View Reviews

Paper ID
383

Paper Title
Joint Audio and Speech Understanding

Reviewer #1

Questions

2. Technical correctness Indicate your opinion about the technical correctness of the approach, including the assumptions, mathematical formulation, algorithms, experimental design, and whether the results are trustworthy.
Technically solid

3. Novelty Novelty may be in the formulation, algorithm, and application. In addition, new findings about a known approach that would be of interest to the community could be also considered novel. If you chose "Not novel", please refer to the existing work in the comments below.
Novel enough - Novel contribution to an existing approach, new experimental findings

4. Clarity of Presentation Take into account the organization of the paper, the writing style and use of language, and the quality of figures and tables.
Clear enough, could benefit from some revision

5. Reproducibility While providing an open source implementation and evaluation on freely accessible data would be ideal for reproducibility, the approach should be considered reproducible as long as enough details are given to reproduce the results even if the code and data are not available.
Should be reproducible - The paper contains enough details for an experienced researcher to reproduce the major findings of the work.

6. Overall recommendation Indicate your overall opinion about whether you would like to see the paper at the workshop.
Accept

7. Brief summary of the paper This part tells the committee what the major contributions are, what the authors did, how they did it, and what the results were. It also helps authors to verify that the reviewer understood their approach and interpretation of the results.
Human-like recognition of speech and audio is attempted in this work. The authors propose a model for universal audio perception using OpenAQA dataset. Several samples have also been presented to demonstrate the performance of their system.

8. Key strengths of the paper
A novel method has been proposed for introducing human-like speech and audio perception task.

9. Main weakness of the paper
NA

10. Detailed Comments for Authors Please supply detailed comments to back up your rankings. The comments will help the committee decide the 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 authors. This should include comments about: - Novelty/Originality Take into account the relevance of the work for the ASRU audience The paper may be a technical contribution in the form of a new algorithm or address new problems or tasks without a novel algorithmic contribution. Judge these papers by the novelty of their insights and the value these insights could have for the community. - Technical Correctness Take into account datasets, baselines, and experimental design. Are enough details provided to be able to reproduce the experiments? - Quality of References Is it a good mix of older and newer papers and covering various authors? Do the authors show a good grasp of the current state of the literature? - Clarity of Presentation The English does not need to be flawless, but the text should be understandable.

Human-like recognition of speech and audio is attempted in this work. The authors propose a model for universal audio perception using OpenAQA dataset. Several samples have also been presented to demonstrate the performance of their system. All the references are cited in the main text and they are good mix of recent and old works.

Reviewer #2

Questions

2. Technical correctness Indicate your opinion about the technical correctness of the approach, including the assumptions, mathematical formulation, algorithms, experimental design, and whether the results are trustworthy.

Technically solid

3. Novelty Novelty may be in the formulation, algorithm, and application. In addition, new findings about a known approach that would be of interest to the community could be also considered novel. If you chose "Not novel", please refer to the existing work in the comments below.

Novel enough - Novel contribution to an existing approach, new experimental findings

4. Clarity of Presentation Take into account the organization of the paper, the writing style and use of language, and the quality of figures and tables.

Clear enough, could benefit from some revision

5. Reproducibility While providing an open source implementation and evaluation on freely accessible data would be ideal for reproducibility, the approach should be considered reproducible as long as enough details are given to reproduce the results even if the code and data are not available.

Should be reproducible - The paper contains enough details for an experienced researcher to reproduce the major findings of the work.

6. Overall recommendation Indicate your overall opinion about whether you would like to see the paper at the workshop.

Accept

7. Brief summary of the paper This part tells the committee what the major contributions are, what the authors did, how they did it, and what the results were. It also helps authors to verify that the reviewer understood their approach and interpretation of the results.

The authors extended the audio QA/instruction dataset, OpenAQA, with speech QA/instruction pairs derived from a number of tasks, to create the large scale Open-ASQA dataset containing both open- and close-ended instructions. The authors then extend LTU to build a LoRA-tuned LLaMA combining Whisper features and transcripts as inputs after TLTR layers to process input prompts containing both speech/audio and text, and reach superior performance on various downstream tasks possibly with zero-shot.

8. Key strengths of the paper

The authors create a large-scale synthetic dataset for instruction following on the speech and audio modality, which would be valuable for future research. Based on that, the authors build a LLaMA-based system following the scheme of LTU that effectively incorporate speech inputs into LLaMA for speech and audio understanding.

9. Main weakness of the paper

1. More than 50% samples in Open-ASQA are from OpenAQA, i.e. audio, not speech, while speech can be rather diverse in both semantic and paralinguistic sense. As shown in Table 4, keeping only the audio or the transcription inputs has varied impacts. Will the focus on audio rather than speech in the training dataset leads to suboptimal results on speech tasks? What if you downsample the data for audio during training to place more focus on speech?
2. The models are only compared with methods like AudioClip on close-ended tasks, and other LTU variants on open-ended tasks. It could be important to evaluate and compare with other recent LLM-enhanced speech models following the same evaluation protocol, e.g. ImageBind-LLM (https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/imagebind_LLM), Speech-LLaMA (https://arxiv.org/abs/2307.03917), and SpeechGPT. It is also valuable to justify the end-to-end modelling, particularly considering that the dataset is created using GPT3.5 with only textual description of audios. How will LLaMA with LoRA perform if the inputs are merely such textual description generated by separate models instead of the actual audio?

10. Detailed Comments for Authors Please supply detailed comments to back up your rankings. The comments will help the committee decide the 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 authors. This should include comments about: - Novelty/Originality Take into account the relevance of the work for the ASRU audience The paper may be a technical contribution in the form of a new algorithm or address new problems or tasks without a novel algorithmic contribution. Judge these papers by the novelty of their insights and the value these insights could have for the community. - Technical Correctness Take into account datasets, baselines, and experimental design. Are enough details provided to be able to reproduce the experiments? - Quality of References Is it a good mix of older and newer papers and covering various authors? Do the authors show a good grasp of the current state of the literature? - Clarity of Presentation The English does not need to be flawless, but the text should be understandable.

Since Whisper features from different layers are incorporated, it will be interesting to know the roles and importance of different layers in the speech encoder across different tasks, e.g. by ablation studies or showing the attention weights in the layer transformer.

Whisper emphasizes its multilingual and speech translation capability. Despite LLaMA being heavily biased towards English, Whisper features may be more language agnostic and leads to stronger multilingual support thanks to the speech translation training (as in https://arxiv.org/abs/2305.09652). I'm curious if the model could also demonstrate some capability understanding non-English speech.

Sec 3.2: How diverse are the generated questions? How different are they compared to each other and the examples in the input prompt? Can you provide some measures (e.g. BLEU, overlapped n-grams)?

Sec 5.1: "Even on audio classification and age classification tasks where {S} is not useful" I guess the author means gender instead of age according to Table 4.

The authors investigated zero-shot learning capabilities, while a key capability of LLM is few-shot incontext learning. Have the authors attempted to investigate that?

The manuscripts mentioned that the code and models will be released but didn't mention the dataset. Will Open-ASQA be released as well?

There are formatting issues in the references, e.g. [2] should be "LLaMA" instead of "Llama".

Reviewer #3

Questions

2. Technical correctness Indicate your opinion about the technical correctness of the approach, including the assumptions, mathematical formulation, algorithms, experimental design, and whether the results are trustworthy.
Technically solid

3. Novelty Novelty may be in the formulation, algorithm, and application. In addition, new findings about a known approach that would be of interest to the community could be also considered novel. If you chose "Not novel", please refer to the existing work in the comments below.
Novel enough - Novel contribution to an existing approach, new experimental findings

4. Clarity of Presentation Take into account the organization of the paper, the writing style and use of language, and the quality of figures and tables.
Very well written

5. Reproducibility While providing an open source implementation and evaluation on freely accessible data would be ideal for reproducibility, the approach should be considered reproducible as long as enough details are given to reproduce the results even if the code and data are not available.
Should be reproducible - The paper contains enough details for an experienced researcher to reproduce the major findings of the work.

6. Overall recommendation Indicate your overall opinion about whether you would like to see the paper at the workshop.
Definite Accept

7. Brief summary of the paper This part tells the committee what the major contributions are, what the authors did, how they did it, and what the results were. It also helps authors to verify that the reviewer understood their approach and interpretation of the results.
This paper presents a audio question answering (QA) system that comprise an LLM (LLaMA 7B), an ASR model (Whisper), and Whisper-based audio tagger (Time- and Layer-wise Transformer; TLTR [3]). On the basis of an important finding [3] that Whisper recognizes not only speech but also general audio events in an implicit manner, TLTR takes an important role as a speech and audio encoder. The propsed system combines TLTR with LLaMA, resulting a large-scale audio QA model trained on the newly-developed "Open-ASQA" dataset. The experiments show the model's reasoning ability as well as instruction following rate.

8. Key strengths of the paper
A well-written paper. What is required is sufficiently described within 8 pages.

9. Main weakness of the paper
The effectiveness of the multi-stage training curriculum (Table 3) is not presented in the experiment section.

10. Detailed Comments for Authors Please supply detailed comments to back up your rankings. The comments will help the committee decide the 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 authors. This should include comments about: - Novelty/Originality Take into account the relevance of the work for the ASRU audience The paper may be a technical contribution in the form of a new algorithm or address new problems or tasks without a novel algorithmic contribution. Judge these papers by the novelty of their insights and the value these insights could have for the community. - Technical Correctness Take into account datasets, baselines, and experimental design. Are enough details provided to be able to reproduce the experiments? - Quality of References Is it a good mix of older and newer papers and covering various authors? Do the authors show a good grasp of the current state of the literature? - Clarity of Presentation The English does not need to be flawless, but the text should be understandable.

It is worth noting that the most part of the model is frozen and only 0.6% (49M) of the entire parameters are trainable.