Author Responses of Paper “Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong Audio Event Taggers”

Reviewer 1

Comment 1: - Key Strength: Analysis of computational load. - Main Weakness: Not appreciated. - Novelty/Originality: Work is novel. - Technical Correctness: Work is complex and detailed. - References: Good.

Response: We thank the reviewer for the positive feedback.

Comment 2: - Clarity of Presentation: In some points the text is not clear. For example, in figure 1 says that in upper figure is shown that Whisper is noticeably more robust when speech is contaminated with background sounds from ESC-50 but I think that this association is not an evidence.

Response: We thank the reviewer for pointing this out. Figure 1 (upper) actually shows that Whisper is noticeably more robust when speech is contaminated with sounds from ESC-50. More specifically, Figure 1 shows the word error rate (WER) of Whisper increases much slower than its counterparts, indicating ESC-50 environment sound has a smaller impact on the Whisper ASR performance, in other words, Whisper is more robust. In the next version of the paper, we will revise the caption of Figure 1 to make this point more clear.

Reviewer 4

Comment 1: This paper provides us with a new insight. Traditionally, it has been assumed that speech recognition models learn noise-invariant representations in order to better distinguish human speech. It has also been the case that models are intentionally fed canonicalized training data, such as in SAT. However, what this paper reveals is that large-scale speech recognition models such as Whisper do not internally produce noise-invariant representations, but rather perform speech recognition with noise information inclusively. In this paper, by adding an additional layer to the Whisper model to train noise type, it achieved better performacen in noise classification task, than previous techniques. I think this is significant.

Response: We thank the reviewer for the very positive feedback and insightful comprehension of our paper.

Reviewer 5

Comment 1: - Key strength: The paper show a counter-intuitive finding that while Whisper is robust against background sounds (noise for ASR), its audio representation is actually not noise-invariant, but instead encodes rich information of non-speech background sounds. And based on this finding, the whisper model is used as a feature extractor for audio event taggers. At the same time, detailed experiments are designed to verify the effectiveness of the method.

Response: We thank the reviewer for the positive feedback and for comprehending our paper well.

Comment 2: Main weakness of the paper: The innovation of the method is weak. Compared with the existing method of using the pre-trained speech model to improve audio event taggers, it only replaces a more effective pre-trained model –Whisper.

Response: We thank the reviewer for the comments. The reviewer is correct in that finetuning a pretrained model with a linear layer has been extensively studied. However, we would like to clarify that the novelty of this work is significant, and not just replacing a pretrained model:

1. Just as the reviewer points out in comment 1, the novelty of this paper is not only a new unified ASR and audio tagging model, but also a new insight that **“noise-robust ASR model can and indeed learns noise-*variant* embeddings”**. To the best of our knowledge, this has not been reported before and its application is not limited to building unified ASR and audio tagging models. For example, it can be useful for future studies on noise-robust ASR design.

2. For unifying the ASR and AT model, different from previous efforts that only train a linear layer on top of the pretrained model representation(s), we design a novel **Time and Layer-wise Transformer(`TL-Tr`)** model. The motivation is that we find different sound classes achieve their best performance using different representation layers. Therefore, ideally, each class should have its own set of weights w.r.t. layers, which motivates us to build an attention mechanism over the *layers*. Performance-wise, the proposed `TL-Tr` brings a significant performance boost compared with previous work, e.g., on ESC-50, the proposed `TL-Tr` model leads to a 4.7% accuracy improvement compared to the conventional linear layer method. We also did a thorough analysis of computation efficiency, and shows the cost of `TL-Tr` is less than 1% of the ASR cost.

We will revise the manuscript to better highlight the novelty of this work.