Acoustic Echo Cancellation and Residual Echo Suppression

Prasad Kamath
University of California San Diego
La Jolla, California
hmojtahed@ucsd.edu

Hamed Mojtahed
University of California San Diego
La Jolla, California
pkamath@ucsd.edu

September 10, 2022

Abstract

Acoustic Echo Cancellation (AEC) is a topic that has gained attention after increasing necessity of telecommunication and video-conferencing. Acoustic echo is resulted by transmission of reflection of audio communicated from Far-end speaker to original speaker. The resulting echo is undesirable and often source of distraction in teleconferencing. Adaptive filtering is the main technique that used in many AEC applications, however, due to the existence of sources of non-linearity in the system it is unable to completely remove the echo and leaves residues in the output audio. Due to the fact that the Deep Neural Network are great tool for adaptation to non-linearities, in this project we have proposed a two-stage techniques that combines the use of adaptive filters for cancellation of linear part of echo and Neural Networks for suppression of non-linear part of residue from the adaptive filter output. In this project, pre-filter and post-filter approaches were investigated, and based on the findings the best echo cancellation was obtained using NLMS post-filtering approach.

1 Introduction

With development and prevalence of telecommunication and video conferencing especially after COVID-19 pandemic the need for clear long distance communication over the Internet has increased. Acoustic Echo is one of the problems that commonly occurs during teleconferencing. Echo is an aversive signal that creates a significant barrier in mutual understanding in communication systems involving voice transmission. Echo signal reflects from boundary surfaces, and other nearby objects, and is generated from the coupling between loudspeakers and microphones of Near-end speaker and relays back the speech back to Far-end. This coupling has adverse effect on the quality of communication systems and the functionality of automated speech recognition in smart speakers. Commonly acoustic echo cancellation methods use adaptive filtering algorithms to identify the impulse response between the loudspeaker and the microphone. To ensure both fast convergence and low computing load, the least mean square (LMS) methods such as LMS, and NLMS are commonly used. When existence of nonlinear distortion in the acoustic echo route is significant, the performance of adaptive approaches suffers significantly. To further reduce the echo, a residual echo suppression module is usually necessary. Due to great ability of deep neutral network to describe nonlinear systems, they have been integrated into residual echo suppression (RES). The typical way to perform RES is by estimating the spectral amplitude of the residual echo using the Far-end signal, and residual signal acquired using adaptive filtering techniques.

1.1 Non-linearities

Nonlinear distortion has become more problematic as consumer devices with smaller components, such as microphones and loudspeakers are driven at higher amplitudes [1]. The performance of linear AEC algorithms are significantly suffered from the distortions from non-linearities. The non-linearities are divided in two categories of with and without memory. Nonlinearities with memory are introduced by mechanical vibrations conveyed from the loudspeaker to the microphone via the device's enclosure [2]. Operating a loud speaker at high amplitude will result in non-linearities with memory [3]. In other hand overdriving amplifiers will result in non-linearites without memory [3]. Resolving the non-linearities with memory are typically more difficult.

1.2 Dataset

For the purpose of AEC experiment in this project two dataset are utilized which are: 1) International Conference on Acoustics, Speech, & Signal Processing (ICASSP) 2022 dataset. 2) Device and Produced Speech (DAPS) dataset.

1.2.1 ICASSP

ICASSP is an AEC challenge that is organized by Microsoft annually. For the purpose of this research we procured the latest ICASSP dataset contest held in 2022. This dataset contains both crowd-sourced and synthetically generated data, with each collection having recordings of Near-end speech, microphone signal, Far-end speech, and echo signals. Both single and double talk instances are included in the dataset [4].

1.2.2 DAPS

DAPS dataset contains alignment of professionally produced studio speech recordings with recordings of the same speech on common consumer devices such as tablet and smartphone. In details there are 15 versions of each audio, which 3 versions are recorded professionally, and 12 versions recorded using consumer device or real-world combinations [5].

2 Methods

In this project we utilized a hybrid approach which involves application of adaptive filter for cancellation of linear part of echo, along with suppression of residue using neural network.

2.1 Pre-filtering and post-filtering

Depending on position of neural network with respect to adaptive filter output we divide the methods used in this project into two sub-categories of pre-filter and post-filter. In pre-filtering approach the neural network is applied to the input audio speech first to remove non-linearites first and then is passed to adaptive filter for removing the linear of the distortion. In contrast in the post-filter approach, which is a popular approach in literature, the adaptive filter first removes the linear part of the distortion from the input audio and then the residue is suppressed by passing the resultant audio output into a neural network.

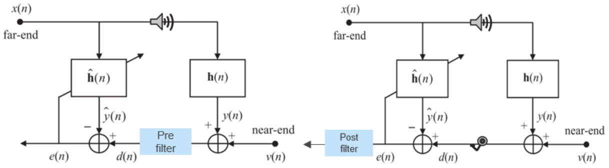

Figure 1: Pre-filtering (left) and post-filtering (right) diagrams. Neural network is shown with blue box.

2.2 Adaptive Filtering

Filters are signal processing tools that by processing a given signal manipulates its information.Adaptive filtering is a solution in signal processing used when there is no fixed specification to the signal processing problem or when time-invariant filter will not lead to single general solution for the problem. An adaptive filter is a time-varying filter that adjust its parameters based on the objective of the problem. Adaptive filters are evaluated by their convergence rate and complexity. One application of adaptive is in echo cancellation in long distance communications with audio transmission. In AEC problems involving adaptive filtering the echo path is estimated by an adaptive filter. This estimate is then subtracted from the microphone signal to produce clean signal, and transmitted to Far-end. The diagram of adaptive filtering used in AEC problem is shown in Figure 2.

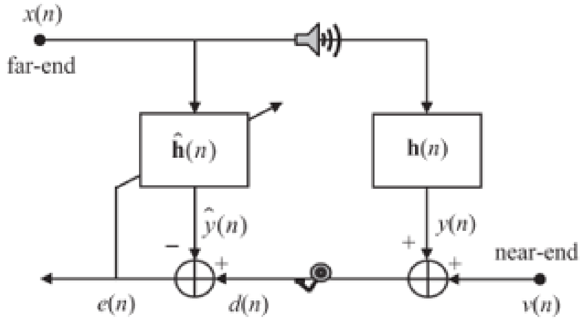

Figure 2: Diagram of acoustic echo canceller using an adaptive filter

The Far-end speech signal is denoted by $x(n)$, which is relayed back as echo. The $v(n)$ is the Near-end speech signal, and $d(n)$ is the microphone input that contains mixture of the Near-end speech and the echo signal $y(n)$ given by:

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

The estimated echo is shown by $\hat{y}(n)$ and is obtained by:

$$\hat{y}(n) = x(n) * \hat{h}(n) \tag{2}$$

The desired signal is denoted by $d(n)$ and obtained by:

$$d(n) = v(n) + h(n) * x(n) \tag{3}$$

Finally, the error signal that includes the clean Near-end speech is obtained by:

$$e(n) = d(n) - \hat{y}(n) \tag{4}$$

2.2.1 Least mean squares filter

Least mean square (LMS) is a widely use adaptive filter algorithm that utilizes stochastic gradient descent [6]. It was first proposed by Widrow and Hoff in 1960. The filter tap coefficients in LMS algorithm are updated according to:

$$w(n+1) = w(n) + 2\mu e(n)x(n) \tag{5}$$

With $x(n)$ denoting the input signal having values shown in Equation 6 and parameter μ denoting the step size.

$$x(n) = [x(n), x(n-1), x(n-2), ..., x(n-N+1)] \tag{6}$$

The output $y(n)$ and error signal $e(n)$ are obtained by:

$$y(n) = w^T(n)x(n) \tag{7}$$

$$e(n) = y(n) - w^T(n)x(n) \tag{8}$$

2.2.2 Normalized least mean square filter

The normalized LMS (NLMS) algorithm is a variant of the LMS method that takes into account signal level variations at the filter input [7]. The step-size parameter is normalized using this method. As the result the algorithm ensure stable and fast convergence. In NLMS algorithm the step size is updated by:

$$\mu(n) = \frac{1}{2x^T(n)x(n)} \tag{9}$$

And the weights are update according to the recursive formula given by:

$$w(n+1) = w(n) + \frac{1}{2x^T(n)x(n)} e(n)x(n) \tag{10}$$

The selection of the step-size in Equation 9 is desirable as it is inversely proportional to input signal energy, and therefore matching to misadjustment. NLMS algorithm is widely used for AEC due to its simplicity and faster convergence. The adaptive step size allows the algorithm to accommodate for variations in the input signal's amplitude, which makes it excellent for feedback path estimation [8].

2.3 Deep Neural Network

2.3.1 Data Pre-processing

For the purpose of the experiments in this project the data was first preporcessed by the pipeline shown in Figure 3. The first step taken in processing is to normalize waveform audio files, then spectrogram of each individual audio file is obtained by computation of short-time Fourier transform (STFT) using predetermined fft points, and Hann window. Next is to split the spectrograms into smaller size spectrogram representing the 300ms delay duration in ehco. The prepossessed data was then divided into test, train and validation.

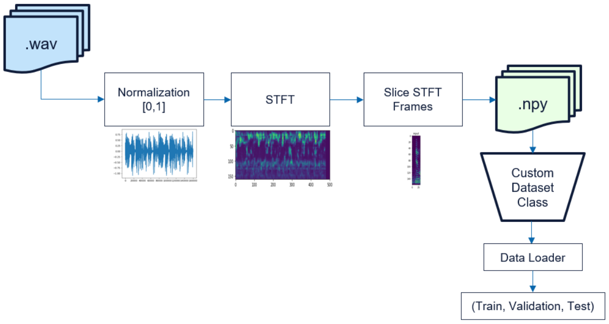

Figure 3: Flow chart showing data preprosssing pipeline

2.3.2 U-Net

The U-Net architecture was proposed by Ronneberger, et. all in 2015 for biomedical Image Segmentation. The U-Net architecture is constructed of a contraction path and an expansion path. The contracting path is consist of convolutions layers applied repeatedly, each followed by a rectified linear unit (ReLU) and a max pooling operation. As the network goes deeper in the the contraction path the number of features increase. The expansive pathway uses a series of up-convolutions and concatenations with high-resolution features from the contracting path to merge feature and spatial information [9]. Because of the U-Net ability to recreate the finest details of audio signals, U-Net networks have previously been used in the audio processing sector for source separation problems. Rodrıguez et. all used U-NET architecture for AEC application by optimizing the hyper-parameters and reducing the number of parameters to meet a latency limit of 40 ms.

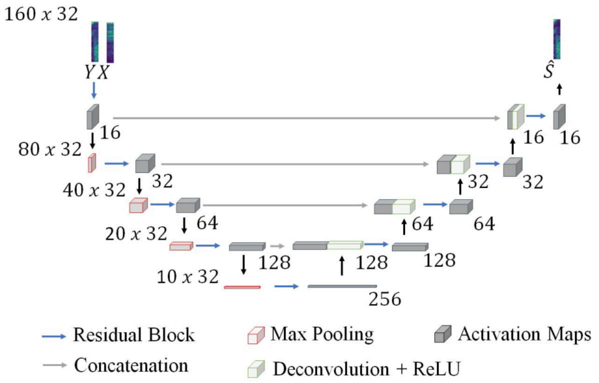

Figure 4: U-Net architecture for acoustic echo cancellation

2.3.3 LSTM

Long short-term memory (LSTM) is one of the solution proposed for the vanishing gradient problem. LSTM cell is constructed from an input gate, an output gate, and a forget. The three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time intervals. Recurrent Neural Networks with Long Short-Term Memory (LSTM) have memory and can use cached hidden states to estimate network parameters in future time steps. LSTM model has the ability to simulate non-linearities with memory, making it more robust than fully convolutional approach. The architecture of the LSTM model [10] is shown in Figure 5

Metric	NLMS (1024) taps	NLMS (Pre-whiten)	LSTM	LSTM (Post-filter)
Mean ERLE (db)	13.72	18.98	18.99	27.26
Max ERLE (db)	29.78	35.23	38.84	54.54
Mean SDR (db)	1.74	4.49	4.79	7.07
Max SDR (db)	10.90	12.03	11.16	13.08

Table 1: Comparison of the Mean/Max ERLE and SDR for the models

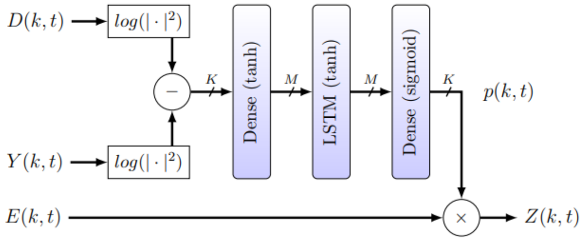

Figure 5: LSTM architecture for acoustic echo cancellation

3 Experiments and Results

3.1 Objective Metrics

The two main metrics used for characterization of the performance of an AEC algorithm are Echo Return Loss Enhancement (ERLE), and Speech to Speech Distortion Ration (SSDR).

3.1.1 ERLE

ERLE is a measure of the residual echo instantaneous power present in near end speech to be transmitted to far end instantaneous power during single talk, and is obtained by:

$$ERLE = 10log_{10}\frac{P_{d(n)}}{P_{e(n)}} \tag{11}$$

ERLE is measured in decibels (dB). The higher ERLE value is indicative of better echo canceller performance. ERLE as a function of time index provides information about the canceller's convergence behavior [11].

3.1.2 SSDR

SSDR measures the distortion in near end speech s(n) introduced by the AEC algorithm during double talk [11].

$$SSDR = 10\log 10\frac{\sum_{n=0}^{N_{DT}-1}|s(n)|^2}{\sum_{n=0}^{N_{DT}-1}|s(n)-e(n)|^2} \tag{12}$$

3.2 Pre-filter: UNET

The UNET-based pre-filter sought to eliminate non-linear features of device-produced audio from the DAPS dataset, but the spectrum was smoothed due to repetitive dimensionality reduction due to multiple max-pool layers, resulting in speech formant information loss. As a result of the distorted reconstructed speech, this design is unsuited for pre-filtering.

3.3 Pre-filter: LSTM

The results obtained utilizing the LSTM model in conjunction with the NLMS-based architecture were much better than those obtained using UNET and NLMS. This is owing to the LSTM network's capability to learn non-linearities with memory, which are common in audio produced by natural devices.

4 Conclusion

In this project we have elucidated the impact of non-linearities on adaptive filter performance in the Acoustic Echo Cancellation application. We found out that fully convolutional networks are not well fitted for modeling non-linearities because non-linearities change over time; hence, LSTM-based architectures that have memory outperform fully convolutional networks. We investigate a two-stage method which uses an adaptive filter as a pre-filter and a DNN as a post-filter. The DNN-only strategy yields good results. Although the DNN pre-filter technique outperforms the baseline NLMS, it distorts the Near-end speech retrieved. Based on the objective metrics post-filtering approach which cascade the NLMS output with DNN produced best echo removal effect.

References

[1] D. Morgan, J. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 686–696, 2001. 2

[2] L. Pfeifenberger and F. Pernkopf, "Nonlinear residual echo suppression using a recurrent neural network." in *Interspeech*, 2020, pp. 3950–3954. 2

[3] 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. 2

[4] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sørensen, and R. Aichner, "Icassp 2022 acoustic echo cancellation challenge," 2022. [Online]. Available: https://arxiv.org/abs/2202.13290 2

[5] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2014. 2

[6] S. S. Haykin, *Adaptive filter theory.* Pearson, 2014. 3

[7] B. Farhang-Boroujeny, *Adaptive filters: theory and applications.* John Wiley & Sons, 2013. 3

[8] 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. 1–19, 2015. 3

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention.* Springer, 2015, pp. 234–241. 4

[10] R. C. Staudemeyer and E. R. Morris, "Understanding lstm–a tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv:1909.09586*, 2019. 4

[11] B. Fang, "A robust residual echo suppression algorithm even during double talk," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP).* IEEE, 2020, pp. 6–9. 5