Interspeech 2018
2-6 September 2018, Hyderabad

Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios

Hao Zhang[1], *DeLiang Wang*[1,2,3]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
[3]Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, China

{zhang.6720, wang.77}@osu.edu

Abstract

Traditional acoustic echo cancellation (AEC) works by identifying an acoustic impulse response using adaptive algorithms. We formulate AEC as a supervised speech separation problem, which separates the loudspeaker signal and the near-end signal so that only the latter is transmitted to the far end. A recurrent neural network with bidirectional long short-term memory (BLSTM) is trained to estimate the ideal ratio mask from features extracted from the mixtures of near-end and far-end signals. A BLSTM estimated mask is then applied to separate and suppress the far-end signal, hence removing the echo. Experimental results show the effectiveness of the proposed method for echo removal in double-talk, background noise, and nonlinear distortion scenarios. In addition, the proposed method can be generalized to untrained speakers.

Index Terms: Acoustic echo cancellation, double-talk, nonlinear distortion, supervised speech separation, ideal ratio mask, long short-term memory

1. Introduction

Acoustic echo arises when a loudspeaker and a microphone are coupled in a communication system such that the microphone picks up the loudspeaker signal plus its reverberation. If not properly handled, a user at the far end of the system hears his or her own voice delayed by the round trip time of the system (i.e. an echo), mixed with the target signal from the near end. The acoustic echo is one of the most annoying problems in speech and signal processing applications, such as teleconferencing, hands-free telephony, and mobile communication. Conventionally, the cancellation of echo is accomplished by adaptively identifying an acoustic impulse response between the loudspeaker and the microphone using a finite impulse response (FIR) filter [1]. Several adaptive algorithms have been proposed in the literature [1] [2]. Among them the normalized least mean square (NLMS) algorithm family [3] is most widely used due to its relatively robust performance and low complexity.

Double-talk is inherent in communication systems as it is typical of conversations when the speakers on both sides talk simultaneously. However, the presence of a near-end speech signal severely degrades the convergence of adaptive algorithms and may cause them to diverge [1]. The standard approach to solve this problem is to use a double-talk-detector (DTD) [4] [5], which inhibits the adaptation during double-talk periods.

The signal received at the microphone contains not only echo and near-end speech but also background noise. It is widely agreed AEC alone is incapable of suppressing background noise. A post filter [6] is usually applied to suppress background noise and residual echos that exist at the output of acoustic echo canceller. Ykhlef and Ykhlef [7] combined the adaptive algorithm with the short-time spectral attenuation based noise suppression technique and obtained a high amount of echo removal in the presence of background noise.

Many studies in the literature model the echo path as a linear system. However, due to the limitations of components such as power amplifiers and loudspeakers, a nonlinear distortion may be introduced to the far-end signal in the practical scenario of AEC. To overcome this problem, some works [8]-[9] proposed to apply a residual echo suppression (RES) to suppress the remaining echo caused by nonlinear distortion. Owing to the capacity of deep learning in modeling complex nonlinear relationships, it can be a powerful alternative to model the nonlinearity of AEC system. Malek and Koldovskỳ [10] modeled the nonlinear system as the Hammerstein model and used a two-layer feed-forward neural network followed by an adaptive filter to identify the model parameters. Recently, Lee *et al.* [11] have employed a deep neural network (DNN) to estimate the RES gain from both the far-end signal and the output of acoustic echo suppression (AES) [12] in order to remove the nonlinear components of echo signal.

The ultimate goal of AEC is to completely remove the far-end signal and the background noise so that only the near-end speech is sent to the far end. From the speech separation point of view, AEC can be naturally considered as a separation problem where the near-end speech is a source to be separated from the microphone recording and sent to the far end. Therefore, instead of estimating the acoustic echo path, we apply supervised speech separation to separate the near-end speech from the microphone signal with the accessible far-end speech as additional information [13]. In this approach, the AEC problem is addressed without performing any double-talk detection or post filtering.

Deep learning has shown great potential for speech separation [14] [15]. The ability of recurrent neural networks (RNNs) to model time varying functions can play an important role in addressing AEC problems. LSTM [16] is a variant of RNN that is developed to deal with the vanishing and exploding problem of traditional RNNs. It can model the temporal dependencies and has shown good performance for speech separation and speech enhancement in noisy conditions [17] [18]. In a recent study, Chen and Wang [19] employed LSTM to investigate speaker generalization for noise-independent models and the evaluation results showed that the LSTM model achieved better speaker generalization than a feed-forward DNN.

In this study, we use bidirectional LSTM (BLSTM) as the supervised learning machine to predict the ideal ratio mask

10.21437/Interspeech.2018-1484

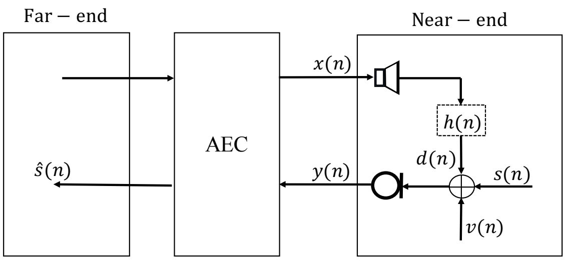

Figure 1: *Diagram of acoustic echo scenario.*

(IRM) from features extracted from mixture signals as well as far-end speech. We also investigate speaker generalization of the proposed method. Experimental results show that the proposed method is capable of removing acoustic echo in the noisy, double-talk and nonlinear distortion scenarios and generalizes well to untrained speakers.

The remainder of this paper is organized as follows. Section 2 presents the BLSTM based method. Experimental results are given in Section 3. Section 4 concludes the paper.

2. Proposed method

2.1. Problem formulation

Let us consider the conventional acoustic signal model, as shown in Fig. 1, where the microphone signal $y(n)$ consists of echo $d(n)$, near-end signal $s(n)$, and background noise $v(n)$:

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

An echo signal is generated by convolving a loudspeaker signal with a room impulse response (RIR). Then echo, near-end speech and background noise are mixed to generate a microphone signal. We formulate AEC as a supervised speech separation problem. As shown in Fig. 2, features extracted from microphone signal and echo are fed to the BLSTM. The estimated magnitude spectrogram of near-end signal is obtained by point-wise multiplying the estimated mask with the spectrogram of microphone signal. Finally, inverse short time Fourier transform (iSTFT) is applied to resynthesize $\hat{s}(n)$ from the phase of microphone signal and the estimated magnitude spectrogram.

2.2. Feature extraction

First the input signals ($y(n)$ and $x(n)$), sampled at 16 kHz, are divided into 20-ms frames with a frame shift of 10-ms. Then a 320-point short time Fourier transform (STFT) is applied to each time frame of the input signals, which results in 161 frequency bins. Finally, the log-magnitude spectral (LOG-MAG) feature [20] is obtained by applying the s logarithm operation to the magnitude responses. In the proposed method, features of microphone signal and far-end signal are concatenated as the input features.Therefore, the dimensionality of the input is $161 \times 2 = 322$.

2.3. Training targets

We use the ideal ratio masks [15] as the training target, which is defined as:

$$\mathrm{IRM}(m,c) = \sqrt{\frac{S^2(m,c)}{S^2(m,c) + D^2(m,c) + V^2(m,c)}} \tag{2}$$

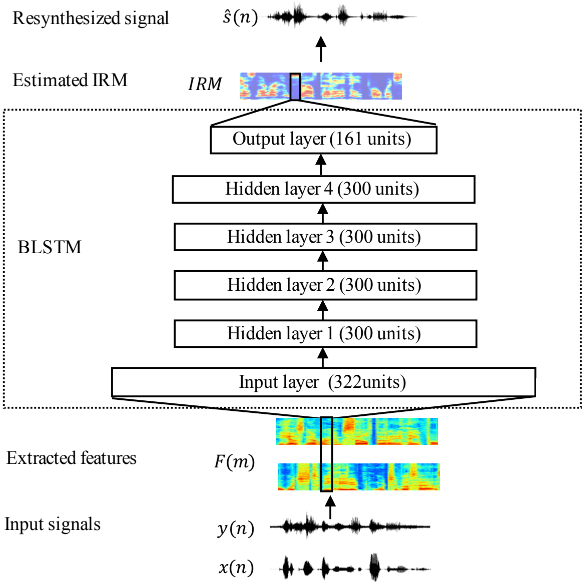

Figure 2: *Diagram of the proposed BLSTM based method.*

where $S^2(.)$, $D^2(.)$, $V^2(.)$ denote the energy of the near-end signal, acoustic echo, and background noise within a T-F unit at time m and frequency c, respectively.

2.4. Learning machines

Fig. 2 shows the BLSTM structure used in this paper. A BLSTM contains two unidirectional LSTMs, one of the LSTMs processes the signal in the forward direction while the other one in the backward direction. A fully connected layer is used for feature extraction. The BLSTM has four hidden layers with 300 units in each layer. The output layer is a fully-connected layer. Since IRM has the value range of $[0, 1]$, we use sigmoid function as the activation function in the output layer. Adam optimizer [21] and mean square error (MSE) cost function are used to train the LSTM. The learning rate is set to 0.0003. The number of training epochs is set to 30.

3. Experimental results

3.1. Performance metrics

Two performance metrics are used in this paper to compare system performance: echo return loss enhancement (ERLE) for single-talk periods (periods without near-end signal) and perceptual evaluation of speech quality (PESQ) for double-talk periods.

ERLE is used to evaluate the echo attenuation achieved by the system [3], which is defined as

$$\mathrm{ERLE} = 10 \log_{10} \left\{ \frac{\mathcal{E}[y^2(n)]}{\mathcal{E}[\hat{s}^2(n)]} \right\} \tag{3}$$

where $\mathcal{E}$ is the statistical expectation operation.

PESQ has a high correlation with subjective scores [22]. It is obtained by comparing the estimated near-end speech $\hat{s}(n)$ with the original speech $s(n)$. The range of PESQ score is from -0.5 to 4.5. A higher score indicates better quality.

In the following experiments, the performance of the conventional AEC methods is measured after processing the signals for around 3 seconds, i.e., the steady-state results.

3.2. Experiment setting

TIMIT dataset [23] is widely used in the literature [24] [5] to evaluate AEC performance. We randomly choose 100 pairs of speakers from the 630 speakers in the TIMIT dataset as the near-end and far-end speakers (40 pairs of male-female, 30 pairs of male-male, and 30 pairs of female-female). There are ten utterances sampled at 16 kHz for each speaker. Three utterances of the same far-end speaker are randomly chosen and concatenated to form a far-end signal. Each utterance of a near-end speaker is then extended to the same size as that of the far-end signal by filling zeros both in front and in rear. An example of how mixtures are generated will be shown later in Figure 3. Seven utterances of these speakers are used to generate mixtures and each near-end signal is mixed with five different far-end signals. So entirely we have 3500 training mixtures. The remaining three utterances are used to generate 300 test mixtures where each near-end signal is mixed with one far-end signal. To investigate the speaker generalization of the proposed method, we randomly chose another10 pairs of speakers (4 pairs of male-female, 3 pairs of male-male, and 3 pairs of female-female) from the rest of the 430 speakers in TIMIT dataset and generate 100 test mixtures of untrained speakers.

Room impulse responses are generated at reverberation time (T_{60}) of 0.2 s using the image method [25]. The length of RIR is set to 512. The simulation room size is $(4, 4, 3)$ m, and a microphone is fixed at the location of $(2, 2, 1.5)$ m. A loudspeaker is placed at 7 random places with 1.5 m distance from the microphone. Thus, 7 RIRs of different locations are generated, of which the first 6 RIRs are used to generate training mixtures and the last one is used to generate test mixtures.

3.3. Performance in double-talk situations

First we evaluate the proposed method in the double-talk situations and compare it with the conventional NLMS algorithm. Each training mixture, $x(n)$, is convolved with an RIR randomly chosen from the 6 RIRs to generate an echo signal $d(n)$. Then $d(n)$ is mixed with $s(n)$ at a signal-to-echo ratio (SER) randomly chosen from $\{-6, -3, 0, 3, 6\}$ dB. The SER level here is evaluated on the double-talk period. It is defined as:

$$\text{SER} = 10 \log_{10} \left\{ \frac{\mathcal{E}[s^2(n)]}{\mathcal{E}[d^2(n)]} \right\} \tag{4}$$

Since the echo path is fixed and there is no background noise or nonlinear distortion, the well known NLMS algorithm combined with the Geigel DTD [4] can work very well in this scenario. The filter size of NLMS is set to 512, which is the same as the length of simulated RIRs. The step size and regularization factor of NLMS algorithm [1] are set to 0.2 and 0.06, respectively. The threshold value of the Geigel DTD is set to 2.

Table 1 shows the average ERLE and PESQ values of these two methods in different SER conditions, where the results of 'None' (or unprocessed results) are calculated by comparing the microphone signal $y(n)$ with near-end speech $s(n)$ in the double-talk periods. The results shown in this table demonstrate that both NLMS and BLSTM methods are capable of removing acoustic echoes.The BLSTM based method outperforms NLMS in terms of ERLE while NLMS outperforms BLSTM in terms of PESQ.

Table 1: *Average ERLE and PESQ values in double-talk situations*

SER		0 dB	3.5 dB	7 dB
ERLE	NLMS	34.63	32.90	30.97
	BLSTM	51.61	50.04	47.42
PESQ	None	1.94	2.14	2.41
	NLMS	4.02	4.01	4.11
	BLSTM	2.74	2.92	3.15

Table 2: *Average ERLE and PESQ values in double-talk and background noise situations with 10 dB SNR*

SER		0 dB	3.5 dB	7 dB
ERLE	NLMS	8.03	6.06	4.14
	NLMS+Post-Filter[7]	23.20	22.79	22.28
	BLSTM	52.41	49.74	47.81
PESQ	None	1.76	1.92	2.03
	NLMS	2.10	2.16	2.20
	NLMS+Post-Filter[7]	2.59	2.66	2.71
	BLSTM	2.62	2.77	2.89

3.4. Performance in double-talk and background noise situations

The second experiment studies scenarios with double-talk and background noise. Since the NLMS with Geigel DTD alone is not capable of dealing with background noise, the frequency domain post-filter based AEC method [7] is employed to suppress the background noise at the output of AEC.

Similarly, each training mixture is mixed at a SER level randomly chosen from $\{-6, -3, 0, 3, 6\}$ dB. A white noise is added to the microphone signal at a SNR level randomly chosen from $\{8, 10, 12, 14\}$ dB. The SNR level here is evaluated on the double-talk period, which is defined as:

$$\text{SNR} = 10 \log_{10} \left\{ \frac{\mathcal{E}[s^2(n)]}{\mathcal{E}[v^2(n)]} \right\} \tag{5}$$

The average ERLE and PESQ values of NLMS, NLMS equipped with the post-filter and the BLSTM based method in different SER conditions with 10 dB SNR level are shown in Table 2. In the 'NLMS+Post-Filter' case, the filter size, step size and regularization factor of NLMS algorithm are set to 512, 0.02 and 0.06, respectively. The threshold value of the Geigel DTD is set to 2. The two forgetting factors of the post-filter are set to 0.99. As can be seen from the table, all of these methods show improvements in terms of PESQ when compared with the unprocessed results. BLSTM outperforms the other two methods in all conditions. In addition, by comparing Table 1 and Table 2, we find that adding the background noise to the microphone signal can seriously impact the performance of NLMS. And the post-filter can improve the performance of NLMS in this scenario.

3.5. Performance in double-talk, background noise and nonlinear distortion situations

The third experiment evaluates the performance of the BLSTM based method in the situations with double-talk, background noise and nonlinear distortion. A far-end signal is processed by the following two steps to simulate the nonlinear distortion introduced by a power amplifier and a loudspeaker.

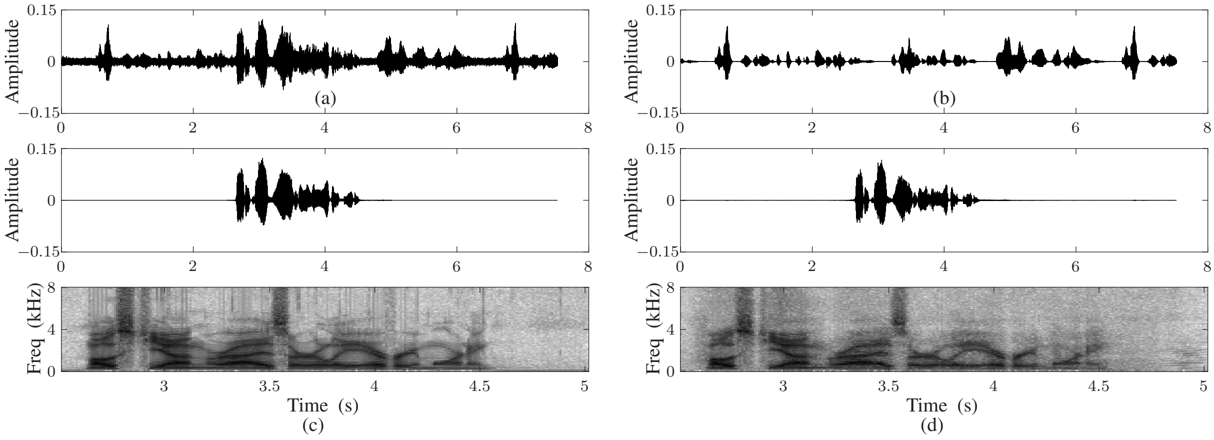

Figure 3: *Waveforms and spectrograms with 3.5 dB SER and 10 dB SNR. (a) microphone signal, (b) echo signal, (c) near-end speech, (d) BLSTM estimated near-end speech.*

First, a hard clipping [26] is applied to the far-end signal to mimic the characteristic of a power amplifier:

$$x_{\texttt{hard}}(n) = \begin{cases} -x_{\texttt{max}} & x(n) < -x_{\texttt{max}} \\ x(n) & |x(n)| \le x_{\texttt{max}} \\ x_{\texttt{max}} & x(n) > x_{\texttt{max}} \end{cases} \qquad (6)$$

where $x_{\texttt{max}}$ is set to 80% of the maximum volume of input signals.

Then the memoryless sigmoidal function [27] is applied to mimic the nonlinear characteristic of loudspeaker:

$$x_{\texttt{NL}}(n) = \gamma\left(\frac{2}{1+\exp(-a\cdot b(n))} - 1\right) \qquad (7)$$

where

$$b(n) = 1.5 \times x_{\texttt{hard}}(n) - 0.3 \times x^2_{\texttt{hard}}(n) \qquad (8)$$

The sigmoid gain γ is set to 4. The sigmoid slop a is set to 4 if $b(n) > 0$ and 0.5 otherwise.

For each training mixture, $x(n)$ is processed to get $x_{\texttt{NL}}(n)$, then this nonlinearly processed far-end signal is convolved with an RIR randomly chosen from the 6 RIRs to generate echo signal $d(n)$. SER is set to 3.5 dB and a white noise is added to the mixture at 10 dB SNR level.

Figure 3 illustrate an echo cancellation example by using the BLSTM based method. It can be seen that the output of the BLSTM based method resembles the clean near-end signal, which indicates that the proposed method can well preserve the near-end signal while suppressing the background noise and echo with nonlinear distortion.

We compare the proposed BLSTM method with the DNN-based residual echo suppression (RES) [11], the results are shown in Table 3. In our implementation of 'AES+DNN', the parameters for the AES and DNN are set to the values given in [11]. The 'SNR=∞' case, which is the situation evaluated in [11], shows that the DNN based RES can deal with the nonlinear component of echo and improve the performance of AES. When it comes to situations with background noise, adding the DNN based RES to AES shows minor improvement in terms of PESQ value. The BLSTM based method alone outperforms the AES+DNN.There is around 5.4 dB improvement in terms of ERLE and 0.5 improvement in terms of PESQ. If we follow the method proposed in [11] and add AES as a preprocessor to the BLSTM system, which is denoted as 'AES+BLSTM', the performance can be further improved. Moreover, it can be seen from Table 3 that the proposed BLSTM method can be generalized to untrained speakers.

Table 3: *Average ERLE and PESQ values in double-talk, background noise and nonlinear distortion situations with 3.5 dB SER, SNR=∞ means no background noise*

		None	AES [12]	AES+DNN [11]
SNR=∞	ERLE	-	11.49	36.59
	PESQ	2.09	2.57	2.71
		None	AES [12]	AES+DNN [11]
SNR=10 dB	ERLE	-	7.50	39.98
	PESQ	1.87	2.12	2.15
		None	BLSTM	AES+BLSTM
SNR=10 dB	ERLE	-	45.44	49.26
	PESQ	1.87	2.67	2.69
		None	BLSTM	AES+BLSTM
SNR=10 dB untrained speakers	ERLE	-	46.30	49.71
	PESQ	1.85	2.63	2.68

4. Conclusion

A BLSTM based supervised acoustic echo cancellation method is proposed to deal with situations with double-talk, background noise and nonlinear distortion. The proposed method shows its capability to remove acoustic echo and generalize to untrained speakers. Future work will apply this method to address other AEC problems such as multichannel communication.

5. Acknowledgement

The authors would like to thank M. Delfarah for providing his LSTM code and K. Tan for commenting on an earlier version. This research started while the first author was interning with Elevoc Technology, and it was supported in part by two NIDCD grants (R01 DC012048 and R01 DC015521).

6. References

[1] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay *et al.*, *Advances in network and acoustic echo cancellation*. Springer, 2001.

[2] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, *A perspective on stereophonic acoustic echo cancellation*. Springer Science & Business Media, 2011, vol. 4.

[3] 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.

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

[5] M. Hamidia and A. Amrouche, “A new robust double-talk detector based on the stockwell transform for acoustic echo cancellation,” *Digital Signal Processing*, vol. 60, pp. 99–112, 2017.

[6] V. Turbin, A. Gilloire, and P. Scalart, “Comparison of three post-filtering algorithms for residual acoustic echo reduction,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1. IEEE, 1997, pp. 307–310.

[7] F. Ykhlef and H. Ykhlef, “A post-filter for acoustic echo cancellation in frequency domain,” in *Complex Systems (WCCS), 2014 Second World Conference on*. IEEE, 2014, pp. 446–450.

[8] F. Kuech and W. Kellermann, “Nonlinear residual echo suppression using a power filter model of the acoustic echo path,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1. IEEE, 2007, pp. 73–76.

[9] A. Schwarz, C. Hofmann, and W. Kellermann, “Spectral feature-based nonlinear residual echo suppression,” in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[10] J. Malek and Z. Koldovskỳ, “Hammerstein model-based nonlinear echo cancellation using a cascade of neural network and adaptive linear filter,” in *Acoustic Signal Enhancement (IWAENC), 2016 IEEE International Workshop on*. IEEE, 2016, pp. 1–5.

[11] C. M. Lee, J. W. Shin, and N. S. Kim, “Dnn-based residual echo suppression,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] F. Yang, M. Wu, and J. Yang, “Stereophonic acoustic echo suppression based on wiener filter in the short-time fourier transform domain,” *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 227–230, 2012.

[13] J. M. Portillo, “Deep Learning applied to Acoustic Echo Cancellation,” Master’s thesis, Aalborg University, 2017.

[14] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: an overview,” *arXiv preprint arXiv:1708.07524*, 2017.

[15] Y. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[18] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, “Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr,” in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[19] 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.

[20] 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.

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

[22] 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 *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[23] 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.

[24] T. S. Wada, B.-H. Juang, and R. A. Sukkar, “Measurement of the effects of nonlinearities on the network-based linear acoustic echo cancellation,” in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–5.

[25] 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.

[26] 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.

[27] 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.