can be extracted at multiple resolutions using downsampling and upsampling blocks, as described in Fig. 3(b). Downsampling (upsampling) consists of $Q$ downsampling (upsampling) blocks, each of which halves (doubles) the time resolution while maintaining the number of features by using DWS convolution. A skip connection exists between downsampling and upsampling streams, and all the convolution layers are followed by the rectified linear unit (ReLU). Finally, the estimated mask is multiplied element-wise with the encoded AEC output representation, and deconvolution was then adopted to reconstruct the estimated near-end speech.

## 4. EXPERIMENTS

### 4.1. Dataset

To evaluate the proposed model, we simulated the environment using a smart speaker with Librispeech [15] and Musan database [16], as in [7]. For the near-end speech, a training set of 72000 utterances lasting for 5 s and a validation set comprising 720 utterances lasting 5 s and a test set comprising 360 utterances lasting 5 s were generated from randomly selected speakers using the Librispeech training, development, and test set, respectively. For the reference (playback) signal, we first randomly split all music datasets into 7 training, 2 validation, and 1 test samples, and training is then commenced based on the validation and test sets over the Musan dataset.

We performed experiments for simulated data with various noise types, near-end signal-to-noise ratios (SNRs), and near-end signal-to-echo ratios (SERs). The noise is randomly sampled from White and Babble noises from the NOISEX-92 database [17], and the relative SNR between the near-end speaker and the noise is randomly sampled between [10, 20] dB, with respect to the near-end speaker for simulating an environment in which smart speakers are used. In total, 72000 pairs (100 hours) of utterances for near-end and reference music signals are prepared as training data. All utterances are sampled at 16 kHz.

To simulate the nonlinearity that arises from power amplifiers and loudspeakers, we adopted a clipping function [7, 18] and a memory-less sigmoidal function [7, 19]. The hard clipping function [18] is defined as

$$x_{hard}(t) = \begin{cases} -x_{max}, & x(t) < -x_{max} \\ x(t), & |x(t)| \leq x_{max} \\ x_{max}, & x(t) > x_{max} \end{cases}, \quad (4)$$

where $x_{max}$ was set to 80% of the maximum volume of $x(t)$. To model the nonlinear characteristics of the loudspeakers, the memoryless sigmoidal function was used:

$$x_{NL}(t) = \gamma \cdot \left( \frac{2}{1 + \exp(-a \cdot b(t))} - 1 \right), \quad (5)$$

where

$$b(n) = 1.5 \times x(t) - 0.3 \times x(t)^2, \quad (6)$$

The sigmoid gain parameter $\gamma$ is set to 2, and the sigmoid slope is set to $a = 4$ if $b(t) > 0$; otherwise, it is set as $a = 0.5$.

To investigate the generalization of the room impulse response (RIR), we extend the echo path after the nonlinearity using 100 randomly selected rooms with lengths widths of within the range [3, 8] m; the height is fixed at 3 m, and the reverberation time is $T_{60} = 200ms$. Assuming a smart speaker scenario, the loudspeaker is placed at a fixed location at a distance of 20 cm from the microphone. In addition, 2 smart speakers were randomly placed in each room, and each randomly placed smart speaker contained 5 near-end speakers. Thus, a total of 1000 RIR pairs were created, among which 800 were used for training, 100 were used for validation, and 100 were used for testing. Room impulse response (RIR) filters were simulated using the image method [20] and the gpuRIR toolbox [21].

### 4.2. Experiment Configurations

All the models were trained for 120 epochs. The learning rate is set to $1e^{-3}$. Early stopping is applied if there are no improvements in the validation loss after 20 successive epochs. Adam [23] is used as the optimizer. For the linear AEC system, we adopted the multi-delay block frequency algorithm, which is proposed in [22], implemented in SpeexDSP [24]. For the encoder and decoder convolutions, we use a filter length of 21 and 512 bases. For each configuration of U-Convblock, the number of U-Convblock $B$ and $Q$ is set to 8 and 5. The other parameters are the same as those in [10]. We implemented all the baselines, i.e., DNN [4], BLSTM [5] and Convtasnet [7]-based RES models (Tasnet-MI), trained under identical conditions. The proposed U-Convblock with Multiple Encoders is denoted as UCME. In addition, multiple inputs of $e(t)$, $\hat{d}(t)$, and $y(t)$ are denoted as UCME-3M; multiple inputs of $e(t)$, $\hat{d}(t)$, and $x(t)$ are denoted as UCME-3R; and multiple inputs of $e(t)$, $\hat{d}(t)$, $x(t)$, and $y(t)$ are denoted as UCME-4.

### 4.3. Evaluation

To evaluate speech quality and intelligibility, we compare the PESQ and STOI scores. As metrics of computational complexity, the real-time factor (RTF), number of giga floating point operations (GFLOPs) executed, and model size are used. All the experiments were analyzed using an Intel Xeon E5-2695 v3 @ 2.30 GHz CPU.

### 4.4. Experimental Results

To examine the benefits of the proposed method, we conduct three experiments: (I) an experiment to determine suitable multiple input combinations, (II) an experiment to compare

| | AEC [22] + UCME-3M | | AEC [22] + UCME-3R | | AEC [22] + UCME-4 | |
|---|---|---|---|---|---|---|
| SER | PESQ | STOI | PESQ | STOI | PESQ | STOI |
| -20dB | 2.75 | 0.85 | 2.91 | 0.88 | **3.01** | **0.90** |
| -15dB | 3.03 | 0.87 | 3.27 | 0.93 | **3.34** | **0.94** |
| -10dB | 3.38 | 0.90 | 3.48 | 0.94 | **3.49** | **0.95** |
| Avg | 3.05 | 0.87 | 3.22 | 0.92 | **3.28** | **0.93** |

**Table 1**: PESQ and STOI scores of the proposed model based on various multiple inputs conditions.

| | | | AEC [22] | | AEC [22] + DNN [4] | | AEC [22] + BLSTM [5] | | AEC [22] + Tasnet-MI [7] | | AEC [22] + UCME-4 (Proposed) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Noise | SNR | SER | PESQ | STOI | PESQ | STOI | PESQ | STOI | PESQ | STOI | PESQ | STOI |
| clean | | -20dB | 1.32 | 0.49 | 1.63 | 0.61 | 1.91 | 0.68 | 2.63 | 0.83 | **3.01** | **0.90** |
| | | -15dB | 1.55 | 0.60 | 2.00 | 0.72 | 2.27 | 0.78 | 2.96 | 0.89 | **3.34** | **0.94** |
| | | -10dB | 2.37 | 0.71 | 2.60 | 0.81 | 1.88 | 0.84 | 3.22 | 0.92 | **3.49** | **0.95** |
| Babble | 10dB | -20dB | 1.36 | 0.47 | 1.59 | 0.59 | 1.89 | 0.66 | 2.45 | 0.79 | **2.45** | **0.82** |
| | | -15dB | 1.54 | 0.59 | 1.97 | 0.71 | 2.24 | 0.76 | 2.82 | 0.87 | **2.82** | **0.88** |
| | | -10dB | 1.87 | 0.70 | 2.33 | 0.80 | 2.58 | 0.84 | 3.12 | 0.91 | **3.19** | **0.93** |
| | 20dB | -20dB | 1.32 | 0.48 | 1.63 | 0.60 | 1.93 | 0.67 | 2.52 | 0.81 | **2.78** | **0.87** |
| | | -15dB | 1.58 | 0.60 | 2.04 | 0.72 | 2.31 | 0.78 | 2.91 | 0.88 | **3.15** | **0.92** |
| | | -10dB | 1.89 | 0.71 | 2.36 | 0.81 | 2.60 | 0.84 | 3.17 | 0.92 | **3.46** | **0.95** |
| White | 10dB | -20dB | 1.32 | 0.47 | 1.58 | 0.58 | 1.88 | 0.65 | 2.48 | 0.79 | **2.66** | **0.85** |
| | | -15dB | 1.52 | 0.59 | 1.98 | 0.71 | 2.26 | 0.76 | 2.86 | 0.87 | **3.04** | **0.91** |
| | | -10dB | 1.86 | 0.71 | 2.35 | 0.80 | 2.59 | 0.84 | 3.13 | 0.91 | **3.36** | **0.94** |
| | 20dB | -20dB | 1.29 | 0.48 | 1.63 | 0.60 | 1.93 | 0.67 | 2.55 | 0.81 | **2.86** | **0.88** |
| | | -15dB | 1.55 | 0.60 | 2.02 | 0.72 | 2.29 | 0.78 | 2.91 | 0.88 | **3.23** | **0.93** |
| | | -10dB | 1.90 | 0.72 | 2.39 | 0.81 | 2.63 | 0.85 | 3.20 | 0.92 | **3.53** | **0.96** |
| Average | | | 1.62 | 0.59 | 2.01 | 0.71 | 2.21 | 0.76 | 2.86 | 0.87 | **3.10** | **0.91** |

**Table 2**: Average PESQ and STOI scores of baselines and proposed model in various noisy and acoustic path situation.

speech quality/intelligibility in noisy environments, and (III) a comparison of computational complexity with respect to the baselines.

#### *4.4.1. Comparison of multiple inputs conditions*

We first studied the impact of a suitable combination of multiple inputs on the performance of the proposed model. The results presented in Table 1 show that UCME-3R outperforms UCME-3M in terms of both PESQ and STOI scores; this implies that the reference signal is an important factor for RES. Moreover, UCME-4 slightly outperformed UCME-3R, which implies that all signals including the microphone and reference signals were effective in improving the performance of RES.

#### *4.4.2. Comparison of performances in various noisy and acoustic echo path situation*

Subsequently, we examined the robustness of the proposed model in an environment involving noise and reverberation. The average PESQ and STOI scores of the AEC system, equipped with the RES as DNN, BLSTM, Convtasnet, and proposed methods in -20, -15, -10 SER with 10, 20 dB SNR level are shown in Table 2. We observe that the proposed UCME-4 RES model yields the best results in terms of the average PESQ and STOI scores. For the proposed model, the average PESQ and STOI scores under all conditions were 3.1 and 0.91, whereas those for the previous best RES model were 2.86 and 0.87, respectively.

#### *4.4.3. Comparison of computational complexity*

| **Model** | **RTF** | **GFLOPs** | **Model Size (MB)** |
|---|---|---|---|
| Tasnet-MI [7] | 1.34 | 7.82 | 20 |
| UCME-4 (Proposed) | 0.53 | 3.99 | 11 |

**Table 3**: Comparison of computational complexity.

In Table 3, We further investigated the RTF, GFLOPs, and model size for the Tasnet-MI and the proposed model. The floating-point operation and model size of the proposed model require less than about half those of baseline. The proposed model has a reasonable 0.53 RTF value for real-time processing. We conclude that the proposed model has not only a competitive PESQ and STOI performance, but also competitive performance in terms of computational complexity.

## 5. CONCLUSION

In this paper, we proposed a novel end-to-end neural network that employs multiple encoders to appropriately incorporate the information among various signals in a latent space, while utilizing efficient U-Convblock as mask estimation network. The proposed system outperforms baselines in environment involving noise and reverberation while requiring relatively low computational resources. In future research, we plan to extend this method to AES scenario in actual recordings over smart audio devices.

## 6. REFERENCES


[1] J. Benesty, T. Gänsler, D. Morgan, S. Gay, and M. Sondhi, *Advances in Network and Acoustic Echo Cancellation*. Springer Science & Business Media, 2001.

[2] S. Y. Lee and N. S. Kim, "A statistical model-based residual echo suppression," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 758–761, 2007.

[3] H. Song and J. W. Shin, "Residual echo suppression considering harmonic distortion and temporal correlation," *Applied Sciences*, vol. 10, no. 15, p. 5291, 2020.

[4] C. M. Lee, J. W. Shin, and N. S. Kim, "Dnn-based residual echo suppression," in *INTERSPEECH*, 2015.

[5] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *INTERSPEECH*, 2018.

[6] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *INTERSPEECH*, 2019.

[7] H. Chen, T. Xiang, K. Chen, and J. Lu, "Nonlinear residual echo suppression based on multi-stream conv-tasnet," in *INTERSPEECH*, 2020.

[8] 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.

[9] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1251–1258, 2017.

[10] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.

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

[12] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[13] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying tasnet: A dissecting approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[14] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[15] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[16] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[17] 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, 1993.

[18] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[19] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-Garcia, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

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

[21] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *arXiv preprint arXiv:1810.11359*, 2018.

[22] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] J.-M. Valin, "Speex: A free codec for free speech," *arXiv preprint arXiv:1602.08668*, 2016.