DEEP ADAPTIVE AEC: HYBRID OF DEEP LEARNING AND ADAPTIVE ACOUSTIC ECHO CANCELLATION

Hao Zhang[1]*, Srivatsan Kandadai*[2]*, Harsha Rao*[2]*, Minje Kim*[3]*, Tarun Pruthi*[2]*, Trausti Kristjansson*[2]

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[2]Amazon Inc., Sunnyvale, CA, USA
[3]Department of Intelligent Systems Engineering, Indiana University, Bloomington, IN, USA

ABSTRACT

In this paper we integrate classic adaptive filtering algorithms with modern deep learning to propose a new approach called deep adaptive AEC. The main idea is to represent the linear adaptive algorithm as a differentiable layer within a deep neural network (DNN) framework. This enables the gradients to flow through the adaptive layer during back propagation and the inner layers of the DNN are trained to estimate the playback reference signal and the time-varying learning factors. The proposed approach combines the power of DNNs with adaptive filters. Experimental results show the effectiveness of the proposed method in scenarios where the echo path changes continuously and signal-to-echo ratio (SER) and signal-to-noise ratio (SNR) are low. Furthermore, compared to fully DNN-based baseline methods, integrating adaptive algorithm consistently improves performance and leads to easier training using smaller models.

Index Terms— Deep learning, acoustic echo cancellation, echo path change, deep adaptive AEC

1. INTRODUCTION

Acoustic echo cancellation (AEC) has received significant attention for several decades [1, 2, 3, 4]. The goal of AEC is to cancel echo caused by playback or far-end speech and transmit only the near-end speech to the far end. Though many algorithms have been proposed in literature [3, 4], it remains a challenging problem, especially in conditions with continuously changing echo-paths in low signal-to-echo ratio (SER) and signal-to-noise ratio (SNR).

Conventionally, AEC is achieved by identifying a linear transfer function between loudspeaker and microphone using adaptive filtering algorithms [3] such as normalized least mean square (NLMS) and affine projection [5, 6, 7]. The performance of these algorithms depend on how well their parameters control the speed of convergence while keeping misalignment in check. Especially during double-talk and echo path change, where convergence rates have to compromise between the two. Furthermore, traditional AEC algorithms are linear and cannot estimate nonlinearities introduced in the echo path by systems like amplifiers and loudspeakers [8, 9].

Deep learning has been utilized recently for solving AEC problems [10, 11]. Its capacity in modeling complex nonlinear relations leads to improved performance [12, 13, 14]. Deep learning based methods formulate AEC as speech separation and work by training a network to directly separate target signal from the microphone signal [15, 16]. They are powerful at handling nonlinear distortions and can achieve joint echo and noise reduction without the need for a double-talk detector (DTD) [17] or post-filtering [18]. However, generalization to untrained situations is crucial for deep learning based methods – complex networks and large-scale training are usually utilized to address this problem [19].

Recent studies have shown the advantages of using differentiable digital signal processing (DDSP) elements inside deep learning networks [20, 21, 22, 23]. A DDSP library that enables integration of signal processing elements with deep learning methods is introduced in [20]. Ramírez et al. trained a network with DDSP for automating audio signal processing [21]. Ivry et al. [24] introduced a nonlinear AEC method that jointly optimizes the network and a standard adaptive filter. It has also been shown that using DDSP within DNN can potentially make training easier and models smaller [20, 22].

In this study, we combine an adaptive linear AEC algorithm with deep learning and propose a new approach, called deep adaptive AEC. Specifically, a DNN model is trained for step size parameter and reference signal estimation, and these estimates are then used by the adaptive AEC to remove echo. The adaptive AEC algorithm is implemented as a differentiable layer with no trainable parameters, hence the gradients can flow through it during training to update the DNN parameters. During the inference stage, the parameters in DNN are fixed while the adaptive filter performs echo cancellation. The proposed method benefits from the adaptive linear filtering algorithm while retaining the power of deep learning. It is worth noting that the proposed approach enables the use of any adaptive algorithm within any DNN framework. In this paper, we utilize the NLMS [25] and a recurrent neural network (RNN) with long short-term memory (LSTM) [26] as the linear AEC and DNN module, respectively. Experimental results show that our proposed approach is effective for echo and noise removal on challenging situations with continuously changing echo paths and low SER and SNR levels.

The remainder of this paper is organized as follows. Section 2 introduces the proposed method. Experimental results are shown in Section 3. Conclusions are presented in Section 4.

2. METHOD DESCRIPTION

2.1. Signal model and classical AEC algorithm

In a typical acoustic signal model, the microphone signal is a mixture of echo, near-end speech, and background noise, and its frequency domain representation is given as:

$$Y_{k,m} = D_{k,m} + S_{k,m} + N_{k,m} \quad (1)$$

where $Y_{k,m}$, $D_{k,m}$, $S_{k,m}$, and $N_{k,m}$ denote the short-time Fourier transform (STFT) of microphone signal, echo, near-end speech, and noise at frame index k and frequency index m, respectively.

The paper describes work performed at Amazon.

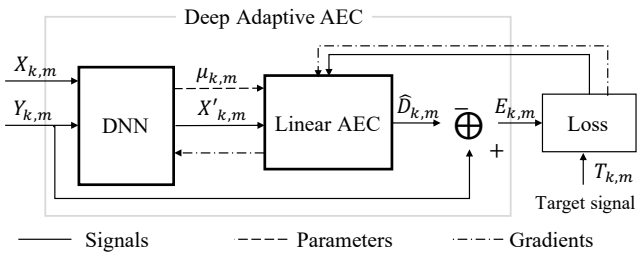

Fig. 1. Diagram of the proposed solution for acoustic echo cancellation.

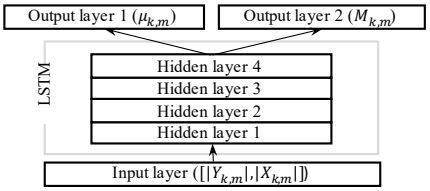

Fig. 2. Diagram of the DNN module.

Given an input signal $Y_{k,m}$ and a reference signal (far-end signal) $X_{k,m}$, traditional AEC algorithms such as NLMS achieve echo removal by updating an adaptive filter to estimate an acoustic echo path denoted by $\hat{\mathbf{W}}_{k,m}$. The estimated echo signal $\hat{D}_{k,m}$ is then subtracted from $Y_{k,m}$ to get the system output (the error signal) $E_{k,m}$.

$$E_{k,m} = Y_{k,m} - \hat{D}_{k,m},\ \hat{D}_{k,m} = \hat{\mathbf{W}}_{k,m}^{H}\mathbf{X}_{k,m} \tag{2}$$

$$\hat{\mathbf{W}}_{k+1,m} = \hat{\mathbf{W}}_{k,m} + \frac{\mu_{k,m}}{\mathbf{X}_{k,m}^{H}\mathbf{X}_{k,m}} E_{k,m}\mathbf{X}_{k,m} \tag{3}$$

where $\hat{\mathbf{W}}_{k,m} = [\hat{W}_{k,m}, \hat{W}_{k-1,m}, \cdots, \hat{W}_{k-L+1,m}]^T$ denotes an adaptive filter of length L, $\mathbf{X}_{k,m}$ is the buffered input, $\mu_{k,m}$ denotes the step size, and the superscript ^{H} represents conjugate transpose.

Classical AEC algorithms are faced with two main challenges. Firstly, the step size $\mu_{k,m}$, which determines the learning rate of the adaptive filter, needs to be chosen carefully to guarantee the convergence of algorithm and achieve acceptable echo removal. Estimation of the optimal step size is further made difficult in situations with continuous echo path variations. Secondly, many traditional algorithms model the echo signal as a linear transform of $X_{k,m}$ and fail to model any nonlinear distortions introduced by the amplifier, loudspeaker and acoustics in the echo path.

2.2. Deep adaptive AEC

A deep adaptive AEC solution, as is shown in Fig. 1, is proposed to address the above challenges. This method achieves echo removal by implementing a linear AEC within a DNN framework where the step size parameter $\mu_{k,m}$ and a reference signal $X'_{k,m}$ for the linear AEC module are estimated by the DNN module.

2.2.1. DNN module

The DNN module takes microphone and far-end signal as inputs to estimate a step size $\mu_{k,m}$ and a reference signal $X'_{k,m}$:

$$\mu_{k,m} = f(Y_{k,m}, X_{k,m}),\ \ X'_{k,m} = g(Y_{k,m}, X_{k,m}) \tag{4}$$

where $f(\cdot)$ and $g(\cdot)$ represent the nonlinear transform functions learned by DNN for estimating $\mu_{k,m}$ and $X'_{k,m}$, respectively.

In this study, we implement the DNN module using an LSTM network, as is shown in Fig. 2. The LSTM has four hidden layers with 300 units in each layer. It is trained to estimate two outputs, step size $\mu_{k,m}$ and a spectral magnitude mask $M_{k,m}$, from the input features. The value range of both $\mu_{k,m}$ and $M_{k,m}$ is [0, 1] and sigmoid is used as the activation function for the output layers. The estimated complex spectrogram of $X'_{k,m}$ is obtained through:

$$X'_{k,m} = |Y_{k,m}| \cdot M_{k,m} \cdot e^{j\theta_{Y_{k,m}}} \tag{5}$$

where $|Y_{k,m}|$ and $\theta_{Y_{k,m}}$ denote the magnitude spectrogram and phase of $Y_{k,m}$, respectively, $\cdot$ denotes point-wise multiplication, and j represents the imaginary unit.

2.2.2. Linear AEC module

The DNN module outputs, together with the microphone signal are provided to the linear AEC module for echo removal. We utilize frequency-domain NLMS as the linear AEC module and replace the step size and reference signal in equations (2) and (3) with the estimated $\mu_{k,m}$ and $X'_{k,m}$. Since the linear AEC module is implemented as a differentiable layer with no trainable parameters, gradients can flow through it and train the DNN parameters.

Note that the functionality of the linear AEC is generalized here. Rather than estimating the real acoustic echo path, the linear AEC in the proposed method serves to estimate a transfer function between the estimated nonlinear reference signal and the echo signal.

2.2.3. Loss function

The loss function for model training is calculated as the mean squared error (MSE) between $E_{k,m}$ and target signal $T_{k,m}$:

$$Loss = \text{MSE}(E_{k,m}, T_{k,m}) \tag{6}$$

The proposed method can be trained to achieve echo removal only or joint echo and noise reduction by using different target signals:

- $T_{k,m} = S_{k,m} + N_{k,m}$: A model trained using this target signal focuses on echo removal without noise reduction (NR), and the estimated $\hat{D}_{k,m}$ approximates echo signal. We denote this model as "Proposed (no NR)".
- $T_{k,m} = S_{k,m}$: A model trained this way achieves joint echo and noise removal, denoted as "Proposed". In this case, the estimated reference signal $X'_{k,m}$ contains noise information in it and the corresponding $\hat{D}_{k,m}$ approximates a mixture of echo and background noise. Hence the final output is an estimate of near-end speech with echo and noise jointly removed from the microphone signal.

2.3. Model training and feasibility analysis

It is worth noting that we have no ground truth for $\mu_{k,m}$ and $X'_{k,m}$ to directly guide the model training. To ensure the effectiveness of the proposed method, the linear AEC module is forced to use the outputs of the DNN module as step size and reference signal to minimize the error signal during training. Through the training of the entire model, the DNN module outputs can be interpreted as a step size and nonlinear reference signal for echo removal. During the inference stage, the parameters of DNN are fixed while the linear AEC is updating its filter coefficients adaptively using the estimated step size and reference signal.

The nonlinear distortions are addressed by using a DNN estimated nonlinear reference signal $X'_{k,m}$ in the linear AEC. By training a DNN to design appropriate time-frequency dependent step size $\mu_{k,m}$, the linear AEC is equipped to model echo path variations.

- From a signal processing perspective, the proposed system can be seen as an adaptive AEC method with its reference signal and step size estimated by a DNN module.
- From a deep learning perspective, the linear AEC module works as a non-trainable layer within a DNN. Integrating this interpretable and more constrained linear AEC elements into the more general and expressive DNN encodes structural knowledge in the model and make model training easier [20].

3. EXPERIMENTS

3.1. Experimental setup

We generate an in house dataset to carry out the experiments and focus on situations with echo path variations in low SER, and low SNR conditions. To record echoes in situations with continuous echo path changes, we use a mobile recording platform and fix a microphone and a loudspeaker on it. Randomly selected music and television/radio files from Spotify, Pandora and Amazon Music are used as far-end signals and played out of the loudspeaker. Since the echo signal are acoustically captured signals, the recordings should have nonlinear distortions introduced by the amplifier and loudspeaker. For generating training dataset, the echoes are recorded in both stationary and mobile scenarios. In the stationary recording session, we place the platform at a random position in the room. As for the mobile case, the platform moves forward and backward continuously while recording the echoes. A total of 260 minutes of recordings are collected with around one third of the recordings recorded in the stationary case and the remaining in the mobile case. We use 200 gender-balanced utterances from TIMIT dataset [27] as the near-end speech. To achieve a noise-independent model, 10000 noises from a sound effect library (http://www.sound-ideas.com) are used for training. Babble noise from the Auditec CD (http://www.auditec.com) is used for testing. Note that the noise used for testing is untrained.

To generate a microphone signal, we randomly select a 10 second signal from the recordings as an echo signal. A clean utterance is padded to the same length and added to the microphone signal at an SER level randomly selected from [-30, 0] dB. A random section of noise is then added to the mixture at an SNR level randomly selected from [-5, 5] dB. In total 5000 microphone signals are generated for training. For testing, we generate 100 microphone signals for each test case using untrained near-end speech, echo, and noise.

For implementation, the signals, sampled at 16 kHz, are windowed into 20 ms frames with a 10-ms overlap between consecutive frames. Then a 320-point STFT is applied to each frame to extract the spectrum. The length of the adaptive filter in frequency-domain NLMS is set to $L = 10$. AMSGrad optimizer [28] and MSE loss are used to train the model. All the networks are trained for 30 epochs with a learning rate of 0.001.

Echo return loss enhancement (ERLE) is used to measure single-talk performance. Perceptual evaluation of speech quality (PESQ) [29] and signal-to-distortion ratio (SDR) are used to evaluate double-talk voice quality. ERLE and SDR are defined as:

$$\text{ERLE} = 10 \log 10 \left[\sum_t y_t^2 / \sum_t e_t^2 \right] \quad (7)$$

$$\text{SDR} = 10 \log 10 \left[\sum_t s_t^2 / \sum_t (s_t - e_t)^2 \right] \quad (8)$$

Table 1. Performance in the presence of double-talk with different SER levels.

SER = -20 dB	PESQ	SDR	ERLE (dB)
Unprocessed	1.28 $\pm$ 0.28	-20.00 $\pm$ 4.97	-
NLMS	1.74 $\pm$ 0.30	-8.55 $\pm$ 2.41	11.58 $\pm$ 2.70
DNN-AEC	1.65 $\pm$ 0.31	1.31 $\pm$ 1.36	46.27 $\pm$ 4.82
DNN-AES	1.68 $\pm$ 0.30	-0.52 $\pm$ 1.42	47.89 $\pm$ 4.52
Proposed	**2.00 $\pm$ 0.28**	**3.89 $\pm$ 1.47**	**52.09 $\pm$ 3.48**
SER = -10 dB	**PESQ**	**SDR**	**ERLE (dB)**
Unprocessed	1.88 $\pm$ 0.22	-10.00 $\pm$ 9.39	-
NLMS	2.32 $\pm$ 0.23	-4.21 $\pm$ 1.47	11.45 $\pm$ 3.10
DNN-AEC	2.30 $\pm$ 0.24	4.72 $\pm$ 1.50	45.79 $\pm$ 4.20
DNN-AES	2.28 $\pm$ 0.22	3.76 $\pm$ 1.63	48.86 $\pm$ 4.34
Proposed	**2.59 $\pm$ 0.21**	**7.79 $\pm$ 1.96**	**51.50 $\pm$ 3.29**

3.2. Performance in double-talk only situations

We first evaluate the performance of the proposed method in situations with double-talk only (without background noise) and compare it with other DNN based methods. Comparison results are presented as mean $\pm$ std and are shown in Table 1. DNN-AEC refers to the fully DNN baseline method that estimates echo signal and then subtracts it from the microphone signal for echo removal [30]. DNN-AES denotes the fully DNN method that directly estimates near-end speech from the microphone signal [15, 16]. For a fair comparison, the DNN structure used in DNN-AEC and DNN-AES is the same as that used in the proposed method. The NLMS algorithm used for comparison is the same as the linear AEC module in the proposed method and a DTD [17] is employed in this NLMS for handling double-talk. It can be seen from the table that all deep learning based methods outperform traditional NLMS algorithm in terms of ERLE. The proposed method achieves better speech quality and echo removal compared to other deep learning based methods.

3.3. Performance in double-talk and background noise

This part studies the performance of the proposed method in the presence of double talk and background noise. Besides DNN-AEC and DNN-AES, DNN based residual echo suppression (RES), denoted as DNN-RES, is utilized as another comparison method. DNN-RES [10, 31] is a popular method for joint echo and noise reduction where DNN is utilized to further suppress residual echo and noise at the output of a traditional AEC algorithm. For comparison purpose, the traditional AEC algorithm and the DNN structure used in the DNN-RES are the same as those used in the proposed method. Table 2 shows the comparison results. In general, NLMS, DNN-AEC, and Proposed (no NR) focus on echo removal without handling noise reduction while the other three methods are used for joint echo and noise removal. It can be seen that the proposed method consistently outperforms other methods for the tasks of echo removal only and joint echo and noise removal.

Spectrograms of a test sample are given in Fig. 3. It is seen that DNN-AES, DNN-RES and the proposed method are capable of removing echo and noise jointly. The output of the proposed method approximates the target near-end speech and has less residual echo and noise during double-talk periods compared to other methods.

3.4. Analysis of the estimated nonlinear reference and step size

In adaptive filter based AEC algorithms, the reference signal should be highly correlated with the microphone signal for efficient adaptation. The ideal reference signal is a microphone signal without any near-end signal and echo removal can be achieved by directly

Table 2. Performance in the presence of double-talk and untrained babble noise with different SNR levels. The SER level for each test sample is randomly selected from [-30,0] dB.

Babble noise, SNR = 0 dB	PESQ	SDR	ERLE (dB)
Unprocessed	1.44 ± 0.45	-14.99 ± 7.10	-
NLMS	1.74 ± 0.31	-7.58 ± 3.01	8.62 ± 3.32
DNN-AEC	1.88 ± 0.27	-0.10 ± 0.97	16.66 ± 7.75
Proposed (no NR)	1.94 ± 0.24	0.31 ± 0.56	17.29 ± 8.37
DNN-AES	1.75 ± 0.41	0.63 ± 2.84	36.35 ± 7.68
DNN-RES	1.78 ± 0.42	-0.13 ± 0.89	46.13 ± 12.34
Proposed	**2.01 ± 0.32**	**3.84 ± 1.78**	**47.69 ± 6.80**
Babble noise, SNR = 5 dB	PESQ	SDR	ERLE (dB)
Unprocessed	1.52 ± 0.46	-15.65 ± 6.99	-
NLMS	1.87 ± 0.36	-7.16 ± 3.45	10.52 ± 2.97
DNN-AEC	2.02 ± 0.33	1.82 ± 2.08	21.09 ± 6.99
Proposed (no NR)	2.09 ± 0.29	3.02 ± 1.59	22.97 ± 7.79
DNN-AES	1.83 ± 0.40	0.71 ± 3.54	41.41 ± 4.05
DNN-RES	1.88 ± 0.48	-0.13 ± 0.68	52.35 ± 8.59
Proposed	**2.12 ± 0.35**	**4.70 ± 2.66**	**52.68 ± 4.90**

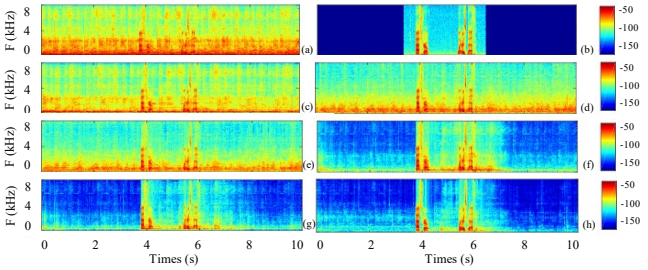

Fig. 3. Spectrograms of a test sample with -10 dB SER and babble noise at 5 dB SNR: (a) microphone signal, (b) target near-end speech, and enhanced speech signals of (c) NLMS, (d) DNN-AEC, (e) Proposed (no NR), (f) DNN-AES, (g) DNN-RES, (h) Proposed.

subtracting the reference signal from the microphone signal. The step size should be small in the presence of interferences to avoid divergence. For situations with echo path variations, the filter should adapt quickly if the echo path changes fast, and vice-versa.

Figure 4 plots the outputs of the DNN module, $M_{k,m}$ and $\mu_{k,m}$, in different situations, where the spectrograms of microphone signal and target near-end speech are provided in Fig. 4 (a) and (b) to show the activities of near-end speech. While Fig. 4 (c) and (d) are the outputs in situations without background noise, the remaining plots are obtained in situations with babble noise. Figure 4 (c) illustrates that the estimated reference signal approximates the microphone signal with near-end speech suppressed from it, which is effective for echo removal. The step size values shown in Fig. 4 (d) tend to be very small during double-talk periods to avoid divergence and then increase immediately following the double-talk section to speed up the convergence of the algorithm. Moreover, the step size values at higher frequencies are relatively larger than those at low frequencies. This is because the power of interference is larger at low frequencies, and also in the mobile case, the acoustic channel changes faster at higher frequencies due to the Doppler effect [32]. The proposed system with "no NR" only focuses on echo removal and it classifies the whole utterance as double-talk due to the presence of strong babble noise. Therefore, the estimated reference signal approximates the echo components in the mixture and the values of step size are close to zero, as shown in Fig. 4 (e) and (f). The outputs for joint echo and noise removal are shown in Fig. 4 (g) and (h), where the estimated reference signal and step size show a similar trend to the no-noise

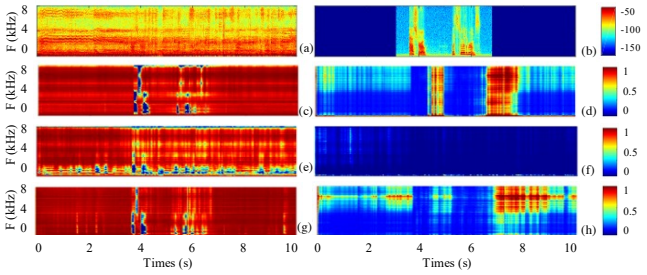

Fig. 4. DNN outputs of a test sample with -20 dB SER and babble noise at 0 dB SNR: (a) $Y_{k,m}$, (b) $S_{k,m}$, (c) and (d) are, respectively, $M_{k,m}$ and $\mu_{k,m}$ of Proposed in no-noise case, (e) $M_{k,m}$ of Proposed (no NR), (f) $\mu_{k,m}$ of Proposed (no NR), (g) $M_{k,m}$ of Proposed, and (h) $\mu_{k,m}$ of Proposed.

Table 3. Performance comparison between DNN-AES and Proposed using different DNN structures with babble noise at 0 dB SNR.

	DNN-AES			Proposed		
DNN	DNN-L	DNN-M	DNN-S	DNN-L	DNN-M	DNN-S
# Parameter	3.2 M	1.2 M	0.3 M	3.2 M	1.2 M	0.3 M
PESQ	1.75	1.72	1.72	2.01	1.98	1.96
SDR	0.63	0.69	0.91	3.84	3.64	3.42
ERLE	36.35	35.32	32.20	45.90	43.96	39.91

case in Fig. 4 (c) and (d). The reference signal, in this case, has noise information in it and the output of the whole system achieves joint echo and noise suppression.

3.5. Comparison results using DNNs with different sizes

We further compare the performance of the proposed method and the DNN-AES method that trained using the same network structure. We gradually shrink the model size and train the two methods using three different DNNs. DNN-L is the large model used in previous experiments. DNN-M is a similar LSTM with 3 hidden layers and 200 units in each layer. DNN-S is a small LSTM with 2 hidden layers and 100 units in each layer. The number of parameters and experimental results are given in Table 3. We note that the performance of both methods reduces gradually by shrinking the model size while the proposed method consistently outperforms DNN-AES. The proposed method trained with a smaller model (DNN-S) achieves better performance than DNN-AES trained with a larger model (DNN-L).

4. CONCLUSIONS

In this paper, we have proposed a deep adaptive filtering based AEC technique to leverage the advantages of traditional and deep learning based AEC methods. The proposed approach enables the integration of traditional AEC elements within deep learning methods to achieve adaptiveness while retaining the power of neural networks. A DNN model is employed to estimate a control parameter and a nonlinear reference signal, which are then used by a differentiable linear AEC module for echo cancellation. Systematic evaluations with ERLE, PESQ, and SDR show the effectiveness and robustness of the proposed method for echo and noise removal in low SER/SNR conditions with continuous echo path variations. In addition, the proposed method trained using a smaller model achieves better performance compared to large fully-DNN based models.

5. REFERENCES

[1] M. M. Sondhi, "An adaptive echo canceller," *Bell System technical journal*, vol. 46, no. 3, pp. 497–511, 1967.

[2] 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.

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

[4] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic press library in signal processing*, vol. 4, pp. 807–877. Elsevier, 2014.

[5] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on speech and audio processing*, vol. 8, no. 5, pp. 508–518, 2000.

[6] S. L. Gay, "The fast affine projection algorithm," in *Acoustic signal processing for telecommunication*, pp. 23–45. Springer, 2000.

[7] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters–an overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, 2000.

[8] A. N. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. IEEE, 1995, pp. 103–106.

[9] M. I Mossi, N. W. D. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *2010 ICASSP*. IEEE, 2010, pp. 313–316.

[10] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in *Proceedings of INTERSPEECH*, 2015.

[11] J. Malek and Z. Koldovskỳ, "Hammerstein model-based nonlinear echo cancelation using a cascade of neural network and adaptive linear filter," in *2016 IWAENC*. IEEE, 2016, pp. 1–5.

[12] M. M. Halimeh, C. Huemmer, and W. Kellermann, "A neural network-based nonlinear acoustic echo canceller," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1827–1831, 2019.

[13] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *2021 ICASSP*. IEEE, 2021, pp. 151–155.

[14] H. Zhang and D. L. Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation," *Proceedings of INTERSPEECH*, pp. 1139–1143, 2021.

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

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

[17] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Processing*, vol. 86, no. 6, pp. 1124–1139, 2006.

[18] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *1997 ICASSP*. IEEE, 1997, vol. 1, pp. 307–310.

[19] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[20] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," *arXiv preprint arXiv:2001.04643*, 2020.

[21] M. A. M. Ramírez, O. Wang, P. Smaragdis, and N. J. Bryan, "Differentiable signal processing with black-box audio effects," in *2021 ICASSP*. IEEE, 2021, pp. 66–70.

[22] B. Kuznetsov, J. D. Parker, and F. Esqueda, "Differentiable IIR filters for machine learning applications," in *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*, 2020, pp. 297–303.

[23] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *2019 ICASSP*. IEEE, 2019, pp. 900–904.

[24] A. Ivry, I. Cohen, and B Berdugo, "Nonlinear Acoustic Echo Cancellation with Deep Learning," in *Proceedings of INTERSPEECH*, 2021, pp. 4773–4777.

[25] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized NLMS algorithms for acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 97, 2015.

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

[27] 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.

[28] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," 2018.

[29] 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 *2001 ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[30] J. Franzen, E. Seidel, and T. Fingscheidt, "Aec in a netshell: on target and topology choices for fcrn acoustic echo cancellation," in *2021 ICASSP*. IEEE, 2021, pp. 156–160.

[31] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation," in *2021 ICASSP*. IEEE, 2021, pp. 121–125.

[32] J. A. Gómez-Tejedor, J. C. Castro-Palacio, and J. A. Monsoriu, "The acoustic Doppler effect applied to the study of linear motions," *European Journal of Physics*, vol. 35, no. 2, pp. 025006, 2014.