Rebuttal letter for Adaptive Speech Quality Aware Complex Neural Network for Acoustic Echo Cancellation with Supervised Contrastive Learning

***To Reviewers* -** We are grateful for all reviewers' careful review and insightful suggestions. We agree that the paper is hard to read and some prerequisite knowledge is required. We feel sorry and will fully re-organize and re-write some parts as reviewers suggested. Furthermore, we provide new experiments for contrastive learning ablation study and new figures for the loss design. We are excited that all the concerns have been addressed and the paper is getting solid and clear. Details are discussed as follows.

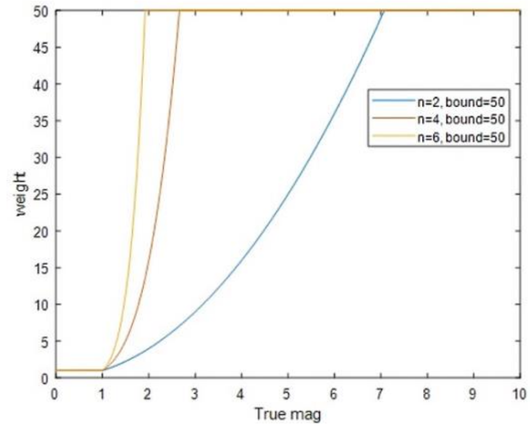
REVIEWER 1

***Question 1**: More experiments should be conducted to verify the effectiveness of contrastive learning.*

Ans: We have performed an additional experiment to show the effectiveness of contrastive learning. With contrastive learning, the PESQ will be improved from 3.12 to 3.14. We will add this result in our result table.

***Question 2**: provide evidence for the loss function design*

Ans: The loss function is designed from the observation that in the frequency domain (complex signal, two masks for real and imaginary signal respectively), the true mask could be larger than 1, which means the energy of clean signal is larger than the noisy speech. This is not well modeled in the previous work. Therefore, we design a speech distortion compensate weight to compensate the signal loss due to the noise. This weight is visualized as the figure below, which will increase the value quickly when the true mask magnitude is larger than 1. This is equivalently to the frequency bin weighted loss and emphasize the loss when the energy of true signal bin is larger than the noisy one. We will re-organize our writing and add the related figure in the final paper.



***Question 3**: suggest authors to highlight the novelty and provide solid validation for the main contribution*

Ans: The contribution of this paper can be summarized from two manifold. 1) we proposed supervised contrastive learning for AEC task, which utilized the pre-training and fine-tuning to improve the performance. 2) we further design a novel adaptive speech aware loss function (speech distortion compensation and echo suppression adaptively loss) to compensate the speech distortion and suppress the residual echo in the speech. As mentioned above, we have also showed that the PESQ value can be improved from 3.14 to 3.16 by adding this loss function design.

REVIEWER 2

***Question 1**: question about the loss function and ERLE results in Table 1*

Ans: For all the loss function, we did not perform weighted summation. All the losses are summed up directly. We will add this explanation in the table caption. The block of results ERLE=1 is mainly due to the wrong results updated to the table, since we are rushing to submit the paper. We are sorry for this and will correct it in the final version.

***Question 2**: question about the architecture design*

Ans: We will revise the architecture figure to make it clearer. It is the residual connection in the feature extraction. The architecture is quite straightforward, 3 modularized network to perform feature extraction, acoustic separation, and mask optimization respectively

***Question 3**: question about the signal preprocessing*

Ans: The signal preprocessing such as random spectral shaping, IR modelling and gains are the same as reference [13]. Interested reader can refer to [13] for details.

***Question 4**: question about the contrastive learning*

Ans: Contrastive learning has two stages: pre-training and fine-tuning. Contrastive learning is only used for pre-training for the audio presentation and then the model is finetuned for the AEC task. Reference [8-9,11] provide more details for contrastive learning

REVIEWER 3

***Question 1**: the inputs and outputs to the neural network described in Section 2.2 are in time-domain signal, while in Figure 1, they are in time-frequency domain.*

Ans: The description of Section 2.2 is a general problem formulation for AEC problem. The problem is usually solved in the time-frequency domain. We will add some explanation to avoid the confusion.

***Question 2**: The variables in the Figure 1, should be explained in the text (B, T, 2N).*

Ans: B, T, 2N refer to batch size, number of audio frames, and the audio signal real and imaginary points after performing FFT transform.

***Question 3**: Is near-end signal the same as microphone signal. It seems that this terminology is not used consequently.*

Ans: Yes. We will review the paper thoroughly and make the terminology consequently.

***Question 4**: Perhaps the description could be started from the perspective of the use of contrastive learning.*

Ans: we think this is a great idea. We will re-organize the methodology parts and discuss the contrastive learning first.