Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 8, No. 1, March 2020, pp. 45~51
ISSN: 2089-3272, DOI: 10.11591/ijeei.v8i1.1070

Double-talk robust acoustic echo canceller based on CNN filter

Haengwoo Lee
Namseoul University/91, Daehak-ro, Cheonan-si, Chung-nam, South Korea

Article Info

Article history:

Received Feb 25, 2019
Revised Nov 11, 2019
Accepted Jan 31, 2020

Keyword:

Neural network
Acoustic echo canceller
Double-talk
Deep learning

ABSTRACT

Conventional acoustic echo cancellation works by using an adaptive algorithm to identify the impulse response of the echo path. In this paper, we use the CNN(convolutional neural network) filter to remove the echo signal from the microphone input signal, so that only the speech signal is transmitted to the far-end. Using the neural network filter, weights are well converged by the general speech signal. Especially, it shows the ability to perform stable operation without divergence even in the double-talk state, in which both parties speak simultaneously. As a result of simulation, this system showed better performance and stable operation compared to the echo canceler of the adaptive filter structure. And, in double-talk, we showed the ERLE in the CNN is about 3 [dB] better than in the general neural network.

Corresponding Author:

Haengwoo Lee,
Department of Electrical and Computer Engineering,
Namseoul University,
91, Daehak-ro, Cheonan-si, Chung-nam, South Korea.
Email: haengwoolee@hanmail.net

1. INTRODUCTION

Acoustic echo is a problem when loudspeaker and near-end signals are combined at the microphone and sent to the far end. The acoustic echo signal disturbs receiving the near-end speeches in the far-end by which the received signals from the far-end in the near-end are emitted through the speaker and then combined with the near-end speeches in the microphone. If this is not properly handled, the far-end speaker will hear his / her voice delayed by the round trip time of the communication system, which is a very unpleasant problem in the hands-free calling. Generally, the elimination of the echo signal is achieved by adaptively converging the acoustic impulse response between the loudspeaker and the microphone using a FIR(finite impulse response) filter[1]. However, this operates normally only in a one-way conversation in which only a far-end signal exists, and in a double-talk interval in which the near-end speeches are also present, the ability to cancel an echo signal suddenly deteriorates. At this time, the presence of the near-end speech signal seriously degrades the convergence of the adaptive algorithm and may even be a factor for the filter coefficient to diverge. Therefore, in order to solve the double-talk problem, there is a method for preventing divergence of coefficients by detecting the double-talk state and stopping the update of the coefficient of the echo canceller[2][3]. However, this method has a relatively long detection time, so that the coefficients of the echo canceller may diverge before which a double-talk is detected. Also, the signal received at the microphone may introduce non-linear distortion into the echo signal due to limitations of components such as power amplifiers and loudspeakers as well as echo and near-end speech. To overcome this problem, a neural network structure that can model complex nonlinear relationships can be a powerful alternative.

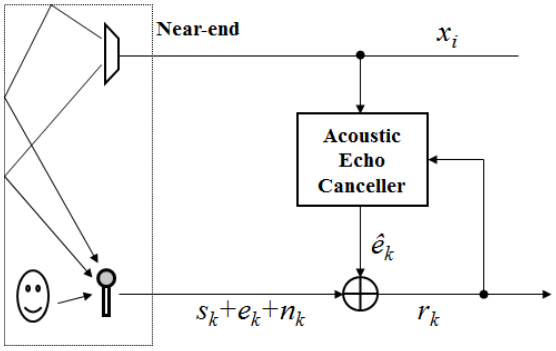

Figure 1. Acoustic echo cancellation system

The final goal of the acoustic echo canceller is to completely remove the echo signal and background noise, and transmit only the near-end speech to the far-end. From the point of view of speech separation, the near-end speech is separated from the microphone signal and transmitted to the far-end, which can naturally be treated as a speech separation problem. Therefore, instead of estimating the acoustic echo path, we apply speech separation learning to separate the near-end speech from the microphone signal together with the far-end speech, which is the raw signal of the echo accessible as additional information. In this approach, the double-talk problem is solved without using a specific sensing circuit. Deep learning shows great potential for speech separation. Experimental results show that the proposed method can remove the acoustic echo well in the noisy double-talk situation.

Computers that have been performing human commands and corresponding simple behaviors have transformed the paradigm of architecture by learning and reinforcing new connections. We experienced this change in 2016, through an artificial intelligence program called AlphaGo. The actual machine learning study began with the development of a multi-layer perceptron learning algorithm[4] in 1988. Through the 1990s, Decision trees, Bayesian networks, Support vector machines in the internet business are started to apply to Information search, Data mining, E-commerce, and recommendation services. In 2006, Geoffrey Hinton announced a way to improve existing neural network technology by adding more layers to the network[5]. As a result, machine learning has evolved into deep learning, and recently, deep learning technology has attracted attention as it shows performance exceeding human ability. In the late 2000s, machine learning contributed greatly to the advancement of the artificial intelligence industry, including the use of Apple's Siri, IBM's Watson, and Google's automatic speech recognizer. Recently, the deep-running model has achieved great results because it has developed a technique for learning multi-layer neural network composed of many layers. An error back-propagation algorithm that learns a multi-layer network can learn a deep network using a large number of layers by learning the lower layer synapses before learning the upper layer. The most commonly used deep-running model is the CNN[6][7][8].

When the near-end and far-end speakers talk simultaneously in an acoustic echo cancellation system, this condition is called double-talk. Without a double-talk detector that works correctly in case of double-talk, the echo canceller cannot reliably cancel echoes and distort speech sent to the far end. To detect the double-talk state, we used cross-correlation energy of the far-end and microphone signals[9] or the variance of the maximum value of the tap in the adaptive filter[10] or zero-crossing rate of error[11]. However, because the detection takes a long time, the coefficients of the filter may diverge. Therefore, in this study, by using neural network technology which is inherently tough to double-talk[12][13][14][15], there is no the need for a double-talk detector. Neural network based acoustic echo cancellers can operate reliably without diverging of weights, even under double-talk conditions.

In this paper, we discuss the learning algorithm of the multi-layer neural network in the following section 2. In section 3, we present the structure of a CNN neural network filter that is suitable for acoustic echo cancellation. Section 4 gives a discussion of the experimental results and analyses. Finally, Conclusion is made in section 5.

2. Learning Algorithm of Multi-Layer Neural Network

A multi-layer perceptron has the structure of a multi-layer forward neural network with one or more hidden layers. Figure 2 shows a multi-layer perceptron consisting of an input layer with l input neurons, a hidden layer with m hidden neurons, and an output layer with n output neurons. The values of the input neurons of the multi-layer perceptron are represented by the l-dimension vector $x = [x_1, x_2, \cdots, x_i, \cdots, x_l]$,

the values of the hidden neurons are represented by the m-dimension vector $a = [a_1, a_2, \cdots, a_j, \cdots, a_m]$, and the values of the output neurons are represented by the n-dimensional vector $a = [a_1, a_2, \cdots, a_k, \cdots, a_n]$. The weight and bias between the input layer and the hidden layer is represented by w_{ij}^1, b_j^1, the weight and bias between the hidden layer and the output layer is expressed by w_{jk}^2, b_k^2. Also, the weighted sum inputted to the j-th hidden neuron is called u_j^h, the weighted sum inputted to the k-th output neuron u_k^o, and the activation function of the hidden neuron is expressed as $\emptyset_h$, the activation function of the output neuron $\emptyset_o$.

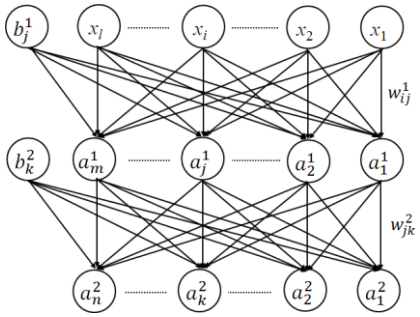

Figure 2. Multi-layer neural network

The output values of the hidden neurons and output neurons can then be expressed by the following equations.

$$a_j^1 = \emptyset(u_j^h) = \emptyset_h\left(\sum_{i=1}^{l} w_{ij}^1 \, x_i + b_j^1\right) \quad (1)$$

$$a_k^2 = \emptyset(u_k^o) = \emptyset_o\left(\sum_{j=1}^{m} w_{jk}^2 \, a_j^1 + b_k^2\right) \quad (2)$$

If we denote all weights and biases as a parameter θ, we can express the value of the k-th output neuron as a function $f_k(x, \theta)$ given the input x.

$$f_k(x, \theta) = a_k^2 = \emptyset_o\left(\sum_{j=1}^{m} w_{jk}^2 \, \emptyset_h\left(\sum_{i=1}^{l} w_{ij}^1 \, x_i^1 + b_j^1\right) + b_k^2\right) \quad (3)$$

The back-propagation learning algorithm was developed by Geoffrey Hinton in the mid-1980s. The supervised learning of the multi-layer perceptron should be based on the target output value and the cost function using the difference of the values outputted by the multi-layer perceptron. When the learning data and the target output value are given as a pair of input and output orders $(x_i, t_i)(i = 1, \cdots, N)$, the error for the whole learning data X can be defined as a mean square error as shown in the following equation.

$$E(X, \theta) = \frac{1}{2N} \sum_{i=1}^{N} \|t_i - f(x_i, \theta)\|^2 \quad (4)$$

In the above equation, the error function $E(X, \theta)$ is set to one value given the data set X and the parameter θ. The data set X is the value given from the outside, and the target to be optimized is θ. Therefore it can be written $E(\theta)$. The back-propagation learning algorithm uses the gradient descent method to find the parameters to minimize the error function $E(\theta)$. The gradient descent method is an algorithm that finds a parameter that minimizes the value of a cost function iteratively.

$$\theta(t+1) = \theta(t) + \Delta\theta(t) = \theta(t) - \eta \frac{\partial E(\theta)}{\partial \theta} \quad (5)$$

Where η is the learning rate that controls the speed of learning. In the multi-layer perceptron, the back-propagation learning uses the error function $E(X, \theta)$ for a data by applying a stochastic gradient descent method of updating a data for each weight.

$$E(x, \theta) = \frac{1}{2}(t_k - a_k^2)^2 = \frac{1}{2}\left(t_k - \emptyset_o\left(\sum_{j=1}^{m} w_{jk}^2 \, a_j^1 + b_k^2\right)\right)^2 \quad (6)$$

The parameter θ, which should be corrected through learning in the above equation, is the weight w_{jk}^2 and bias b_k^2 between the hidden layer and the output layer, and the weight w_{ij}^1 and bias b_j^1 between the input layer and the hidden layer. Therefore, if the error function is partially differentiated by the output-side parameter, it is as follows.

$$\frac{\partial E}{\partial w_{jk}^2} = \frac{\partial E}{\partial u_k^o}\frac{\partial u_k^o}{\partial w_{jk}^2} = -\phi_o'(u_k^o)(t_k - a_k^2)a_j^1 = \delta_k a_j^1 \quad (7)$$

$$\frac{\partial E}{\partial b_k^2} = \frac{\partial E}{\partial u_k^o}\frac{\partial u_k^o}{\partial b_k^2} = -\phi_o'(u_k^o)(t_k - a_k^2) = \delta_k \quad (8)$$

Here, $\phi_o'(u_k^o)$ is the derivative value of the activation function of the output neuron, and is generally a unit step function $\phi_o'(u_k^o) = u(t)$ since the ReLU function ($\max\{0, u_k^o\}$) is widely used. And then δ_k is the effect of the output neuron on the error. Next, the error function is partially differentiated by the input-side parameters as follows.

$$\frac{\partial E}{\partial w_{ij}^1} = \frac{\partial E}{\partial u_j^h}\frac{\partial u_j^h}{\partial w_{ij}^1} = \phi_h'(u_j^h)\sum_{j=1}^m w_{jk}^2\,\delta_k x_i = \delta_j x_i \quad (9)$$

$$\frac{\partial E}{\partial b_j^1} = \frac{\partial E}{\partial u_j^h}\frac{\partial u_j^h}{\partial b_j^1} = \phi_h'(u_j^h)\sum_{j=1}^m w_{jk}^2\,\delta_k = \delta_j \quad (10)$$

Taken together, it can be seen that the parameters between the input layer and the hidden layer are affected by the sum of multiplication the weights between the hidden and output layers by the effect δ_k of each output neuron on the error. Since the error of the output neuron propagates backward to the hidden neuron and influences the parameter control of the hidden neuron, the gradient descent learning method of the multi-layer perceptron is named as the error back-propagation learning algorithm, and finally each parameter is updated by the following equation.

$$w_{jk}^2(t+1) = w_{jk}^2(t) + \eta\phi_o'(u_k^o)(t_k - a_k^2)a_j^1 \quad (11)$$

$$b_k^2(t+1) = b_k^2(t) + \eta\phi_o'(u_k^o)(t_k - a_k^2) \quad (12)$$

$$w_{ij}^1(t+1) = w_{ij}^1(t) - \eta\phi_h'(u_j^h)\sum_{j=1}^m w_{jk}^2\,\delta_k x_i \quad (13)$$

$$b_j^1(t+1) = b_j^1(t) - \eta\phi_h'(u_j^h)\sum_{j=1}^m w_{jk}^2\,\delta_k \quad (14)$$

3. Acoustic echo cancellation based on CNN filter

The CNN-based neural network filter not only expresses the features of the time domain of the signal but also the frequency domain. The CNN structure is composed of convolution kernels with various widths according to frequency bands, and it can read complex nonlinear characteristics depending on the number of hidden layers. In this paper, we use a network with a 2 hidden layer and a 5-band kernel. The width of the kernel is composed of 4, 8, 16, 32, 64 data considering that the speech signal has high energy in the low frequency band. Figure 3 shows an example with a kernel width of 4, and input data of adjacent neurons are superimposed as in a network with a different kernel width.

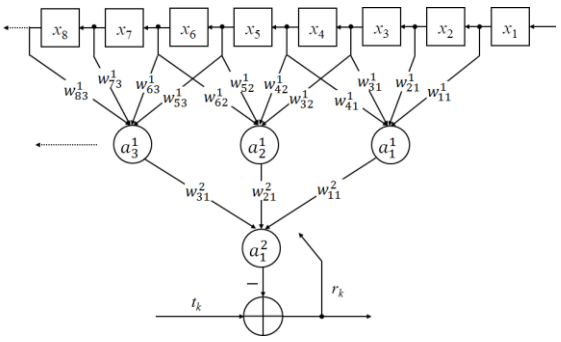

Figure 3. Acoustic echo canceller with CNN filter

For convenience, assuming that the bias is zero and using ReLU(Rectified Linear Unit), which is often used as an activation function, the output of each layer is obtained as follows.

$$a_j^1 = \phi_{ReLU}(u_j^h) = \begin{cases} \sum_{i=1}^{l} w_{ij}^1 x_i, & u_j^h > 0 \\ 0, & Otherwise \end{cases} \quad (15)$$

$$a_k^2 = \phi_{ReLU}(u_k^o) = \begin{cases} \sum_{j=1}^{m} w_{jk}^2 a_j^1, & u_k^o > 0 \\ 0, & Otherwise \end{cases} \quad (16)$$

The error value of the output neuron with respect to the target value is $r_k = t_k - a_k^2$ and the derivative value of the activation function is $\phi'(u) = 1\ (for\ u > 0)$. Using the NLMS (normalized least mean square) algorithm, the weights of each layer for $u > 0$ are updated as follows:

$$w_{jk}^2(t+1) = w_{jk}^2(t) + \eta r_k \frac{a_j^1}{E\{a_m^1\}} \quad (17)$$

$$w_{ij}^1(t+1) = w_{ij}^1(t) + \eta \sum_{j=1}^{m} w_{jk}^2 r_k \frac{x_i}{E\{x_l\}} \quad (18)$$

4. Simulation results

In order to evaluate the performance of the proposed acoustic echo canceller, simulations were performed using a Python program. The speech signal is sampled at 8 kHz and represented as 8 bits. The simulation room size is (3×3×2) m and the room impulse response has a reverberation time of 64 ms. Echo canceling performance uses ERLE(echo return loss enhancement) defined as follows.

$$ERLE[dB] = 10\log\left(\frac{E\{(x_i)^2\}}{E\{(r_k)^2\}}\right) \quad (19)$$

Here, $E\{\cdot\}$ represents a probabilistic expectation value, and x_i is the microphone input signal in which the echo signal is mixed with the near-end speaker voice and noise. Figure 4 shows the ERLE characteristics and residual error curves for two types of echo cancellers when the echo is generated by the white noise at the same time the near-end talker is speaking. The ERLE curve of the FIR filter structure, indicated by the dotted line (F) in the figure above, is strongly influenced by the speech of the near-end. On the other hand, the ERLE curve of the CNN filter structure indicated by the solid line (N) increases steadily, up to 25 [dB] regardless of the speech of the near-end. The lower figure shows the residual error of the CNN filter structure. It decreases continuously in a short time and converges to the minimum value.

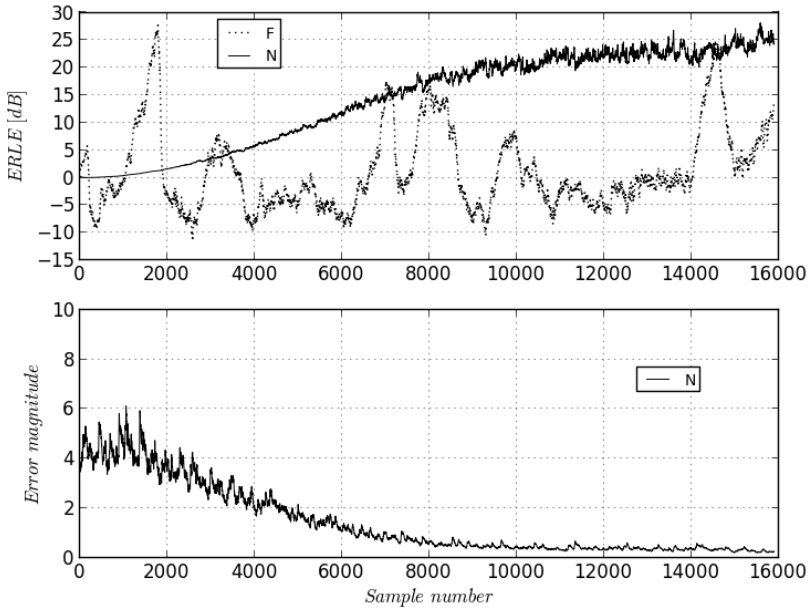

Figure 4. ERLE and error in double-talk with far-end white noise and near-end speech

The curve fo upper figure in Figure 5 shows that the ERLE curve oscillates largely due to the speech of the near-end talker when the echo is generated by the speech of the far-end talker in the FIR filter structure. Here, the light curve (S) represents the speech of the near-end talker and the thick curve (E) represents the ERLE value. And the curve of lower figure is the mean error value after echo cancellation.

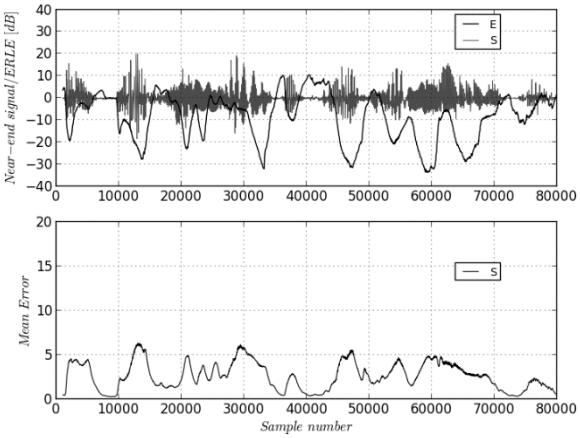

Figure 5. ERLE and mean error in double-talk with speeches in FIR filter

Figure 6 shows the characteristics of a CNN filter when a double-talk occurs. The curve of upper figure shows the mic-induced speech of the near-end talker, and the curve of middle figure shows the ERLE value, which is constantly increased to 15 [dB] while being interrupted by the speech of the near-end talker. And the curve of lower figure shows that the mean error value is kept very small regardless of the speech of the near-end talker.

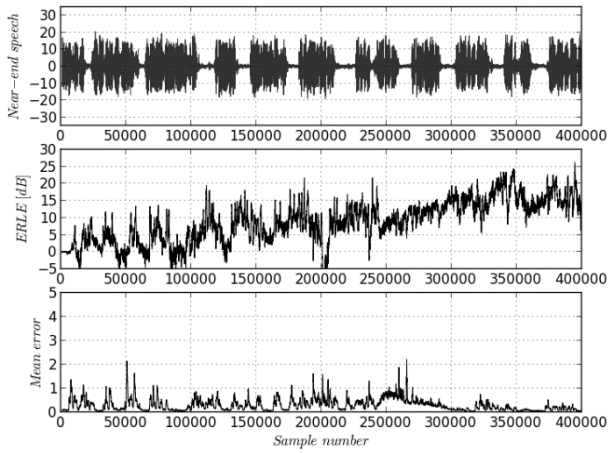

Figure 6. ERLE and mean error in double-talk with speeches in neural network filter

Figure 7 shows the Comparison of the ERLE and the mean error in double-talk in the general neural network and the CNN. The curves of upper figure show the ERLEs which have about 3 [dB] better in the CNN than the general neural network. The curves of lower figure show that the mean error value in the CNN gets slightly smaller than in the general neural network.

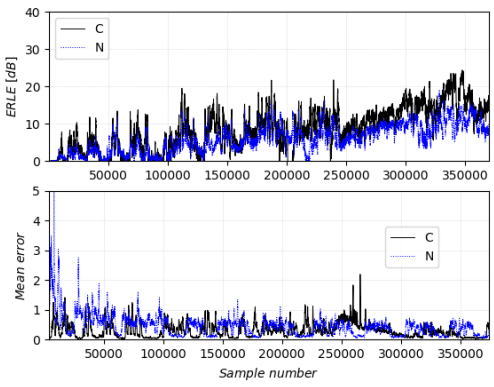

Figure 7. Comparison of ERLE and mean error in double-talk in general neural network and CNN

5. Conclusion

In this paper, we show that the CNN filter removes the echo signals well in the acoustic echo canceller despite the double-talk. Using the neural network filter, weights are well converged on the general speech signal. Especially, it shows the ability to perform stable operation without divergence even in double-talk state. Therefore, the acoustic echo canceller can always update the weights regardless of the double-talk.

To compare the performance, We simulated the ERLE in double-talk in the general neural network and the CNN. The results show the ERLEs in the CNN are about 3 [dB] better than in the general neural network.

ACKNOWLEDGEMENTS

Funding of this paper was provided by Namseoul University.

REFERENCES

[1] Wenbin Hsu, Frank Chui, and David A. Hodges, "An Acoustic Echo Canceler", IEEE J. of Solid-state Circuits, vol.24, no.6, pp.1639-1646, Dec. 1989.
[2] S. Minami, T. Kawasaki, "A Double Talk Detection Method for an Echo Canceller", ICC'85, pp.1492-1497, 1985.
[3] Hua Ye, Bo-Xiu Wu, "A New Double-Talk Detection Algorithm Based on the Orthogonality Theorem", IEEE Trans. on Comm., vol.39, no.11, pp.1542-1545, Nov. 1991.
[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Cognitive modeling, vol.5, p.3, 1988.
[5] G. Hinton, R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol.313, no.5786, pp.504–507, Jul. 2006.
[6] LISA Lab, "Convolutional Neural Networks - Deep Learning 0.1 documentation," 2013.
[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol.86, no.11, pp.2278-2324, Nov. 1998.
[8] Jürgen Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol.61, pp.85–117, 2015.
[9] Mahfoud Hamidia, Abderrahmane Amrouche, "Double-talk Detection Using Signal Energy for Acoustic Echo Cancellation," Workshop of the Speech Communication and Signal Processing Lab, Dec. 2016.
[10] Sonika, Sanjeev Dhull, "Double Talk Detection in Acoustic Echo Cancellation based on Variance Impulse Response," International Journal of Electronics and Communication Engineering, vol.4, no.5, pp.537-542, 2011.
[11] Muhammad Z. Ikram, "Double-talk detection in acoustic echo cancellers using zero-crossings rate," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, pp.1121-1125, Apr. 2015.
[12] Hao Zhang, DeLiang Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," Interspeech 2018, pp.3239-3243, Sep. 2018.
[13] Mehdi Bekrani, Andy W. H. Khong, and Mojtaba Lotfizad, "Neural Network Based Adaptive Echo Cancellation for Stereophonic Teleconferencing Application," IEEE International Conference on Multimedia and Expo(ICME) 2010, pp.1172-1177, 2010.
[14] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in Proc. Interspeech, vol.1, pp.1775-1779, Sep. 2015.
[15] Michael Muller, Jakub Jansky, Marek Bohac, and Zbynek Koldovsky, "Linear acoustic echo cancellation using deep neural networks and convex reconstruction of incomplete transfer function," IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics(ECMSM) 2017, pp.1-6, May 2017.