Paper Title
Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong Audio Event Taggers

Track Name
INTERSPEECH 2023 Main Track

Reviewer #1

Questions

1. Reviewer confidence (1-3) By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Conta Meta-Reviewer".
2: Confident

2. Technical correctness of the work (1-4) Please evaluate the scientific and/or technical correctness of th work. If experiments are presented please consider if enough details are provided on the datasets, baselir and experimental design to allow the experiments to be reproduced or equivalent experiments run. If you give a score of 1 or 2, please provide further explanation in your Detailed Comments.
4: Technically solid

3. Clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing and quality of figures, tables etc. If you give a score of 1 or 2 please provide further explanation in your Detailed Comments.
3. Clear enough, could benefit from some revision

4. Overall recommendation (1-6)
5: Accept: I think this paper should be accepted

5. Detailed Comments for Authors Please supply detailed comments to back up your rankings. These comments will be forwarded to the authors of the paper. The comments will help the committee decide th outcome of the paper, and will help justify this decision for the authors. If the paper is accepted, the comments should guide the authors in making revisions for a final manuscript. Hence, the more detailed you make your comments, the more useful your review will be - both for the committee and for the author (min. 120 words). Particularly, provide feedback on the following topics: * Key Strength of the paper * Mair Weakness of the paper * Novelty/Originality, taking into account the relevance of the work for the Interspeech audience * Technical Correctness, is the work technically and/or scientifically solid? Are sufficient details provided to allow any experiments to be reproduced or equivalent experiments run? * Quality of References, is it a good mix of older and newer papers? Do the authors show a good grasp of tl current state of the literature? Do they also cite other papers apart from their own work? * Clarity of Presentation, the English does not need to be flawless, but the text should be understandable
-- Key Strength of the paper --
Analysis of computational load.
-- Main Weakness of the paper --
Not appreciated.
-- Novelty/Originality, taking into account the relevance of the work for the Interspeech audience –
Work is novel.
-- Technical Correctness, is the work technically and/or scientifically solid? Are sufficient details provided to allov any experiments to be reproduced or equivalent experiments run? --
Work is complex and detailed.
-- Quality of References, is it a good mix of older and newer papers? Do the authors show a good grasp of the current state of the literature? Do they also cite other papers apart from their own work? --
References are good.
-- Clarity of Presentation, the English does not need to be flawless, but the text should be understandable --
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.

Reviewer #4

Questions

1. Reviewer confidence (1-3) By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Conta Meta-Reviewer".
2: Confident

2. Technical correctness of the work (1-4) Please evaluate the scientific and/or technical correctness of th work. If experiments are presented please consider if enough details are provided on the datasets, baselir and experimental design to allow the experiments to be reproduced or equivalent experiments run. If you give a score of 1 or 2, please provide further explanation in your Detailed Comments.
4: Technically solid

3. Clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing and quality of figures, tables etc. If you give a score of 1 or 2 please provide further explanation in your Detailed Comments.
3. Clear enough, could benefit from some revision

4. Overall recommendation (1-6)
5: Accept: I think this paper should be accepted

5. Detailed Comments for Authors Please supply detailed comments to back up your rankings. These comments will be forwarded to the authors of the paper. The comments will help the committee decide th outcome of the paper, and will help justify this decision for the authors. If the paper is accepted, the comments should guide the authors in making revisions for a final manuscript. Hence, the more detailed you make your comments, the more useful your review will be - both for the committee and for the author (min. 120 words). Particularly, provide feedback on the following topics: * Key Strength of the paper * Mair Weakness of the paper * Novelty/Originality, taking into account the relevance of the work for the Interspeech audience * Technical Correctness, is the work technically and/or scientifically solid? Are sufficient details provided to allow any experiments to be reproduced or equivalent experiments run? * Quality of References, is it a good mix of older and newer papers? Do the authors show a good grasp of tl current state of the literature? Do they also cite other papers apart from their own work? * Clarity of Presentation, the English does not need to be flawless, but the text should be understandable
This paper provides us with a new insight.
Traditionally, it has been assumed that speech recognition models learn noise-invariant representations in order t 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 internall 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 performac in noise classification task, than previous techniques.
I think this is significant.

Reviewer #5

Questions

1. Reviewer confidence (1-3) By selecting a value, you confirm that you are competent to evaluate this

Meta-Reviewer".
3. Very Confident

2. Technical correctness of the work (1-4) Please evaluate the scientific and/or technical correctness of th work. If experiments are presented please consider if enough details are provided on the datasets, baselir and experimental design to allow the experiments to be reproduced or equivalent experiments run. If you give a score of 1 or 2, please provide further explanation in your Detailed Comments.
4: Technically solid

3. Clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing and quality of figures, tables etc. If you give a score of 1 or 2 please provide further explanation in your Detailed Comments.
3. Clear enough, could benefit from some revision

4. Overall recommendation (1-6)
4: Weak Accept: I am leaning to accept this paper

5. Detailed Comments for Authors Please supply detailed comments to back up your rankings. These comments will be forwarded to the authors of the paper. The comments will help the committee decide th outcome of the paper, and will help justify this decision for the authors. If the paper is accepted, the comments should guide the authors in making revisions for a final manuscript. Hence, the more detailed you make your comments, the more useful your review will be - both for the committee and for the author (min. 120 words). Particularly, provide feedback on the following topics: * Key Strength of the paper * Mair Weakness of the paper * Novelty/Originality, taking into account the relevance of the work for the Interspeech audience * Technical Correctness, is the work technically and/or scientifically solid? Are sufficient details provided to allow any experiments to be reproduced or equivalent experiments run? * Quality of References, is it a good mix of older and newer papers? Do the authors show a good grasp of tl current state of the literature? Do they also cite other papers apart from their own work? * Clarity of Presentation, the English does not need to be flawless, but the text should be understandable
Key strength of paper:
The paper show a counter-intuitive finding that while Whisper is robust against background sounds (noise for AS 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.
.

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

Technical Correctness, is the work technically and/or scientifically solid? Are sufficient details provided to allow a experiments to be reproduced or equivalent experiments run?
The technical correctness of the work is solid enough.

Quality of References, is it a good mix of older and newer papers? Do the authors show a good grasp of the current state of the literature? Do they also cite other papers apart from their own work?
The quality of References is high enough, covering the latest progress and important work in the field.

Clarity of Presentation, the English does not need to be flawless, but the text should be understandable
The English expression of this paper is clear enough and the text is easy to understand.