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

# A Robust and Cascaded Acoustic Echo Cancellation Based on Deep Learning

*Chenggang Zhang, Xueliang Zhang*

Department of Computer Science, Inner Mongolia University, China
21809006@mail.imu.edu.cn, cszxl@imu.edu.cn

## Abstract

Acoustic echo cancellation (AEC) is used to cancel feedback between a loudspeaker and a microphone. Ideally, AEC is a linear problem and can be solved by adaptive filtering. However, in practice, two important problems severely affect the performance of AEC, i.e. 1) double-talk problem and 2) nonlinear distortion mainly caused by loudspeakers and/or power amplifiers. Considering these two problems in AEC, we propose a novel cascaded AEC which integrates adaptive filtering and deep learning. Specifically, two long short-term memory networks (LSTM) are employed for double-talk detection (DTD) and nonlinearity modeling, respectively. The adaptive filtering is employed to remove the linear part of echo. Experimental results show that the proposed method outperforms conventional methods in terms of the objective evaluation metrics by a considerable margin in the matched scenario. Moreover, the proposed method has much better generalization ability in the unmatched scenarios, compared with end-to-end deep learning method.

**Index Terms**: Acoustic echo cancellation, double-talk detection, deep learning, long short-term memory

## 1. Introduction

Acoustic echo widely exists due to coupling of the loudspeaker and the microphone in the process of communication with full-duplex hands-free devices, such as mobile telephony and teleconferencing system[1, 2, 3]. The microphone of these devices which captures signals coming from its own loudspeaker can produce uncomfortable echoes that seriously disturb the normal communication. So an important issue that has to be addressed is the acoustic echo cancellation (AEC). Ideally, AEC can completely remove acoustic echoes and transmit only the near-end speech to the far-end. However, one of the major challenges of AEC is to make it generalize well under such conditions as double-talk, background noise and nonlinear distortion. This study focuses on the generalization ability of AEC algorithm in different scenarios, especially in low signal-to-echo ratio (SER) conditions.

Although traditional AEC methods have been proposed to deal with double-talk and noise in the past decades, most of those methods are based either on correlation between signals, or on statistical properties of speech and noise [4, 5, 6]. They often fail to track non-stationary distortion in unexpected acoustic conditions, so the performance is severely affected by related signal characteristics.

In the recent years, deep learning has achieved remarkable results in the fields of speech recognition, and speech separation etc.[7, 8, 9]. More recently, Zhang and Wang [10] formulated AEC as a supervised speech separation problem, in which echo is considered as a special interfering noise. And they employed a end-to-end deep learning structure to deal with the problem. Lately, Zhang *et al.* [11] further developed a deep learning algorithm which considered the impacts of nonlinear distortions and additive noise. For learning-based algorithm [10, 11, 12, 13, 14], the performance often drops in the unmatched conditions (unseen samples in training stage) that is called generalization problem. This problem is even serious for AEC, because many factors can cause unmatched scenarios, e.g. microphone, loudspeaker, environment noise and far-end signals [15, 16, 17]. To improve the generalization, a direct way is to collect as much training data as possible. However, it pays huge cost.

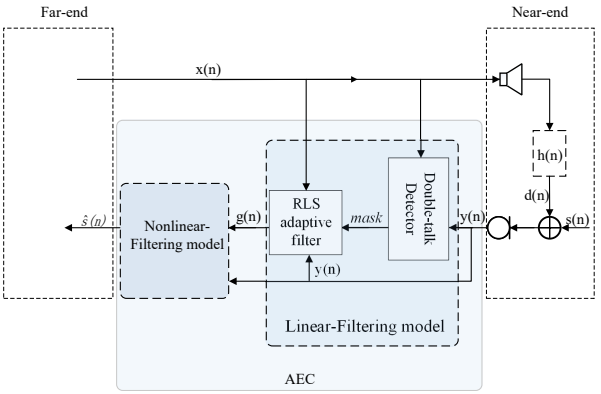


Figure 1: *Block diagram of the proposed method in acoustic echo scenario.*

In this paper, we propose a cascaded algorithm which combines conventional adaptive filtering with deep learning. The proposed algorithm consists of a linear-filtering model (LFM) and a nonlinear-filtering model (NLM). In LFM, a LSTM is employed as double-talk detector (DTD) to improve the performance of adaptive filtering. With the output of the LFM, another LSTM is trained to suppress the residual echo in the output of the LFM. Experimental results show that the proposed method outperforms the traditional methods in terms of the objective evaluation metrics in the matched scenario. Moreover, we also find that the proposed method has good generalization ability in the unmatched scenarios.

The rest of this paper is organized as follows. In Section 2, we introduce the AEC system and present the proposed method. The experimental setups are presented in Section 3. Experimental results and discussion are given in Section 4. Finally, Section 5 concludes the paper.

## 2. Algorithm description

### 2.1. System overview

The single-channel AEC method we proposed is depicted in Figure 1. The microphone received signal $y(n)$ consists of near-end speech signal $s(n)$ and echo signal $d(n)$ which is generated



http://dx.doi.org/10.21437/Interspeech.2020-1260

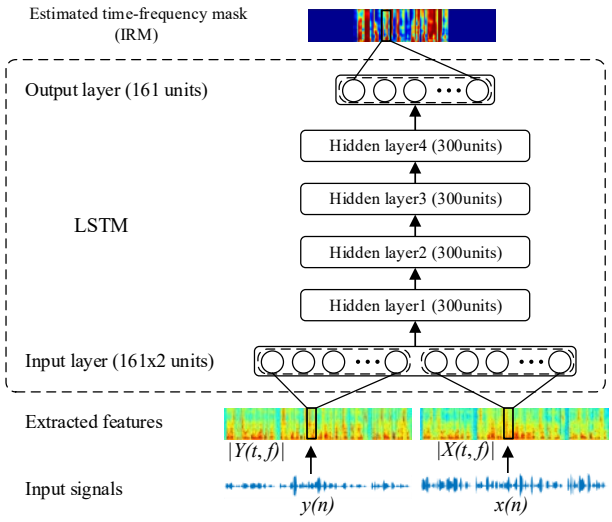


Figure 2: *Network architecture of LSTM for time-frequency mask estimation.*

by convolving a far-end signal $x(n)$ with a room impulse response (RIR) [18]:

$$d(n) = x(n) * h(n) \tag{1}$$

where * denotes the convolution operation, and $h(n)$ is the transfer function of actual echo transmission path. So, the $y(n)$ is obtained by:

$$y(n) = d(n) + s(n) \tag{2}$$

The goal of AEC is to obtain $s(n)$ by estimating $h(n)$ using $y(n)$ and $x(n)$. From Eq. (1) and (2), if there is no near-end signal, $h(n)$ is quite easy to estimate by using adaptive filtering algorithms, e.g. least mean square (LMS), normalized least mean square (NLMS) and recursive least square (RLS) [19, 20].

### 2.2. Linear-Filtering model (LFM)

However, when near-end signal and echo appear simultaneously, estimating $h(n)$ becomes complicated. This is called double-talk problem. A common strategy is to stop updating $h(n)$ when double talk happens. So, the accuracy of DTD has huge impact on performance and convergence speed of AEC. In this subsection, we introduce the approach for linear part of AEC. The LFM consists of deep learning-based DTD and RLS adaptive filtering.

#### *2.2.1. Double-talk detection*

The most effective way is to detect the double talk in time-frequency unit level considering both performance and convergence speed. So, we employ deep neural network to estimate the time-frequency mask which is widely used in speech enhancement at present [21, 22]. The training target is defined by Eq. (3) :

$$IRM(t,f) = \sqrt{\frac{|D(t,f)|^2}{|S(t,f)|^2 + |D(t,f)|^2}} \tag{3}$$

where $|S(t,f)|$ and $|D(t,f)|$ denote the time-frequency (T-F) unit of magnitude spectra at time $t$ and frequency $f$ of $s(n)$ and $d(n)$, respectively.

All input signals are sampled to 16 kHz, and then divided into frames with 20 $ms$ window length and 10 $ms$ offset, and *Hanning* window is used. We apply the short-time Fourier transformation (STFT) magnitude spectrum, only the first 161 frequency bins are used. In fact, IRM can be viewed as the probability of echo appearing at T-F units. If IRM is close to 1, it means no near-end signal showing up. Otherwise, it means that double talk happens.

To estimate the IRM, we use a recurrent neural network with four LSTM layers with 300 units in each layer, which is shown in Figure 2. A fully connected layer used for feature extraction is taken as an input layer. The magnitude spectra of $y(n)$ and $x(n)$ are concatenated as the input features which the dimension is 161 × 2 = 322, and then fed into LSTM. We use *sigmoid* activation function in the output layer which is fully connected, and its dimension is 161, corresponding to a frame of the estimated mask. Adam optimizer [23] is used to update the weights of LSTM, and the mean squared error (MSE) is used as the loss function. The learning rate, number of training epochs and batch size are set to 0.0003, 50 and 32, respectively.

#### *2.2.2. Adaptive filtering*

RLS has an important feature that its convergence speed is much faster than that of the standard LMS filter [19, 20], a frequency-domain RLS adaptive filter with DTD is employed to remove the linear echo components in microphone signal. The process can be described as follows.

$X(t,f)$ and $Y(t,f)$ are the frequency-domain counterparts of $x(n)$ and $y(n)$ at time-frame $t$ and frequency bin $f$ respectively, and $n$ being the time index. The cost function is a sum of squared errors, as given below:

$$E(t,f) = \sum_{\nu=0}^{t-1} \beta^{\nu} \left| Y(t-\nu,f) - W^T(t-\nu,f) X(t-\nu,f) \right|^2 \tag{4}$$

where $^T$ is the transpose operations, $\beta$ is the forgetting factor, and $W$ is the weight matrix. The optimization is to find a $W$ to minimize the $E$ :

$$W(t,f) = H^{-1}(t,f) P(t,f) \tag{5}$$

where, $H$ and $P$ denotes the covariance matrix which are updated by Eq. (6) and (7) if $IRM(t,f) > LC$. $LC$ is local criterion.

$$H(t,f) = \beta H(t-1,f) + X^T(t,f) X(t,f) \tag{6}$$

$$P(t,f) = \beta P(t-1,f) + Y^T(t,f) X(t,f) \tag{7}$$

If $IRM(t,f) \leq LC$, the update formulas are Eq. (8) and (9) :

$$H(t,f) = H(t-1,f) \tag{8}$$

$$P(t,f) = P(t-1,f) \tag{9}$$

It means that we do not update the parameters when double talk happens. The estimated frequency-domain echo signal $\hat{D}(t,f)$ is obtained by:

$$\hat{D}(t,f) = (H^{-1}(t,f) P(t,f))^T X(t,f) \tag{10}$$

then, the estimated frequency-domain near-end signal $G(t,f)$ is given by:

$$G(t,f) = Y(t,f) - \hat{D}(t,f) \tag{11}$$

Accordingly, the LFM output time-domain signal $g(n)$ is synthesized from $G(t,f)$, using the inverse STFT (iSTFT) [24].

It should be noticed that Eq. (10) includes an operation of matrix inverse, which is time consuming particular for large matrix. In practice, matrix inverse can be avoided by recursive algorithm (the details of derivation shown in [20]).

### 2.3. Nonlinear-Filtering model (NFM)

Due to the nonlinearity of the speaker and/or amplifier, there still exists residual echoes after LFM. In common, post-processing module is required. In order to remove the residual echoes, we train another LSTM which has the same structure as the one used in double-talk detection except for inputs and training target. The inputs for the second LSTM are $|G(t,f)|$and $|Y(t,f)|$ which are the magnitude spectra of $g(n)$ and $y(n)$. The training target is phase sensitive mask (PSM) [25, 26], as given below:

$$\begin{aligned} PSM(t,f) &= \mathrm{Re}\left\{\frac{|S(t,f)|\,e^{j\theta_s}}{|G(t,f)|\,e^{j\theta g}}\right\} \\ &= \frac{|S(t,f)|}{|G(t,f)|}\cos(\theta_s-\theta_g) \end{aligned} \tag{12}$$

where $|S(t,f)|$and $|G(t,f)|$ denote magnitude spectra of $s(n)$ and $g(n)$, $\theta_s$ and $\theta_g$ denote the *phases* in the T-F unit, respectively. $Re\{\cdot\}$ computes the real component. In the test stage, the estimated magnitude spectrum of near-end signal $|\hat{S}(t,f)|$ is obtained by:

$$|\hat{S}(t,f)| = PSM(t,f)|G(t,f)| \tag{13}$$

Finally, the estimated time-domain near-end speech signal $\hat{s}(n)$ is re-synthesized from $|\hat{S}(t,f)|$ combined with the *phase* of $G(t,f)$, using the iSTFT.

## 3. Experimental setups

### 3.1. Evaluation metrics

We use two metrics to evaluate the AEC performance: the echo return loss enhancement (ERLE) [27] for single-talk periods and the perceptual evaluation of speech quality (PESQ) [28] for double-talk periods.

The ERLE measures the echo attenuation between microphone signal $y(n)$ and estimated near-end speech $\hat{s}(n)$, which is defined as:

$$\mathrm{ERLE} = 10\log_{10}\left\{\frac{E\left[y^2(n)\right]}{E\left[\hat{s}^2(n)\right]}\right\}(\mathrm{dB}) \tag{14}$$

where $E[\cdot]$ denotes the statistical expectation operation.

PESQ uses a cognitive model to compute the disturbance between the target speech and the processed speech, and it ranges from -0.5 to 4.5. The larger the score, the better the processed speech quality.

### 3.2. Datasets preparation

We use the TIMIT corpus [29], which consists of 630 speakers, each containing 10 utterances, for a total of 6300 utterances that sampled at 16 kHz. We first select 100 pairs of speakers as far-end and near-end signals, respectively. For each pair, we randomly select three utterances and concatenate them to form the far-end signal. The near-end signal has the same length as the far-end signal by adding zeros at both front and rear of the signal. We generate 5200 pairs of signals in total. 4000, 900 and 300 utterances are used for training, validation and testing respectively. It should be mentioned that the speakers in test set don't appear in training and validation sets.

We generate 7 different RIRs using the similar way reported in the literature [10]. All room impulse responses are generated by the image method [30] with reverberation time ($T_{60}$) being 200 $ms$, and the reflection order of RIR is set to 512. The simulation room size ($length \times width \times height$) is ($4 \times 4 \times 3$) $m$. Microphone is fixed at the center location of the room. A loudspeaker is placed at 7 random locations with 1.5 $m$ distance from the microphone. And we randomly choose 6 RIRs to generate echo signals for training, and use the remaining RIR for testing.

For training and validation sets, we generate the microphone signals at SER level randomly chosen from {-6, -3, 0, 3, 6}dB, by mixing the near-end speech signal and echo signal. The SER level here is evaluated in the double-talk period. It is defined as:

$$\mathrm{SER} = 10\log_{10}\left\{\frac{E\left[s^2(n)\right]}{E\left[d^2(n)\right]}\right\}(\mathrm{dB}) \tag{15}$$

And for testing mixtures, we generate the microphone signals at four different SER levels {-10, -5, 0, 5}dB.

### 3.3. Comparison methods and parameter settings

We compare our approach with two AEC algorithms. 1) NCC-NLMS: the conventional NLMS combined with the normalized Cross Correlation DTD [5]. The filter size is set to 512, the step size and the regularization factor are set to 0.2 and 0.06, respectively. 2) End-to-end learning method: directly estimate the PSM of near-end signal using far-end and microphone signals as input by LSTM. The LSTM has four hidden layers with 300 units in each layer. A fully connected layer used for feature extraction taken as LSTM input layer that have 322 units. $Sigmoid$ activation function used in a fully connected output layer that have 161 units.

## 4. Evaluation and comparison

### 4.1. Performance in double-talk situations

For the first experiment, we evaluate the proposed method in double-talk situations which can be treated as the matched scenario.

Table 1: *Average ERLE and PESQ scores in speech echo.*

| **SER** | | **-10dB** | **-5dB** | **0dB** | **5dB** |
|---|---|---|---|---|---|
| ERLE | None | — | — | — | — |
| | NCC-NLMS | 13.86 | 13.82 | 13.76 | 13.68 |
| | LSTM(end-to-end) | 20.48 | 23.36 | 28.68 | 28.09 |
| | LFM-NFM(pro.) | **31.56** | **36.59** | **38.65** | **38.42** |
| PESQ | None | 1.21 | 1.59 | 1.95 | 2.28 |
| | NCC-NLMS | 1.84 | 2.16 | 2.51 | 2.82 |
| | LSTM(end-to-end) | 1.49 | 1.81 | 2.16 | 2.49 |
| | LFM-NFM(pro.) | **2.31** | **2.79** | **3.18** | **3.50** |

Table 1 shows the average ERLE and PESQ scores of these methods in different SER conditions, where the results of 'None' (i.e. unprocessed speech) are calculated by comparing the $y(n)$ with $s(n)$ in the double-talk periods. The best scores in

each case are highlighted by boldface. In this table, the results demonstrate that all methods are capable of removing acoustic echoes. Taking the 0 dB SER case for example, it shows that going from LFM-NFM to NCC-NLMS improves ERLE by 24.89 and PESQ by 0.67. And our proposed algorithm significantly outperforms others in both metrics.

### 4.2. Performance of music echo

In training stage, the far-end signals are speech. In practice, music is also a very common echo. This experiment is to evaluate the generalization performance of AEC to music signals. We use GTZAN music library (available at http://marsyas.info), which contains 1000 different songs in 10 different genres with 100 songs in each genre and each song lasting about 30 seconds. We randomly select 300 songs, and resample at 16kHz.

Table 2: *Average ERLE and PESQ scores of music echo.*

| **SER** | | **-10dB** | **-5dB** | **0dB** | **5dB** |
|---|---|---|---|---|---|
| ERLE | None | — | — | — | — |
| | NCC-NLMS | 17.57 | 17.49 | 17.42 | 17.25 |
| | LSTM(end-to-end) | ***29.17*** | ***26.16*** | *21.72* | *17.02* |
| | LFM-NFM(pro.) | 22.72 | 23.57 | **24.47** | **25.47** |
| PESQ | None | 1.17 | 1.53 | 1.88 | 2.21 |
| | NCC-NLMS | 1.46 | 1.79 | 2.13 | 2.47 |
| | LSTM(end-to-end) | *0.98* | *1.20* | *1.46* | *1.58* |
| | LFM-NFM(pro.) | **2.38** | **2.70** | **2.93** | **3.05** |

The results of these methods in different SER conditions with background music echoes are shown in Table 2. Note that in the table, when the score of the PESQ of each algorithm is lower than 'None', we consider the algorithm to be invalid and the scores are shown with *italics*. From the table, we can see that LSTM is invalid for untrained music echoes. We can also find that the NCC-NLMS works well when dealing with non-stationary echo. In the case of 0 dB SER, it shows that the proposed LFM-NFM improves ERLE by 7.05 and PESQ by 0.8 to compared with NCC-NLMS. The proposed method consistently outperforms the conventional methods, and the performance generalizes well in untrained music echoes and SERs conditions.

### 4.3. Performance in unseen condition with nonlinear echo

In practice, loudspeaker and amplifier often cause nonlinearity. To test the generalization ability of the proposed algorithm, we follow the approach in literature [10, 12] to simulate the seriously nonlinear distortion echo captured by the microphone after passing through a power amplifier, a loudspeaker and acoustic transmission in order.

Firstly, the nonlinearity of power amplifier can be modeled using the hard-clipping way [31] by:

$$x_{hard}(n) = \begin{cases} -x_{\max} & x(n) < -x_{\max} \\ x(n) & |x(n)| \le x_{\max} \\ x_{\max} & x(n) > x_{\max} \end{cases} \tag{16}$$

where $x_{hard}(n)$ is the outputs of hard-clipping, and $x_{\max}(n)$ is set to 80% of the maximum value of the input signal. Then, in order to simulate an asymmetric loudspeaker distortion, we apply the following memoryless $sigmoid$ nonlinearity function [32] to the far-end signal:

$$x_{\mathrm{NL}}(n) = \gamma \left( \frac{1}{1 + e^{(-p \cdot q(n))}} - \frac{1}{2} \right) \tag{17}$$

where

$$q(n) = 1.5 \times x_{hard}(n) - 0.3 \times x_{hard}^2(n) \tag{18}$$

and the parameter $\gamma$ is the $sigmoid\ gain$ and it is set equal to 2, $p$ represents the $sigmoid\ slope$ and shown as:

$$p = \begin{cases} 4 & q(n) > 0 \\ 0.5 & q(n) \le 0 \end{cases} \tag{19}$$

Accordingly, the nonlinear distortion echo signals are generated by the $x_{\mathrm{NL}}(n)$ convolving with RIRs.

It should be mentioned that we do not add any nonlinearity in training stage.

Table 3: *Average ERLE and PESQ scores in nonlinear situations.*

| **SER** | | **-10dB** | **-5dB** | **0dB** | **5dB** |
|---|---|---|---|---|---|
| ERLE | None | — | — | — | — |
| | SVAF | 11.68 | 11.25 | 10.11 | 9.86 |
| | LSTM(end-to-end) | **12.61** | 14.62 | 14.34 | 14.16 |
| | LFM-NFM(pro.) | 12.22 | **15.34** | **17.41** | **18.45** |
| PESQ | None | 1.17 | 1.53 | 1.88 | 2.21 |
| | SVAF | 1.29 | 1.33 | 2.09 | 2.23 |
| | LSTM(end-to-end) | 1.25 | 1.56 | 1.92 | 2.30 |
| | LFM-NFM(pro.) | **1.53** | **1.94** | **2.31** | **2.67** |

Since the NLMS is not capable of dealing with nonlinear distortions, second-order Volterra adaptive filter (SVAF) [33] is employed to cancel the nonlinear echo in the microphone signal. The length of first order Volterra kernel is set to 512, second order length is 64, and the learning rate are set to 0.2 and 0.1, respectively. Table 3 shows the average ERLE and PESQ scores of these methods in different SER conditions with nonlinear distortions. Although the LSTM method has the strongest suppression of noise (12.61 dB for ERLE), its damage to near-end signal is the most serious (1.25 for PESQ). It can also be seen that the proposed method performs best in this nonlinear situation.

## 5. Conclusions

In this study, we propose a cascaded method to make AEC more robust. Different from traditional algorithms, the AEC problem is treated as a supervised learning task by predicting IRM for double-talk detection and PSM residual echo suppression. Experimental results show that proposed method outperforms the traditional methods in terms of the objective evaluation metrics in the matched scenario. Moreover, the results also show that the proposed method can significantly improve the removal of acoustic echo in the unmatched scenarios, and has good generalization performance, especially in low SER conditions, which is a promising sign for the practical use of AEC.

## 6. Acknowledgment

The authors would like to thank Hao Li and Hao Zhang for their helpful comments. This research was supported by the China National Nature Science Foundation (No.61876214).

# 7. References


[1] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. An application of very-high-order adaptive filters," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, 1999.

[2] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. John Wiley & Sons, 2005, vol. 40.

[3] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5. IEEE, 2006, pp. 281–284.

[4] D. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Transactions on Communications*, vol. 26, no. 5, pp. 647–653, 1978.

[5] M. Iqbal, J. Stokes, and S. Grant, "Normalized double-talk detection based on microphone and aec error cross-correlation," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, Jul. 2007, pp. 360–363.

[6] T. Gansler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communications*, vol. 44, no. 11, pp. 1421–1427, 1996.

[7] H. Geoffrey, D. Li, Y. Dong, E. George, and A.-r. Mohamed, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

[9] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

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

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

[12] C. Lee, J. Shin, and N. Kim, "Dnn-based residual echo suppression," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] Q. Lei, H. Chen, J. Hou, L. Chen, and L. Dai, "Deep neural network based regression approach for acoustic echo cancellation," in *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing - ICMSSP 2019*. ACM Press, 2019, pp. 94–98.

[14] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE, 2018, pp. 231–235.

[15] J. Costa, A. Lagrange, and A. Arliaud, "Acoustic echo cancellation using nonlinear cascade filters," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 5. IEEE, 2003, pp. V–389.

[16] G. Lazzarin, S. Pupolin, and A. Sarti, "Nonlinearity compensation in digital radio systems," *IEEE Transactions on Communications*, vol. 42, no. 234, pp. 988–999, 1994.

[17] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21–32, Jan. 1998.

[18] E. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[19] S. Haykin, *Adaptive filter theory*. Pearson Education India, 2005.

[20] W. Liu and S. Weiss, *Wideband beamforming: concepts and techniques*. Chichester, West Sussex, U.K.; Hoboken, N.J.: Wiley, 2010.

[21] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[22] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

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

[24] P. Loizou, *Speech enhancement: theory and practice*, 2nd ed. Boca Raton, Fla: CRC Press, 2013.

[25] H. Erdogan, J. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[26] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4390–4394.

[27] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 4, pp. 807–877.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2001, pp. 749–752.

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

[30] J. Allen and D. 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.

[31] 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.

[32] D. Comminiello, M. Scarpiniti, L. 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.

[33] A. Stenger, L. Trautmann, and R. Rabenstein, "Nonlinear acoustic echo cancellation with 2nd order adaptive volterra filters," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 2. IEEE, 1999, pp. 877–880.