# DNN-based Residual Echo Suppression

*Chul Min Lee*[1], *Jong Won Shin*[2] *and Nam Soo Kim*[1]

[1]Department of Electrical and Computer Engineering and INMC
Seoul National University, Seoul, Korea
[2]School of Information and Communications,
Gwangju Institute of Science and Technology, Gwangju, Korea
cmlee@hi.snu.ac.kr, jwshin@gist.ac.kr, nkim@snu.ac.kr

## Abstract

Due to the limitations of power amplifiers or loudspeakers, the echo signals captured in the microphones are not in a linear relationship with the far-end signals even when the echo path is perfectly linear. The nonlinear components of the echo cannot be successfully removed by a linear acoustic echo canceller. Residual echo suppression (RES) is a technique to suppress the remained echo after acoustic echo suppression (AES). Conventional approaches compute RES gain using Wiener filter or spectral subtraction method based on the estimated statistics on related signals. In this paper, we propose a deep neural network (DNN)-based RES gain estimation based on both the far-end and the AES output signals in all frequency bins. A DNN architecture, which is suitable to model a complicated nonlinear mapping between high-dimensional vectors, is employed as a regression function from these signals to the optimal RES gain. The proposed method can suppress the residual components without any explicit double-talk detectors. The experimental results show that our proposed approach outperforms a conventional method in terms of the echo return loss enhancement (ERLE) for single-talk periods and the perceptual evaluation of speech quality (PESQ) score for double-talk periods.

**Index Terms**: acoustic echo suppression, residual echo suppression, nonlinear echo, deep neural networks, optimal gain regression

## 1. Introduction

Acoustic echo cancellation (AEC) or suppression (AES) is a technique to reduce the echo originated from acoustic coupling between loudspeakers and microphones [1, 2, 3, 4]. Although there have been many techniques which are prove to suppress the echo successfully, there still exists some amount of residual echo at the outputs of these methods. One of the reasons for which the AEC or AES suffer is that the echo signal is not a linear function of the far-end digital signal even when the echo path is perfectly linear. The power amplifiers and loudspeakers, especially cheap and small ones, can be the sources of this nonlinearity.

To overcome this problem, several residual echo suppression (RES) filters have been applied to the output of the AEC or AES to suppress remaining echo [5, 6, 7, 8]. The authors in [5] and [6] proposed RES methods to estimate the signal-to-echo ratio (SER) and then apply Wiener filters or spectral subtraction in the frequency domain. In [7], subband filtering based on the spectral subtraction was combined with a truncated Taylor series expansion of acoustic echo path for the estimation of power spectral density of the echo. In [8], a RES algorithm using a magnitude regression model of residual echo based on modeling inter-frequency dependencies of far-end and the echo signals was proposed. Recently, a RES in which the residual echo was estimated from the far-end signal using an artificial neural network (ANN) was proposed [9]. The inputs of the ANN were the far-end signal in the given frequency bin, the power of the signal and the sum of the frequency components that can bring about harmonic distortion, and the final spectral gain was the Wiener filtering gain. However, these methods did not consider the nonlinear characteristic between residual echo and far-end signals in the all frequency bins.

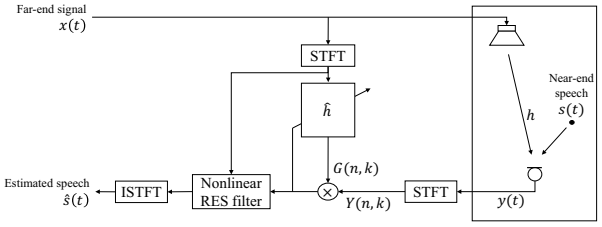


Figure 1: *Schematic diagram of AES system with nonlinear RES post-filter.*

In this paper, we proposed a residual echo suppression using deep neural networks (DNNs) which estimate the optimal RES gain based on both the far-end and the output signals of AES in all frequency bins. DNN structures can learn the complicated mapping among high-dimensional vectors and have been already successfully applied to automatic speech recognition and speech enhancement area [10, 11, 12, 13]. We expect that these architectures can accommodate to model a nonlinear regression function from these signals to optimal RES gain based on DNN training using multi-condition data even though the room impulse responses (RIRs) used in the training do not match the RIRs for the test. We evaluate overall performance using two objective measures in matched and mismatched conditions for various RIRs, SER, clipping type, and level of nonlinearity in loudspeaker. These metrics are echo return loss enhancement (ERLE) for single-talk periods and the ITU-T Recommendation P. 862 perceptual evaluation of speech quality (PESQ) [14] for double-talk periods. Experimental results showed that the proposed method demonstrated improved speech quality and echo suppression compared with a conventional algorithm with ANN-based residual echo estimation and Wiener filtering gain function [9].

## 2. Acoustic echo suppression system with nonlinear RES filter

Acoustic echo suppression (AES) [1, 2, 3, 4] provides an attractive alternative to acoustic echo cancellation (AEC) techniques for telecommunication in low-complexity systems to suppress the acoustic echo. A single-channel AES system is depicted in Figure 1. The far-end signal $x(t)$ at time index $t$ is generated by the source signal through the acoustic impulse response in the transmission room. Let $y(t)$ be the input signal including near-end speech $s(t)$ in the receiving room and $Y(n,k)$ is the short-time Fourier Transform (STFT) coefficient of $y(t)$ for $k$-th frequency bin at the $n$-th frame. The spectral gain function to suppress the echo, $G(n,k)$, is obtained from the Wiener filtering or spectral subtraction in each frequency bin. However, due to limitations of linear echo modeling, the echo component may still remain in the output of AES including a considerable amount of nonlinear echo degrading the quality of the near-end speech. To improve the output of AES, additional nonlinear residual echo suppression (RES) filter can be applied to the remaining signal. Using the residual echo suppression gain $G_{res}(n,k)$, the final estimated speech in the frequency domain, $\hat{S}(n,k)$ is given by,

$$\hat{S}(n,k) = \{G(n,k) \cdot G_{res}(n,k)\}Y(n,k). \quad (1)$$

When the power amplifiers and loudspeakers introduce severe nonlinearity, it is very important to calculate $\hat{G}_{res}(n,k)$ accurately in accordance with the nonlinearity of residual echo.

## 3. Residual echo suppression using deep neural networks

Various RES methods have been developed to suppress residual echo effectively [5, 6, 7, 8, 9]. However, these may not describe the exact nonlinear property of the residual echo signal due to the difficulty of constructing highly complex functions. Recently, in speech recognition and enhancement areas, deep neural network (DNN) structures have been employed as a powerful tool to find the complicated mapping or functions and shown better performance than other conventional methods [10, 11]. The main reason may be a breakthrough in DNN by using the stacked restricted Boltzmann machines (RBMs) accompanied with greedy layer-wise unsupervised learning to initialize the DNN parameters. After the unsupervised pre-training stage, a supervised learning algorithm is carried out to fine-tune the weights of the DNN using back propagation and stochastic gradient descent method. The detailed procedures about pre-training and fine-tuning processes are described in [12, 13]. In [9], an artificial neural network (ANN) was utilized to estimate the residual echo from the far-end signal, but the structure was not flexible enough because the input feature of the ANN was constructed by the knowledge of the harmonic distortion and the final gain function was the Wiener filter gain.

In this paper, we proposed an optimal gain regression based on DNN for RES. The DNN structure is employed to successfully represent a complex nonlinear regression function for optimal gain in the RES process. The gain $G_{res,opt}(n,k)$ is defined as follows,

$$G_{res,opt}(n,k) = \max\left\{\min\left(1, \frac{|S(n,k)|}{|Y_{AES}(n,k)|}\right), G_{min}\right\}, \quad (2)$$

where $S(n,k)$ and $Y_{AES}(n,k)$ are the STFT coefficients of clean near-end speech and the AES output signal and $G_{min} = 10^{-4}$ is introduced to reduce artifacts. As for the input, the far-end and the residual echo spectra are used. The relationship between the residual echo and the RES gain may be less dependent on the acoustic echo path than that between the input microphone signal and the gain. Therefore, the DNN may be able to identify the nonlinear relationship among the residual echo, far-end signal and RES gain through multi-condition training although the DB in the process is made by applying just a few echo paths. A DNN system for the proposed method is illustrated in Figure 2. This structure consists of a Gaussian-Bernoulli RBM and two Bernoulli-Bernoulli RBMs. The nodes of each hidden layer and the output layer in the DNN are modeled by the sigmoid function. The inputs of the DNN model are pairs of the residual echo and far-end signal represented by the magnitude spectra in the STFT domain. When taking $N$-point STFT, the dimension of the input feature vector considering $T$ consecutive frames of the residual echo and far-end signal is $(N/2+1) \times 2 \times T$, while the output of the DNN is a $(N/2+1)$-dimensional RES gain vector. These are normalized to have zero mean and unit variance. The phase of the estimated speech are kept the same as that of the AES output because the phase information is not crucial for human auditory system.

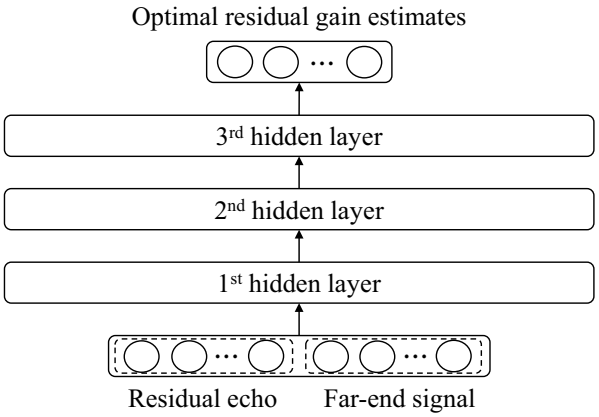


Figure 2: *A DNN structure for the proposed RES.*

In the DNN training, we firstly try to learn a deep generative model for spectra of the residual echo and far-end signal as a pre-training stage. The RBMs can be trained layer-by-layer in an unsupervised greedy fashion using contrastive divergence (CD). The parameters of each RBM are updated in this process. Afterwards, in the fine-tuning stage, back-propagation algorithm with the minimum mean squared error (MMSE) function between the estimated RES gain and the optimal gain, is employed to train to the DNN. The optimal gain for RES, $G_{res,opt}(n,k)$, can be computed using the AES output and near-end speech signals through equation (2). A stochastic gradient descent algorithm is performed in mini-batches to improve learning convergence as follows,

$$MMSE = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}(G_{res,opt}(m,k) - \hat{G}_{res}(m,k))^2 \quad (3)$$

where $M$ and $K$ is the mini-batch size and the total number of frequency bins, respectively. Then, the estimates of weights and bias can be updated iteratively. Some of the conventional methods were based on the independence assumption among each frequency bins or dependency of only a few of adjacent bins [5, 6, 7]. In contrast, the proposed work can take account

of nonlinear mapping between the optimal RES gain and features extracted from the output of AES and the far-end signal in a whole frequency range. In addition, the proposed method does not require any double-talk detectors as the training DB includes both near-end speech and echo signal. Thus, we believe that the proposed method can improve the echo estimation compared with other conventional methods.

## 4. Experimental results

To assess the performance of the proposed DNN-based RES, we conducted several simulations under various conditions. From the TIMIT database, we created 450 files (4036 s) of microphone signals for each room impulse response (RIR) from a location of a loudspeaker to the microphone illustrated in Figure 3 to construct residual echo DB. These files were sampled at 16 kHz. To simulate the echo signal captured by the microphone after passing through a power amp, a loudspeaker and acoustic transmission in order, we subsequently performed three processes on the far-end signals: clipping, applying the simulation model of a nonlinear loudspeaker, and convolution with RIRs. Artificial clipping [15] is made by

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

or

$$x_{soft}(n) = \frac{x_{max} x(n)}{\sqrt[\rho]{|x_{max}|^{\rho} + |x(n)|^{\rho}}} \quad (5)$$

where $x_{hard}$ and $x_{soft}$ are the outputs of hard and soft clipping, respectively, and $x_{max}$ is the maximum value of the output signal. For soft clipping, the value of $\rho$ was set to 2. To mimic the nonlinear loudspeaker characteristic, the memoryless sigmoidal function [16] was applied as follows:

$$x_{NL}(n) = \gamma\Big(\frac{2}{1 + \exp\left(-a \cdot b(n)\right)} - 1\Big) \quad (6)$$

where

$$b(n) = 1.5 \times x(n) - 0.3 \times x(n)^2. \quad (7)$$

The parameter $\gamma$ is the sigmoid gain which was set to $\gamma = 4$. The sigmoid slope value $a$ was chosen as $a = 4$ if $b(n) > 0$ and $a = 1/2$ otherwise. A receiving room was designed as a small office room with dimensions $4\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$. By using image method [17], the RIRs from 7 loudspeaker locations to the microphone in the receiving room depicted in Figure 3 were generated with reverberation time $T_{60} = 200\,\mathrm{ms}$. The length of the RIRs was set to 512. The echo level measured at the microphone was on average 3.5 dB lower than that of the near-end speech. For performance evaluation, echo return loss enhancement (ERLE) and perceptual evaluation of speech quality (PESQ) [14] were used as objective measures. The ERLE measure is defined by

$$\mathrm{ERLE}(t) = 10 \log_{10} \left[ \frac{E[y^2(t)]}{E[\hat{s}^2(t)]} \right] (\mathrm{dB}). \quad (8)$$

First, we applied the conventional acoustic echo suppression (AES) technique [4] on the whole data set. The AES in [4] was slightly modified so that it becomes a single channel AES by eliminating the second-channel echo estimation. The parameters for the AES were set to the values shown in [4]. Although the AES proposed in [4] was shown to reduce the linear echo effectively, the average ERLE for the tested data was approximately 9 dB due to the severe nonlinear distortion.

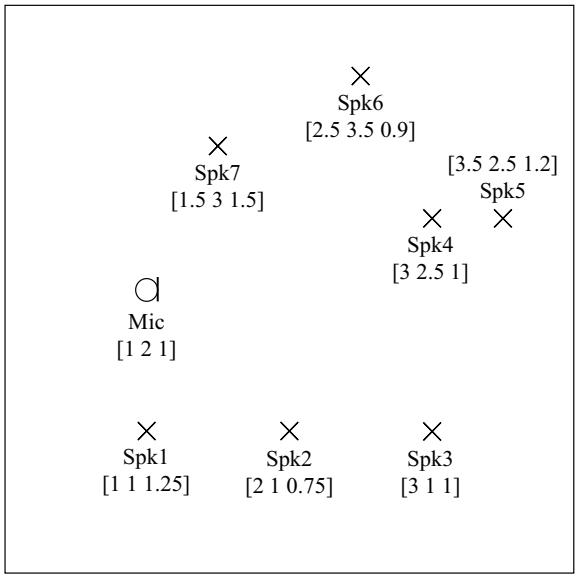


Figure 3: *Locations of one microphone and 7 loudspeakers in a simulated receiving room of* $4\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$ *for constructing echo DB.*

To compare the proposed method with a conventional RES technique, we implemented the ANN-based RES using spectral features [9]. The uniform 128-point STFT analysis-synthesis filter bank was used with 75 % overlap. The offline estimator for the RES is a network with two log-sigmoid hidden nodes. The magnitude spectra of far-end signal and the average over all subbands up to the half of the current band were used as the inputs. The training was performed on 30 files (267 s) of the residual echo applying RIRs from the locations of Spk1, Spk2, and Spk3 to that of the Mic in Figure 3. The parameters were set as follows: $\lambda = 0.95$ and $\mu = 5.0$. For double-talk detection, we applied the manually marked information on this method. We have also tried training larger DB or taking 256-point STFT, but neither of them could bring about performance improvement.

For the training of the proposed technique, the total 1200 files (10774 s) established in the locations of Spk1, Spk2, and Spk3 were used to train a DNN. The frame length was set to 256 samples with 50 % overlap. A 256-point STFT was applied to each frame. Each hidden and the output layer had 2048 and 129 nodes, respectively. The final input vector consisted of the current frame and the previous two frames and therefore becomes a 774-dimensional vector. The number of epoch for each layer of RBM pre-training was 20. Learning rate of the pre-training was 0.0005. In the fine-tuning, learning rate was set to 0.1 for the first 10 epochs, then decreased by 10 % after each epoch. Total iteration number was 50 and the mini-batch size $M$ was set to 256. For the tests at each location, we used two sets of 50 files (445s) for the single-talk and double-talk tests, respectively. The near-end speech was also chosen from the TIMIT database.

In Table 1, the overall results of the ERLEs for single-talk periods and PESQ scores for double-talk periods are shown, where the test data are obtained in all 7 positions of the loud-

Table 1: ERLE and PESQ Score obtained in the matched and mismatched conditions for various RIRs.

| RES type | | | None | ANN[9] | Proposed (DNN) |
|---|---|---|---|---|---|
| ERLE | Matched | Spk1 | 9.12 | 21.65 | 38.07 |
| | | Spk2 | 9.80 | 22.68 | 38.05 |
| | | Spk3 | 9.09 | 21.73 | 36.52 |
| | Mismatched | Spk4 | 9.65 | 20.68 | 27.98 |
| | | Spk5 | 8.75 | 23.09 | 30.82 |
| | | Spk6 | 10.00 | 21.69 | 26.37 |
| | | Spk7 | 9.83 | 21.87 | 23.96 |
| PESQ | Matched | Spk1 | 2.62 | 2.71 | 3.01 |
| | | Spk2 | 2.68 | 2.72 | 3.05 |
| | | Spk3 | 2.69 | 2.74 | 3.05 |
| | Mismatched | Spk4 | 2.68 | 2.75 | 3.01 |
| | | Spk5 | 2.61 | 2.67 | 2.93 |
| | | Spk6 | 2.71 | 2.74 | 3.00 |
| | | Spk7 | 2.73 | 2.77 | 3.00 |

Table 2: ERLE and PESQ Score in the mismatched conditions for effects of other factors at the location of Spk4.

| RES type | | None | ANN [9] | Proposed (DNN) |
|---|---|---|---|---|
| ERLE | SER 0 dB | 9.65 | 21.87 | 26.24 |
| | HC (70 %) | 9.64 | 21.80 | 26.20 |
| | SC (80 %) | 9.59 | 21.47 | 26.71 |
| | SC (70 %) | 9.56 | 21.56 | 26.33 |
| PESQ | SER 0 dB | 2.46 | 2.54 | 2.77 |
| | HC (70 %) | 2.66 | 2.74 | 2.99 |
| | SC (80 %) | 2.58 | 2.66 | 2.90 |
| | SC (70 %) | 2.55 | 2.63 | 2.88 |

speaker by using hard clipping at the 80 % of the maximum volume of the input signal. From the whole results, the proposed method based on DNN showed better performance than the conventional RES in both the matched and mismatched conditions. Especially, it could be seen that the proposed RES preserved the near-end speech much better as seen from the comparison of the PESQ scores. These results were obtained by training with only a few of the RIR cases, which may support our assumption that the mapping from the far-end signal and the residual echo to the RES gain is not substantially affected by the acoustic environment.

To investigate the effects of other factors such as signal-to-echo ratio, clipping types and amount of nonlinearity of loudspeakers on the RES algorithms, we additionally tested several cases corresponding to other mismatched conditions at the location of Spk4. For this test, we applied the same models used in the previous test, which were trained with DB at Spk1, Spk2, and Spk3 positions with 80 % hard clipping on each method. The performance of the proposed RES was compared to that of the conventional RES in Table 2. Signal-to-echo ratio (SER) 0 dB means that the near-end speech to echo ratio level was on average 0 dB. HC ($l$ %) and SC ($l$ %) indicate the hard and soft clipping with $l$ % of the maximum amplitude of the input signal, respectively. Comparing the output of our method to unprocessed signal, at least 0.3 point improvement of the PESQ score was found. In all 4 cases, the proposed method outperformed the conventional RES and was not affected by the various mismatched factors.

An example of $ERLE$ variation over time is given in conjunction with the corresponding unprocessed echo waveform in Figure 4. The proposed algorithm attenuated the residual echo components more efficiently than the conventional RES.

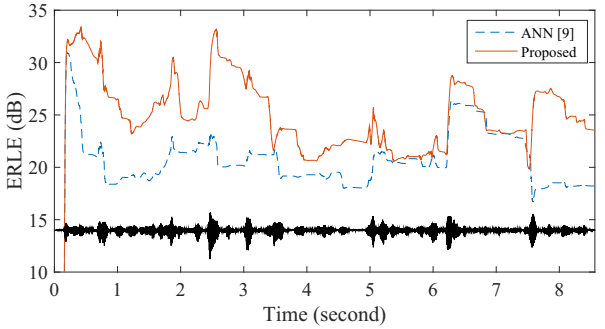


Figure 4: *Comparison of ERLE at the location of Spk4 in a single-talk situation.*

## 5. Conclusions

In this paper, we have proposed an optimal gain regression employing DNN for nonlinear residual echo suppression in the STFT domain. It is shown that the DNN-based regression can represent complex mapping among the optimal gain, residual echo, and far-end signal in the whole frequency bins. Moreover, the proposed method can suppress the residual components without any explicit double-talk detectors. The proposed RES outperformed the conventional one in terms of ERLE for single-talk situations and PESQ score for double-talk situations.

## 6. Acknowledgements

This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP (Institute for Information & communications Technology Promotion).

## 7. References

[1] C. Avendano, "Acoustic echo suppression in the STFT domain," *in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 175–178.

[2] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May. 2006, vol. 5, pp. 281-284.

[3] Y. S. Park and J. H. Chang, "Frequency domain acoustic echo suppression based on soft decision," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 53-56, Jan. 2009.

[4] C. M. Lee, J. W. Shin, and N. S. Kim, "Stereophonic acoustic echo suppression incorporating spectro-temporal correlations," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 316–320, Mar. 2014.

[5] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction", *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp.307–310.

[6] S. Y. Lee and N. S. Kim, "A statistical model based residual echo suppression," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 758-761, Oct. 2007.

[7] F. Kuech and W. Kellermann, "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, pp. 73–76.

[8] D. Bendersky, J. Stokes, and H. Malvar, "Nonlinear residual acoustic echo suppression for high levels of harmonic distortion," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 261–264.

[9] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," *in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2013

[10] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14-22, Jan. 2012.

[11] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, Vol. 21, No. 1, pp. 65–68, Jan. 2014.

[12] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, Jul. 2006.

[13] R. Salakhutdinov and G. Hinton, "Using deep belief nets to learn covariance kernels for Gaussian processes," *in Proc. Advances in Neural Inform. Process. Syst.*, 2007, vol. 20, pp. 1-8.

[14] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Rec. P. 862, 2000.

[15] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2065-2079, Sep. 2012

[16] D. Comminiello et al., "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1502-1512, Jul. 2013.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943-950, Apr. 1979.