Adaptive Speech Quality-Aware Complex Neural Network for Acoustic Echo Cancellation using Supervised Contrastive Learning

Anonymous submission to INTERSPEECH 2023

Abstract

This paper proposes an adaptive speech quality complex neural network for acoustic echo cancellation (AEC) in a multi-stage framework. The proposed model comprises a modularized neural network that focuses on feature extraction, acoustic separation, and mask optimization. To enhance performance, we adopt a contrastive learning framework during the pre-training stage, followed by fine-tuning using a complex neural network with novel speech quality aware loss functions. The proposed model is trained using 72 hours for pre-training and 72 hours for fine-tuning. Experimental results demonstrate that the proposed model outperforms the state-of-the-art methods. Our approach provides a novel framework for AEC by leveraging the benefits of both the contrastive learning framework and the adaptive speech quality aware complex neural network.

1. Introduction

Acoustic echoes are a common issue in audio communication, particularly when using speakerphones or in conference settings. In such situations, it is common for speakers to hear the echo of their own voices, which can be very annoying and distracting. Traditional methods to address this issue involve modeling the echo impulse response between the speaker and microphone as a finite impulse response (FIR) filter and adaptively adjusting it using normalized least mean square (NLMS) algorithm [1]. However, this method works well only in far-end single-only scenarios. In double talk scenarios where the near-end and far-end speech are present simultaneously, the adaptive filter may not converge, resulting in relatively large residual echoes [2].

Deep learning-based solutions have shown promising results for acoustic echo cancellation [3, 2, 4]. For example, [2] proposed a deep long short-term memory network (DRLTN) to remove echo from both the frequency and time domains. However, this architecture lacks consideration for phase information. To address this issue, [4] developed a complex encoder-decoder neural network using complex conv2D and LSTM to incorporate phase information. Nevertheless, complex neural networks can be computationally expensive, making them challenging to implement on hardware. In contrast, pseudo complex neural networks, such as DPCRNN [5], which take the real and imaginary values as two input channels, are more desirable.

Contrastive learning is an effective training method that learns robust representations in the supervised or unsupervised setting [6, 7]. It involves creating positive and negative pairs and training a target model to increase the similarity between positive pairs and decrease it for negative pairs. While contrastive learning for audio has been studied for vector representation, such as [8], which proposed a self-supervised pre-training approach for learning a general-purpose representation of audio for classification-based sub-stream tasks, contrastive learning for regression-based tasks has not been fully explored. This paper proposes a new self-supervised contrastive learning approach for audio regression tasks and demonstrates its performance using acoustic echo cancellation as an example.

In this paper, we propose a complex neural network for acoustic echo cancellation, which is modularized to perform input feature extraction, speech separation, and output predicted speech mask enhancement. We also introduce a supervised contrastive learning loss for regression-based tasks and propose novel adaptive speech-aware loss functions to compensate for speech distortion and suppress residual echoes in the speech. Finally, we demonstrate the effectiveness of our proposed method using the ICASSP-2022 AEC dataset [9], which shows that the proposed loss achieves better performance. It improves the PESQ by around 0.5 and 0.75 compared to the ICASSP-2022 Microsoft competition baseline [9] and DTLN [2], respectively.

2. Methodology

2.1. Problem Formulation

Acoustic echo cancellation is a signal processing task that aims to remove the acoustic echo component from a mixed audio signal captured by a microphone. The mixed microphone signal, $y(n)$ can be modeled as the sum of near-end speech, acoustic echo, and near-end noise, i.e.,

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

where $s(n)$, $d(n)$, and $v(n)$ are the near-end speech, echo, and noise components, respectively, and n is the time index. The echo component $d(n)$ is the result of the convolution between far-end signal and the impulse response. This work can be reformulated as a noise suppression problem in the frequency domain, where the input mixed signal $Y(k, f)$ is transformed using a frequency mask $M(k, f)$ to obtain the enhanced signal, $M(k, f)|Y(k, f)|e^{j\phi_y(k,f)}$.

2.2. Contrastive Learning Pre-training Framework

Contrastive learning has gained much attention recently in computer vision [10, 11]. It is a technique that enables the model to learn the difference between data pairs and extract better image representation by maximizing the distance of similar data pairs. In the speech domain, recent studies have shown that contrastive learning can be used to pretrain an encoder to help downstream tasks achieve better results [12]. Motivated by these findings, we propose a contrastive learning pretraining framework for

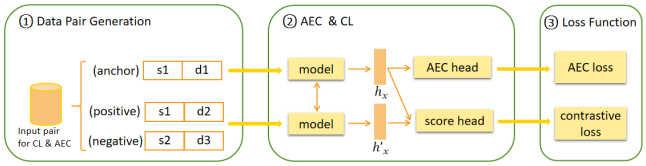


Figure 1: *Contrastive Learning Pre-training Framework.*

Acoustic Echo Cancellation (AEC) models to enhance their performance.

The proposed contrastive learning (CL) framework is shown in Figure 1 and consists of three parts: data pair generation, AEC complex neural network, and the CL loss function. In the data generation stage, we perform data augmentation to enable contrastive learning, creating an anchor audio and its corresponding positive and negative audio pairs. The positive audio pairs share the same near-end audio as the anchor audio, but have different far-end audio. The negative audio pairs are different from both the near-end and far-end audio of the anchor audio. By maximizing and minimizing the similarity between the positive and negative pairs, our model can learn to distinguish between near-end and far-end audio in the microphone signal. An additional CL score head controls the training processing of both AEC and CL. The anchor audio and contrastive audio share the same weight as the AEC model, with h_x and h'_x representing the features of the anchor and contrastive audio pair. The similarity between these features is calculated in the scoring head, with only the anchor feature used in the AEC head. Binary cross-entropy is used as as the loss function for the contrastive loss, which is expressed as follows:

$$\mathcal{L}_{CL} = -\log \frac{\exp(sim(x, x^+))}{\sum_{x^- \in X^-(x) \cup \{x^+\}} \exp(sim(x, x^-))} \quad (2)$$

where $sim(x, x^+)$ is the similarity score between the anchor feature and the positive feature at the frame level, and $sim(x, x^-)$ is the similarity score between the anchor feature and each negative feature.

2.3. AEC Fine-tuning Framework

As illustrated in Figure 2, our deep complex AEC network consists of three modules, namely, the feature extraction network, acoustic separation network, and mask optimization network. To incorporate both the phase and amplitude information of the signal, the input to the network is the stack of the complex form with real and imaginary parts of the microphone signal and far-end signals in the time-frequency domain. B, T, and 2N are representations of the dimensions of microphone and far-end signals, corresponding to batch size, number of audio frames, and the real and imaginary components of the audio signals after performing FFT transform.

2.3.1. Feature Extraction Network (FEN)

In FEN, a Convolutional Gated Recurrent Unit (CGRU) structure is utilized for extracting shallow feature representations of the signal. Specifically, the CGRU module comprises a one-dimensional convolution layer (Conv1D) and one GRU layer. The CGRU structure is designed to capture the characteristic information of the signal, while the Conv1D layer is responsible for capturing local contextual information. By adding a residual connection that concatenates the input with the output of the CGRU module, we obtain an extracted feature mask that is able to preserve the input signal information. The skip connection is shown to be advantageous in helping the CGRU module better understand the characteristic information of the signal.

2.3.2. Acoustic Separation Network (ASN)

ASN is designed to extract the near-end speech signal by applying two GRU layers on the input estimated feature mask, resulting in an output estimated near-end speech mask with half of the dimension size on the last dimension axis.

2.3.3. Mask Optimization Network (MON)

In MON, both the estimated near-end speech mask and the original microphone input signal are concatenated as inputs to generate a better near-end speech mask. By leveraging the microphone signal information, MON can differentiate between the desirable and undesirable signal components and therefore achieve more efficient suppression of the acoustic echo. The output of MON, estimated near-end remask, is obtained by adding its input, thereby enhancing the performance in complex acoustic scenarios. This approach allows the network to effectively leverage temporal dependencies in the input signal and gradually refine the output estimate.

2.4. Loss Functions Design

2.4.1. Mask Mean Square Error (MaskMSE) Loss

The acoustic separation network and the mask optimization network aim to estimate the near-end speech masks, denoted as $\hat{M}complex$, which can be used to obtain the separated near-end speech signal by multiplying the estimated mask and the microphone signal. The true masks, denoted as $Mcomplex$, can be generated using the given true label of the near-end speech $S(n)$ and the microphone input signal $Y(n)$.

To compare the estimated mask with the true mask, the Mask Mean Square Error (MaskMSE) Loss function is commonly used as the objective function. The MaskMSE Loss is defined as the mean square error between the estimated complex mask, $\hat{M}complex$, and the true mask, $Mcomplex$. The complex mask can be expressed in terms of its real and imaginary components, denoted as M_R and M_I, respectively:

$$M_{complex} = M_R + iM_I \quad (3)$$

The real and imaginary components of the complex mask can be calculated as follows:

$$M_R = \frac{S_R Y_R + S_I Y_I}{{Y_R}^2 + {Y_I}^2}, \quad M_I = \frac{S_I Y_R - S_R Y_I}{{Y_R}^2 + {Y_I}^2} \quad (4)$$

where SR and SI represent the real and imaginary parts of the near-end speech signal, $s(n)$, and YR and YI represent the real and imaginary parts of the microphone input signal, $y(n)$, respectively.

The MaskMSE Loss is computed as the squared difference between the true and estimated complex masks:

$$\mathcal{L}_M = |M_{complex} - \hat{M}_{complex}|^2 = |M_R - \hat{M}_R + i(M_I - \hat{M}_I)|^2 \quad (5)$$

2.4.2. Error Reduction Loss

The error signal $E(n)$ is defined as the difference between the microphone input signal $Y(n)$ and the true label near-end

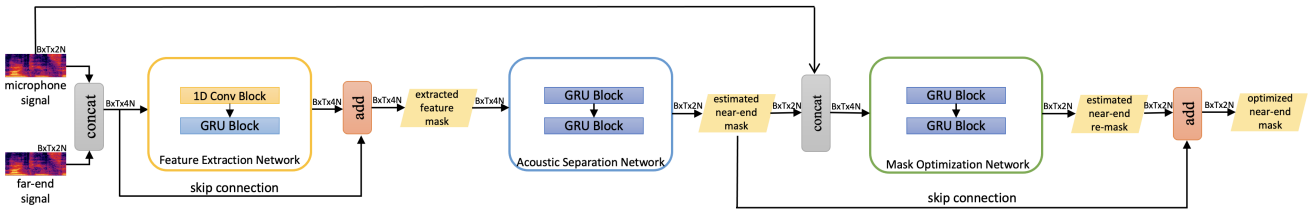
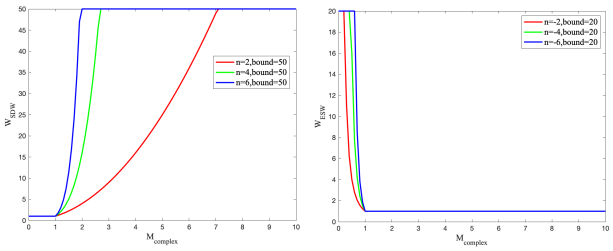


Figure 2: *AEC Fine-tuning Framework.*



(a) (b)

Figure 3: *(a) An example of different values of n on graph of W_{SDW} vs. $M_{complex}$ for Equation (8); and (b) an example of different values of n on graph of W_{ESW} vs. $M_{complex}$ for Equation (10).*

speech signal $S(n)$. Ideally, the same estimated complex mask should reduce all error signals to zero. This error can be minimized by multiplying the modulus of the complex estimated mask and the modulus of the error signal, as follows:

$$E_{complex} = Y_{complex} - S_{complex} \tag{6}$$

$$\mathcal{L}_E = |\hat{M}_{complex}||E_{complex}| \tag{7}$$

2.4.3. Speech Distortion Compensation Weighted (SDW) Loss

The Speech Distortion Compensation Weighted (SDW) Loss is an adaptive weight designed to introduce non-linear weights on the MaskMSE, in order to compensate for speech distortions in the estimated speech. The SDW weight is calculated based on a speech quality-aware criterion that considers the dominant contribution of clean speech over noisy speech at certain frequency bins. More specifically, the speech distortion compensation weight is defined as:

$$\mathcal{W}_{SDW} = min(max(1, |M_{complex}|)^n, bound) \tag{8}$$

where $|M_{complex}|$ represents the magnitude of the complex-valued mask, n and $bound$ are positive numbers greater than 1, which control the shape and range of the SDW function.

The SDW weight is then used to modify the MaskMSE loss, denoted as L_M, as follows:

$$\mathcal{L}_{SDW} = W_{SDW} L_M \tag{9}$$

The SDW loss enables the model to effectively handle speech distortions that arise due to factors such as reverberation and noise, while maintaining a good speech quality. The choice of the SDW parameters n and $bound$ can have a significant impact on the performance of the SDW loss and should be carefully tuned based on the specific task and dataset.

2.4.4. Echo Suppression Compensation Weighted (ESW) Loss

The Echo Suppression Compensation Weighted (ESW) Loss is an adaptive speech quality-aware weight that introduces non-linear weights on MaskMSE, in order to suppress the echoes in the estimated speech. The ESW is applied at certain frequency bins, where the true mask $M_{complex}$ is less than 1, indicating that the noisy speech dominates over the clean speech. The echo suppression compensation weight applies a weighted power function on MaskMSE, which helps to suppress the echoes of the estimated speech.

$$\mathcal{W}_{ESW} = min(min(1, |M_{complex}|)^n, bound) \tag{10}$$

where n is a negative number less than 1, and $bound$ is a positive number.

$$\mathcal{L}_{ESW} = W_{ESW} L_M \tag{11}$$

The ESW loss effectively suppresses the echoes in the estimated speech, resulting in a clearer and more natural sounding speech. The ESW loss, together with the SDW loss, enables the proposed AEC system to achieve high-quality speech enhancement, even in the presence of complex acoustic environments with strong echoes and noise.

The total loss is calculated as the weighted sum of the AEC loss and contrastive loss, as follows: $\mathcal{L}_{ALL} = \mathcal{L}_{AEC} + \alpha\mathcal{L}_{CL}$. where $L_{AEC} = L_M + L_E + L_{SDW} + L_{ESW}$, α is the weight parameter that controls the contribution of the contrastive loss, set to 1 in the CL Pre-training stage and 0 in the AEC Fine-tuning stage.

3. Experiments

3.1. Dataset

The proposed AEC system was trained and validated on a dataset consisting of 90 hours of audio, of which 72 hours were used for training and the remaining 18 hours for validation. The clean speech used for generating the far-end and near-end audio was obtained from multilingual data provided by Reddy et al. [13]. The dataset comprises speech in French, German, Italian, Mandarin, English and others. To further enhance the quality of the audio, the DTLN model [14] was used to remove possible noise before generating the far-end and near-end audio.

To create the echo, far-end and near-end signals, we follow exactly the same data process as [2] and keep 20% of the near-end audio discarded to create the near-end only scenario. All signals were normalized to (-25, 0) dB and segmented into 4-second segments before being passed into the network. The proposed dataset provides a diverse range of acoustic environments and enables the evaluation of the AEC system under different challenging conditions.

Table 1: *Performance comparison of different methods at different noise level (SNR) input, in terms of PESQ and ESTOI. None stands for the input audio without any processing.*

Noise	Model	PESQ					ESTOI				
		10db	15db	20db	25db	30db	10db	15db	20db	25db	30db
Room Noise	None	2.265	2.287	2.310	2.324	2.335	0.886	0.887	0.889	0.890	0.891
	DTLN [2]	2.766	2.769	2.777	2.777	2.775	0.902	0.902	0.902	0.903	0.990
	Baseline[9]	3.025	3.028	3.030	3.032	3.031	0.919	0.919	0.920	0.921	0.921
	CL Pretraining	3.117	3.116	3.117	3.121	3.120	0.931	0.931	0.932	0.932	0.932
	AEC Finetuning	3.163	3.234	3.352	3.425	3.462	0.930	0.938	0.942	0.945	0.946
Music Noise	None	1.516	1.551	2.325	2.338	2.347	0.673	0.679	0.889	0.891	0.891
	DTLN [2]	1.841	1.860	2.763	2.763	2.763	0.732	0.737	0.902	0.901	0.902
	Baseline [9]	1.915	1.937	3.019	3.020	3.021	0.750	0.757	0.919	0.920	0.920
	CL Pretraining	1.981	2.007	3.099	3.103	3.102	0.770	0.777	0.930	0.930	0.930
	AEC Finetuning	3.374	3.463	3.498	3.510	3.511	0.822	0.826	0.828	0.829	0.940

3.2. Experiment Set-up

The input of our architecture, denoted as $y(n)$, is sampled at a rate of 16 kHz and taken in frames of 512 points (32 ms) with 8 ms overlapping consecutive frames. After the Fast Fourier transform (FFT), this leads to a 513-dimensional spectral feature in each frame, considering both real and imaginary values. The total input features are 1026, including both microphone signals and far-end (reference) signals. The overall neural network architecture consists of three components, namely FEN, ASN, and MON. The network layers are a combination of Conv1D with a filter number of 128 and GRU with a number of units of 512.

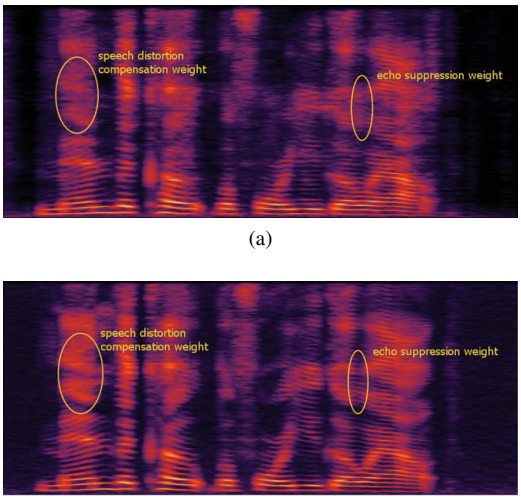
To train the proposed AEC system, a two-stage training strategy is employed. In the pre-train stage, both the contrastive loss and AEC loss are used with an initial learning rate of 0.001. In the fine-tuning stage, the CL framework is removed and only the AEC loss is employed with an initial learning rate of 10^{-5}. The ReduceLROnPlateau learning rate scheduler in Tensorflow is used to monitor the validation loss, and if the model is not improved in 5 epochs, the learning rate is reduced by a factor of 0.1 until it reaches 1e-10. The model is trained with the Adam optimizer [15].

To evaluate the performance of the proposed AEC system and compare it with the state-of-the-art approaches, DTLN [2] and the ICASSP-2022 Microsoft competition baseline [9] are used as baselines and denoted as baseline in Table 1.

4. Results and Discussion

The performance of different methods under different noise levels, SNR, is summarized in Table 1 in terms of Perceptual Evaluation of Speech Quality (PESQ) [16] and Extended Short-Time Objective Intelligibility (ESTOI) [17]. PESQ and ESTOI evaluate the speech quality and intelligibility of the enhanced speech, respectively. Two types of noise, room noise, and music noise, are added for evaluation.

Our pre-training and fine-tuning AEC model outperform DTLN [2] and ICASSP-2022 Microsoft competition baseline [9] under various noise levels. For example, under 30 dB input noise level, the pre-training model can achieve 3.204 PESQ values in the room noise compared to DTLN's 2.775 and the baseline model's 3.0275. In the fine-tuning stage, the performance can be further improved to 3.4624 PESQ values in the room noise and 3.5112 in the music noise under 30 dB noise of inputs. Additionally, the model has been evaluated using the latest evaluation metric, Perceptual Objective Listening Quality Assessment (POLQA), which provides a new measurement standard for predicting Mean Opinion Scores (MOS) and outperforms the older PESQ standard. Our best model achieved an average of 3.84 POLQA values under different input noise levels, outperforming the state-of-the-art result (3.22 POLQA).



(a)

(b)

Figure 4: *An example of audio spectrum in frequency-time domain before (a) and and after (b) using SDW and ESW.*

Figure 4 visualizes the benefit of the adaptive speech quality aware method. It shows the difference in the spectrum of an example of enhanced audio before and after using SDW and ESW. By applying them, the model can remove residual echo while maintaining speech quality by compensating for distortion in both low and high-frequency bands.

5. Conclusions

The paper presents a novel approach to speech enhancement and echo cancellation through the use of an adaptive speech quality aware complex neural network with a supervised contrastive learning framework. The proposed contrastive learning pre-training framework for AEC models enables the model to learn the difference between data pairs and extract better audio representations, resulting in improved performance in downstream tasks.The proposed AEC fine-tuning framework consists of three neural network blocks, FEN, ASN, and MON, each specifically designed to target different tasks of AEC. The adaptive speech quality aware loss function results in significant benefits to remove speech echoes and improve speech quality.The experiments show that the proposed framework outperforms the state-of-the-art results in terms of average PESQ and ESTOI values under different noise levels.

6. References

[1] J. Benesty, M. M. Sondhi, Y. Huang *et al.*, *Springer handbook of speech processing*. Springer, 2008, vol. 1.

[2] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7138–7142.

[3] A. Fazel, M. El-Khamy, and J. Lee, "Cad-aec: Context-aware deep acoustic echo cancellation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6919–6923.

[4] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "Ft-lstm based complex network for joint acoustic echo cancellation and speech enhancement," *arXiv preprint arXiv:2106.07577*, 2021.

[5] X. Le, H. Chen, K. Chen, and J. Lu, "Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement," *arXiv preprint arXiv:2107.05429*, 2021.

[6] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Conference on Neural Information Processing Systems*, 2020.

[7] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, "A theoretical analysis of contrastive unsupervised representation learning," in *International Conference on Machine Learning*, 2019.

[8] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," 2020.

[9] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sørensen, and R. Aichne, "Icassp 2022 acoustic echo cancellation challenge," 2022.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[12] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[13] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[14] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation lstm network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

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

[16] "ITU-T P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." 2001.

[17] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1–1, 11 2016.