INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Deep Multitask Acoustic Echo Cancellation

Amin Fazel, Mostafa El-Khamy, Jungwon Lee

Samsung Semiconductor Inc., USA
{amin.fazel, mostafa.e, jungwon2.lee}@samsung.com

Abstract

Acoustic echo cancellation or suppression methods aim to suppress the echo originated from acoustic coupling between loudspeakers and microphones. Conventional approaches estimate echo using adaptive filtering. Due to the nonlinearities in the acoustic path of far-end signal, further post-processing is needed to attenuate these nonlinear components. In this paper, we propose a novel architecture based on deep gated recurrent neural networks to estimate the near-end signal from the microphone signal. The proposed architecture is trained using multitask learning to learn the auxiliary task of estimating the echo in order to improve the main task of estimating the clean near-end speech signal. Experimental results show that our proposed deep learning based method outperforms the existing methods for unseen speakers 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: Non-linear acoustic echo cancellation, deep learning, gated recurrent neural network, recurrent neural network, gated recurrent unit

1. Introduction

Acoustic echo is generated when the far-end signal playing out of a loudspeaker is coupled back to the microphone in the near-end point. Therefore, the far-end user hears a mixture of near-end signal and a delayed and modified version of his own voice, known as the acoustic echo. The goal of Acoustic echo canceller (AEC) or suppressor (AES) is to reduce this echo while leaving the speech of near-end user undistorted. Conventional methods address this problem by estimating the acoustic path with an adaptive filter [1]. Since most of these methods assume a linear relationship between acoustic echo and far-end signal, a nonlinear post-filtering is typically applied to suppress the remained residual echoes [2][3].

Neural network has been used as a nonlinear post-filtering in the past [4]. However, at the time, constraints in computational power and size of training data resulted in relatively small network implementation and limited overall AEC performance. Recent advancements in deep learning have shown great potential in various speech processing tasks [5][6][7][8], but not many works considered acoustic echo cancellation. Lee et al. [9] used a deep neural network (DNN) to estimate the gain of residual echo suppression. Recurrent neural networks (RNNs) have shown great success in sequence modeling tasks such as natural language processing (NLP), especially if they are used in an encoder-decoder framework [10] or as a sequence to sequence learning machine [11]. RNNs are particularly powerful in these frameworks as they are highly capable of modeling rich context dependencies that are inherent in these tasks. Recently, Zhang and Wang [12] used a bidirectional long-short term memory (BLSTM) to predict a mask from features of microphone and far-end signals, which is then used to resynthesize the near-end speech signal. To enable echo estimation, conventional methods typically require a double-talk detector (DTD) to halt the filter adaptation during double-talk periods, when both the near-end and far-end signals are simultaneously present. In comparison, some deep learning based echo canceller systems do not need a separate DTD module for cancelation of the acoustic echoes [9] [12].

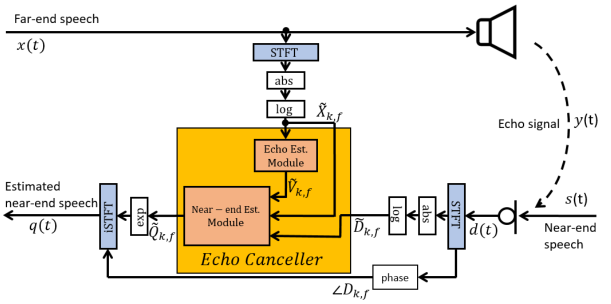

Figure 1: *Diagram of proposed deep multitask acoustic echo cancellation.*

In this paper, we propose a new recurrent network for acoustic echo cancellation. More specifically, we use deep gated recurrent unit (GRU) [10][13] networks in an encoder-decoder architecture to map the spectral features of the microphone and far-end signals to a hyperspace, and then decode the target spectral features of the near-end signal from the encoded hyperspace. The proposed architecture is trained using multitask learning to learn the auxiliary task of estimating the echo in order to improve the main task of estimating the clean near-end speech signal. The experimental results show that our proposed model can cancel acoustic echo in both single-talk and double-talk periods with nonlinear distortions without requiring a separate DTD.

The remainder of this paper is structured as follows. We first provide a formal definition of the problem in section 2. We then present our context-aware multitask recurrent network in section 3, followed by experimental settings and results in section 4. Finally, we conclude in section 5.

2. Problem statement

Let *v(t)* be an arbitrary time-domain signal at time *t*. The system model and proposed solution are illustrated in Figure 1. The microphone signal $d(t)$ consists of near-end speech signal $s(t)$ and acoustic echo $y(t)$:

$$d(t) = s(t) + y(t). \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2019-2908

The acoustic echo signal is a modified version of far-end speech signal $x(t)$ and includes room impulse response (RIR) and loudspeaker distortion.

The acoustic echo cancellation problem is to retrieve the clean near-end signal after removing any echo due to the far-end signal. Conventional systems estimate a model of the echo path with a linear adaptive filter and then subtract the estimated echo from the microphone signal. A residual echo suppressor (RES) can be further applied to improve the near-end signal. RES is typically realized by Wiener filter or spectral subtraction in the frequency domain. The final output of AEC system is estimated near-end signal $q(t)$.

The echo return loss enhancement metric (ERLE) is often used to evaluate the echo reduction that is achieved by the system during the single-talk situations when there is no near-end signal. ERLE is defined as:

$$ERLE(\mathrm{dB}) = 10\log_{10}\frac{E\{d^2(t)\}}{E\{\mathrm{q}^2(t)\}} \quad (2)$$

where E is the statistical expectation operation which is realized by averaging.

To evaluate the performance of the system during the double-talk periods, the perceptual evaluation of speech quality (PESQ) is often used. The PESQ is calculated by comparing the estimated near-end speech against the ground-truth near-end speech during the double-talk periods only. The PESQ score ranges from -0.5 to 4.5 and a higher score indicates better quality.

We assume the audio signals are sampled at 16 kHz. The spectral feature vectors are computed using a 512-point short time Fourier transform (STFT) with a frame shift of 256-point (16ms). The 512-point STFT magnitude vector is reduced to 257-point by removing the conjugate symmetric half. The final logarithmic magnitude spectral feature vector is extracted by applying the logarithmic operation to the STFT magnitude. The input features are standardized to have zero mean and unit variance using the scalars calculated from the training data. The STFT complex-valued spectrum of *v(t)* at frame k and frequency bin *f* is denoted by $V_{k,f}$. Its phase is denoted by $\angle V_{k,f}$ and its logarithmic magnitude is denoted by $\tilde{V}_{k,f}$. Let $\widetilde{\boldsymbol{V}}_k$ be the vector of logarithmic magnitudes at all frequency bins and frame k.

3. Proposed method

In this paper, we propose to estimate the near-end speech signal with a context-aware multitask gated RNN. Specifically, we use logarithmic spectral features of far-end speech x and microphone d as inputs. The target outputs include logarithmic spectral features of ground-truth echo signal y and near-end speech signal s. The proposed architecture estimates both near-end speech and echo signals by jointly optimizing a weighted loss. The information from estimating echo is used to better estimate the near-end speech. To our knowledge, this paper is first to propose a multitask network for AEC.

Figure 2 describes the proposed multitask AEC framework. The echo estimation module constitutes of two-layer stacked GRU networks and is trained to generate an estimate $\widetilde{\boldsymbol{V}}_k$ of the echo signal. The output of the last GRU layer from this network is fed into another three-layer stacked GRU networks along with $\widetilde{\boldsymbol{D}}_k$ and $\widetilde{\boldsymbol{X}}_k$ and are regressed to $\widetilde{\boldsymbol{Q}}_k$ which is an estimate of logarithmic spectrum magnitude of the near end signal. The time domain signal can be generated from $\widetilde{\boldsymbol{Q}}_k$ and the phase of the microphone signal using inverse short time Fourier transform (iSTFT) or the Griffin-Lim algorithm [14]. For simplicity in this work, we only show the results using iSTFT reconstruction.

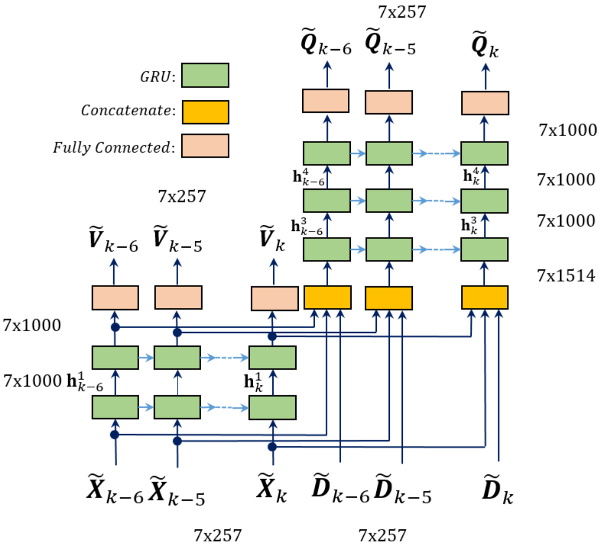
7x257
$\widetilde{\boldsymbol{Q}}_{k-6}$
$\widetilde{\boldsymbol{Q}}_{k-5}$
$\widetilde{\boldsymbol{Q}}_{k}$
GRU:
Concatenate:
Fully Connected:
7x1000
$\mathbf{h}^4_{k-6}$
$\mathbf{h}^4_{k}$
7x1000
7x257
$\mathbf{h}^3_{k-6}$
$\mathbf{h}^3_{k}$
7x1000
$\widetilde{\boldsymbol{V}}_{k-6}$
$\widetilde{\boldsymbol{V}}_{k-5}$
$\widetilde{\boldsymbol{V}}_{k}$
7x1514
7x1000
7x1000
$\mathbf{h}^1_{k-6}$
$\mathbf{h}^1_{k}$
$\widetilde{\boldsymbol{X}}_{k-6}$
$\widetilde{\boldsymbol{X}}_{k-5}$
$\widetilde{\boldsymbol{X}}_{k}$
$\widetilde{\boldsymbol{D}}_{k-6}$
$\widetilde{\boldsymbol{D}}_{k-5}$
$\widetilde{\boldsymbol{D}}_{k}$
7x257
7x257

Figure 2: *Proposed multitask unrolled GRU networks for echo and near-end signal estimation.*

3.1. Causal context-aware inputs and outputs

It has been shown in previous studies that using past and/or future frames can help the estimation of current frame for speech processing applications [15]. However, a fixed context window is typically used as the input to a fully-connected layer [16]. In these methods, the contextual information can be lost after the first layer as the information flows through deeper layers. In his work, we used the context features for both inputs and outputs of our networks in order to keep the contextual information throughout the network. The input features for current time consists of the feature vector of current frame and vectors of six previous frames. Causal windows are chosen to prevent extra latency. Seven frames with 50% overlap creates a receptive filed of 112ms which is long enough for processing the speech signal. To incorporate the context awareness, we deployed unrolled deep GRU networks with 7 time-steps for both the echo estimation module and the near-end estimation module, as shown in Figure 2.

Outputs of the network also consist of current frame feature vector and six previous frames. During the training, each one of these frames are optimized against their own targets. This helps the model to learn the weights based on the context of the targets. In the inference time, the last frame is only considered as the output of the model.

3.2. Multitask GRU based AEC

The underlying model architecture of our proposed AEC method consists of a variant of GRU. More specifically, the GRUs have the following output activation:

$$\mathbf{h}_k = \mathbf{z}_k \odot \mathbf{h}_{k-1} + (1 - \mathbf{z}_k) \odot \hat{\mathbf{h}}_k \quad (3)$$

where ⊙ is an element-wise multiplication, and the update gates $\mathbf{z}_k$ are

$$\mathbf{z}_k = \sigma\left(\mathbf{W}_z\widetilde{\boldsymbol{X}}_k + \mathbf{U}_z\mathbf{h}_{k-1}\right) \tag{4}$$

where σ is a sigmoid function. The candidate hidden state $\hat{\mathbf{h}}_k$ is computed by

$$\hat{\mathbf{h}}_k = \mathrm{elu}\left(\mathbf{W}\widetilde{\boldsymbol{X}}_k + \mathbf{U}(\mathbf{r}_k \odot \mathbf{h}_{k-1})\right) \tag{5}$$

where elu is exponential linear unit function and reset gates $\mathbf{r}_k$ are computed by

$$\mathbf{r}_k = \sigma\left(\mathbf{W}_r\widetilde{\boldsymbol{X}}_k + \mathbf{U}_r\mathbf{h}_{k-1}\right) \tag{6}$$

where $\mathbf{U}$, $\mathbf{W}$, $\mathbf{U}_r$, $\mathbf{W}_r$, $\mathbf{U}_z$, and $\mathbf{W}_z$ are the internal weight matrices of the GRUs.

Our deep learning AEC model consists of two stacked GRU networks. The first stack takes context-aware frames $\widetilde{\boldsymbol{X}}_k$ as inputs to each GRU and estimates $\widetilde{\boldsymbol{V}}_k$ using a fully connected (FC) output layer with linear activation. The outputs of last GRU layer from the first stack get concatenated with the context-aware frames $\widetilde{\boldsymbol{X}}_k$ and $\widetilde{\boldsymbol{D}}_k$ to create inputs of 7×1514 dimension for the first GRU layer of second stack. Second stack consists of 3 GRU layers and a FC layer with linear activation to estimate the context-ware frames $\widetilde{\boldsymbol{Q}}_k$ of estimated near-end speech. In Figure 2, $\widetilde{\boldsymbol{X}}_k$ is a feature vector of size 257 for the frame *k,* and $\mathbf{h}_k^1$ is the output vector of GRU in the first layer with the size of 1000. The output dimensions of each layer is shown in Figure 2.

All models were trained using AMSGrad optimization [17] (Adam variant [18]) by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-3}$ for 100 epochs, with a batch size of 100. The weights of all layers are initialized with Xavier method [19] and the biases are set to zero. We set the learning rate to 0.0003. To avoid overfitting, we use L2 regularization for all the weights with a regularization constant of 0.000001.

3.3. Weighted loss function

A common loss function for speech processing application is mean square error (MSE) [12] that is calculated between ground-truth source s and network estimated output q in the feature domain (usually STFT). Since estimating the echo path signal is expected to provide more information to determine the network weights (as in the convolutional solutions), we propose a weighted loss function in order to use that information. This function is jointly optimized:

$$loss_k = \beta \textstyle\sum_{n=0}^{6} \left\|\widetilde{\boldsymbol{S}}_{k-n} - \widetilde{\boldsymbol{Q}}_{k-n}\right\|_1 + (1-\beta) \textstyle\sum_{n=0}^{6} \left\|\widetilde{\boldsymbol{Y}}_{k-n} - \widetilde{\boldsymbol{V}}_{k-n}\right\|_1 \tag{7}$$

where β is the weighting factor.

4. Experimental Results

4.1. Dataset Preparation

We used TIMIT dataset [20] to evaluate AEC performance. We created the dataset similar to the one reported in [12], specifically the following steps have been taken: From 630 speakers of TIMIT, 100 pairs of speakers (40 male-female, 30 male-male, 30 female-female) are randomly chosen to be used as the far-end and near-end speakers. Three utterances of the same far-end speaker are randomly chosen and concatenated to create 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 zeroes both in front and in rear. Seven utterances of near-end speakers are used to generate 3500 training mixtures where each near-end signal is mixed with five different far-end signal. From the remaining 430 speakers, we randomly picked another 100 pairs of speakers as the far-end and near-end speakers. We followed the same procedure as described above, but this time only three utterances of near-end speakers are used to generate 300 testing mixtures where each near-end signal is mixed with one far-end signal. Therefore, the testing mixtures are from untrained speakers.

The following processes were applied to the far-end signal to model the nonlinear acoustic path as in [21]. For the nonlinear model of acoustic path, we first applied the hard clipping to simulate the power amplifier of loudspeaker (x_{max} is set to 80% of the maximum volume of input signal):

$$x_{clip}(t) = \begin{cases} -x_{max} \; if \; x(t) < -x_{max} \\ x(t) \quad if \; |x(t)| \le x_{max} . \\ \; x_{max} \quad if \; x(t) > x_{max} \end{cases} \tag{8}$$

Then, to simulate the loudspeaker distortion, we applied the following sigmoidal function:

$$x_{nl}(t) = 4\left(\frac{2}{1+\exp(-a.b(t))} - 1\right) \tag{9}$$

where $b(t) = 1.5x_{clip}(t) - 0.3x_{clip}(t)^2$ and $a = 4$ if $b(t) > 0$ and $a = 0.5$ otherwise. Finally, the output of sigmoidal function is convolved with a randomly chosen RIR $g(t)$ in order to simulate the acoustic transmission of far-end signal in the room:

$$y_{nl}(t) = x_{nl}(t) * g(t) \tag{10}$$

where $*$ indicates convolution. The length of RIRs is set to 512, the simulation room size is 4m×4m×3m, and a microphone is fixed at the location of [2 2 1.5] m. A loudspeaker is placed at seven random places with 1.5 m distance from the microphone. The RIRs are generated using image method [22] with reverberation time (T_{60}) of 200ms. From 7 RIRs, we used the first six RIRs to generate training data and the last one is used to generate testing data. We also modelled a linear acoustic path by only convolving the far-end signal with RIR to generate the echo signal, clipping and loudspeaker distortion are not applied for this model:

$$y_{lin}(t) = x(t) * g(t). \tag{11}$$

For training mixtures, we generated the microphone signals at signal to echo ratio (SER) level randomly chosen from {-6, -3, 0, 3, 6}dB by mixing the near-end speech signal and echo signal. The SER level is calculated on the double-talk period as:

$$\mathrm{SER(dB)} = 10\log_{10}\frac{E\{s^2(t)\}}{E\{y^2(t)\}}. \tag{12}$$

For test mixtures, we generated the microphone signals at three different SER levels (0dB, 3.5dB, and 7dB). The PESQ scores of unprocessed test mixtures for linear model are 1.87, 2.11, and 2.34 and for nonlinear model are 1.78, 2.03, and 2.26 at SER levels 0dB, 3.5dB, and 7dB, respectively. The unprocessed PESQ scores are calculated by comparing the microphone signal against near-end signal during the double-talk period.

4.2. Numerical results

As the benchmark system, we used a frequency domain normalized least mean square (NLMS) as an AES [23]. A DTD is used based on the energy of microphone signal and far-end signal. We further applied a RES algorithm based on the method presented in [24]. We also compare our results against the bidirectional long short-term memory (BLSTM) method that was reported in [12].

Table 1: *ERLE and PESQ scores in linear model of acoustic path.*

	Method	Testing SER (dB)		
		0	3.5	7
ERLE (dB)	AES+RES	29.38	25.88	21.97
	BLSTM [12]	51.61	50.04	47.42
	CA Single task GRU	62.88	61.81	60.11
	CA Multitask GRU	64.66	64.16	62.26
PESQ gain	AES+RES	0.93	0.81	0.68
	BLSTM [12]	0.80	0.78	0.74
	CA Single task GRU	0.98	0.95	0.93
	CA Multitask GRU	1.04	1.02	0.99

We first evaluated our proposed method using linear model of acoustic path. Table 1 shows the average ERLE values and PESQ gains for the conventional benchmark, BLSTM, and our proposed context-aware multitask GRU which is denoted as "CA Multitask GRU". The PESQ gain is calculated as the difference of PESQ value of each method with respect to its unprocessed PESQ value. This table also shows the results for context-aware single task GRU (denoted as "CA Single task GRU") that only uses the second stack of GRU layers with $\widetilde{\boldsymbol{D}}_k$ and $\widetilde{\boldsymbol{X}}_k$ as the inputs where the loss function is calculated by only penalizing the network outputs against ground-truth feature vector of near-end speech. The results show that multitask GRU outperforms single task GRU in terms of both PESQ and ERLE. It also shows that the proposed method outperforms both conventional AES+RES and BLSTM methods in all conditions.

Table 2: *ERLE and PESQ scores in nonlinear model of acoustic path.*

	Method	Testing SER (dB)		
		0	3.5	7
ERLE (dB)	AES+RES	16.76	14.26	12.33
	AES+DNN [9]	-	36.59	-
	CA Multitask GRU	61.79	60.52	59.47
PESQ gain	AES+RES	0.54	0.43	0.31
	AES+DNN [9]	-	0.62	-
	CA Multitask GRU	0.84	0.83	0.81

We further studied the impact of nonlinear model of acoustic path on our proposed method. In this set of experiments, we used $y_{nl}(t)$ in generating the microphone signals, therefore our model contains both power amplifier clipping and loudspeaker distortions. We again compared results of our method against conventional AES+RES. We also compared our results against the AES that uses DNN-based RES that was proposed in [9] and denoted as 'AES+DNN'. The results show that the proposed method outperforms the other two methods in both PESQ and ERLE. Spectrograms in Figure 3 illustrate an AEC example using our proposed deep multitask AEC in nonlinear model of acoustic path with 0dB SER. Evidently, the proposed method achieves a superior echo reduction without significant near-end distortion.

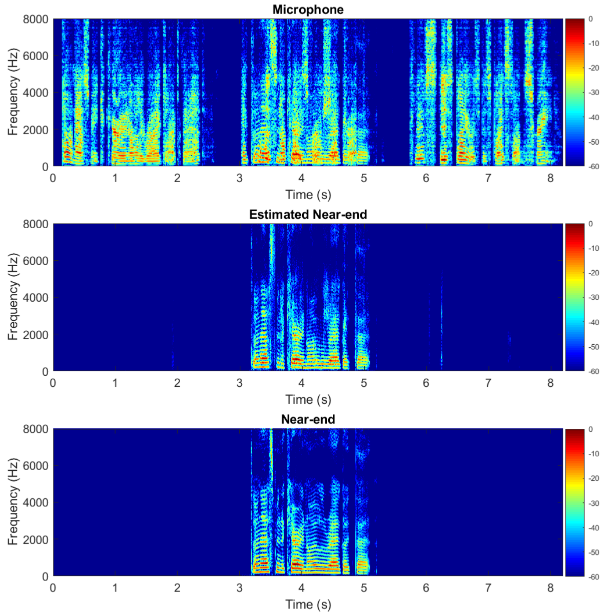

Figure 3: *Spectrograms of microphone, estimated near-end, and near-end signals in nonlinear model of acoustic path and 0dB SER.*

We also evaluated the performance of our proposed method in presence of additive noise and nonlinear model of acoustic path. When generating the training data, we added a white noise at 10dB SNR level with nonlinear acoustic path at 3.5dB SER level. We compared our method against a conventional AES+RES. Our multitask based method outperforms the conventional method by a large margin as shown in Table 3.

Table 3: *ERLE and PESQ scores in nonlinear model of acoustic path (SER=3.5dB) and additive noise (SNR =10dB).*

ERLE (dB)	AES+RES	10.13
	CA Multitask GRU	46.12
PESQ	None	1.80
	AES+RES	2.01
	CA Multitask GRU	2.50

5. Conclusions

We proposed a novel deep multitask recurrent neural network for AEC which performs well in both single-talk and double-talk periods. We demonstrate the benefit of end-to-end multitask learning of both the echo and the near-end signal simultaneously. We also demonstrate the benefit of having low latency causal context windows to improve the context-awareness when estimating the clean near-end signal. When compared on reference datasets, our proposed multitask AEC network can reduce the echo more significantly than other published methods, and is robust to additive background noises. As future work, we intend to explore AEC in environments with more severe background noises.

6. References

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach, Adaptive and learning systems for signal processing, communications, and control*. Hoboken, N.J, USA: Wiley-Interscience, 2004.

[2] 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, 1998.

[3] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 307–310.

[4] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *Proc. Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1-4.

[5] G. Hinton, L. Deng, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp.82–97, 2012.

[6] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Annual Conference of the International Speech Communication Association,* 2013, pp. 555–559.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[9] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in *Proc. Annual Conference of the International Speech Communication Association*, 2015, pp. 1775–1779.

[10] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwen, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proc. Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Advances Neural Information Processing Systems*, 2014, pp. 3104–3112.

[12] H. Zhang and D. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Proc. Annual Conference of the International Speech Communication Association*, 2018, pp. 3239-3243.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *in Proc. NIPS Deep Learning Workshop,* 2014.

[14] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] F. Santos and T. H. Falk., "Speech Dereverberation With Context-Aware Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no 7, pp. 1236–1246, 2018.

[16] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[17] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *International Conference on Learning Representations (ICLR)*, 2018.

[18] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[19] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249-256.

[20] . 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.

[21] 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.

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

[23] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 13, no. 5, pp. 1048–1062, 2005.

[24] R. Martin and S. Gustafsson, "The echo shaping approach to acoustic echo control", *Speech Communication*, vol. 20, no. 3-4, pp. 181-190, 1996.