JOINT AUDIO AND SPEECH UNDERSTANDING

Yuan Gong¹, Alexander H. Liu¹, Hongyin Luo¹, Leonid Karlinsky², James Glass¹
¹MIT CSAIL, USA ²MIT-IBM Watson AI Lab, USA

ABSTRACT

Humans are surrounded by audio signals that include both speech and non-speech sounds. The recognition and understanding of speech and non-speech audio events, along with a profound comprehension of the relationship between them, constitute fundamental cognitive capabilities. For the first time, we build a machine learning model, called LTU–AS, that has a conceptually similar universal audio perception and advanced reasoning ability. Specifically, by integrating Whisper [1] as a perception module and LLaMA [2] as a reasoning module, LTU–AS can simultaneously recognize and jointly understand spoken text, speech paralinguistics, and non-speech audio events - almost everything perceivable from audio signals.

1. INTRODUCTION

Humans live in a multifarious environment of audio signals, encompassing both speech and a wide variety of non-speech sounds. The ability to accurately discern, interpret, and integrate these speech and non-speech audio elements, in conjunction with a profound understanding of the interrelationships they entail, represents a fundamental cognitive capability of humans. When we hear “watch out!” and a car horn simultaneously, we can infer the danger. If we hear birds chirping and someone says “that’s a rare one,” we know there is an unusual bird nearby. Understanding music usually requires paying attention to both the lyrics and the melody.

However, most existing machine learning models can only recognize either speech or audio events. Further, while being strong in audio or speech perception, these models possess limited reasoning and understanding capabilities. This motivates us to build a joint audio and speech understanding model that is able to simultaneously recognize and jointly understand speech and audio events. Particularly, as shown in Figure 1, our model integrates pretrained Whisper [1] automatic speech recognizer (ASR) and a time and layerwise Transformer (TLTR) [3] as the perception module and LLaMA [2] large language model (LLM) as the reasoning module. In addition, we formulate the training data as (audio, question, answer) (AQA) tuples, which allows us to combine 13 audio and speech datasets of various tasks with different label sets into a single 9.6M Open-ASQA dataset, among which 6.9 million data are open-ended AQA tuples generated by GPT [4] with audio instruction generation [5].

call our model LTU–AS (listen to, think of, and understand audio and speech). Performance-wise, we show LTU–AS is strong on all audio/speech tasks. But more importantly, as shown in Fig. 1 and Table 6, LTU–AS can answer free-form open-ended questions about the audio and speech with an instruction following rate over 95% (evaluated by GPT-4), and exhibits emerging joint audio and speech reasoning ability.

Related Work: LTU–AS substantially improves the recent audio large language model LTU [5] that only understands non-speech audio. Particularly, LTU–AS adopts Whisper [1] and TLTR [3] as the audio encoder instead of the AST [6] audio encoder in LTU. This change enables LTU–AS to recognize both speech and audio events. We also augment the LTU OpenAQA-5M dataset with 4 million speech and audio/speech understanding AQAs in creating the 9.6M Open-ASQA dataset. There are a few recent efforts on joint audio and speech recognition [7, 8, 9, 10] but none of them exhibit advanced joint reasoning ability. Other recent audio LLMs [11, 12, 13, 14] primarily focus on only speech. To the best of our knowledge, LTU–AS is the first joint audio and speech understanding model.
2. LTU-AS MODEL ARCHITECTURE

2.1. Design Overview

The architecture of LTU-AS is depicted in Fig. 1. The system input is a pair of audio and question in natural language form. The audio is first input to the Whisper audio encoder. Then, the output of the Whisper encoder is fed to the Whisper decoder to transcribe it to discrete spoken text (if there is no speech, then the output of the decoder will be empty, which is as expected). Meanwhile, we feed the output of all 32 Whisper encoder intermediate layers to an AudioSet-pretrained Time and Layer-Wise Transformer (TLTR) [3] to encode "soft" audio events and speech paralinguistic information, and then project to a series of continuous audio tokens \{A\} with a linear layer.

During training, the entire Whisper model is frozen. Only the TLTR model and projection layer are trainable. This design is due to a few reasons: First, training a large language model as an automatic speech recognizer (ASR) can be very expensive but the benefit is unclear [14]; we thus choose to freeze the entire Whisper model to inherit its strong and robust ASR ability. Second, although the Whisper encoder encodes rich audio events and speech paralinguistic information [3, 15, 16], it encodes information in the representations of different layers. Since we anticipate LTU-AS being a universal perception model, we use the TLTR model to apply attention mechanisms over both time and layers.

The key advantage of this setting is that the audio is encoded to both text and continuous tokens, so both linguistic and non-linguistic information are kept. We then tokenize and embed the spoken text and input question to a sequence of text tokens \{S\} and \{Q\}, respectively. Finally, we concatenate and input \{A\}, \{S\}, and \{Q\} to the LLaMA LLM. Due to the computational limit, we trim the length of the audio token \{A\} to 25 (corresponding to 10 seconds of audio), but allow \{S\} and \{Q\} to be of variable length.

2.2. Audio Encoder

Whisper [1] is a recently proposed robust ASR model that features a standard Transformer [17]-based encoder-decoder architecture trained with a massive 680k hour labeled speech corpus recorded in diverse conditions. Notably, it was found that the Whisper encoder features not only encode linguistic information, but also encode rich general background sound information [3] and paralinguistic and other information (e.g., emotion [15] and language development [16]). In this paper, we use the Whisper-large model whose encoder and decoder are both 32-layer, 1280-dimensional Transformer networks.

Time and Layer-Wise Transformer (TLTR): We use the AudioSet pretrained TLTR for Whisper proposed in [3], originally for audio event detection. We empirically find there is no need to pretrain it further on speech classification tasks before training together with LTU-AS. Whisper and TLTR pool the audio with a factor of 40, i.e., for each 10-second audio (1000 frames), the length of the TLTR output is 25 (2.5Hz).

Projection Layer: We use a single linear layer to project the TLTR output from 1280-dimensional to 4096-dimensional to match the embedding dimension of the LLaMA LLM.

2.3. LLaMA Large Language Model

We use the LLaMA-7B LLM [2] with Vicuna [18] instruction following for fine-tuning. To mitigate catastrophic forgetting [19] and save computation, we freeze the entire LLaMA model and adopt Low-rank Adaptation [20] (LoRA), which introduces a small set of auxiliary learnable weights on top of the pre-trained LLaMA model. Specifically, we inject LoRA adapters (rank=8 and α=16) to the projection layers for all keys and queries in all LLaMA self-attention layers [17].

Training Objective: As an audio LLM, LTU-AS is trained on the next token prediction task conditioning on the past tokens and the reference audio, i.e., maximizing \( P(O_t \mid O_{1:t-1}, A, S, Q), \) through cross-entropy for all \( 1 < t \leq T \) given the tokenized ground truth text sequence (i.e., output) \( O_{1:T} \) and the reference audio token \( A \), spoken text \( S \), and question \( Q \). This training objective allows us to unify nearly all audio and speech tasks except audio/speech generation into a single training framework.

Generation Setting: We use a plain generation setting of Temperature=0.1, Top K=500, and Top P=0.95 with a repetition penalty of 1.1 [21, 22] for all tasks.

Model Parameters: As a LLM, LTU-AS has about 8.5 billion parameters. However, only 49 million parameters are actually trainable (40M for TLTR, 4.2M for LoRA adapters, and 5M for the projection layer), which is only about 0.6% of the total number of parameters. This significantly lowers the computation requirement to train LTU-AS. Practically, LTU-AS is trained on 4× A6000 GPUs for about 80 hours.

3. THE OPEN-ASQA DATASET

We aim to build LTU-AS to address a wide range of open-ended audio and speech tasks, and understand the audio and speech jointly. To achieve this objective, we need a training dataset to provide such joint audio and speech supervision. Unfortunately, there is no existing dataset that meets our needs. The closest one is the OpenAQA dataset used to train LTU [5], which is an audio question-answering dataset consisting of 1.9 million closed-ended and 3.7 million open-ended AQAAs. However, OpenAQA lacks speech related, and joint audio-speech questions. Therefore, on the basis of OpenAQA-5M, we add an additional 2.7 million speech-related AQAAs (0.9 million closed-ended and 1.8 million open-ended) and 1.2 million joint audio and speech AQAs (almost all open-ended), and build a new 9.6M Open-ASQA dataset. Note that we do not collect new audio and speech data, but instead relabel 13 existing public datasets summarized in Table 2. For all these datasets, we only include data marked as training and validation samples and exclude any data marked as test or evaluation.
As with OpenAQA, all Open–ASQA samples are formatted as (audio, question, answer) tuples, where “audio” and “question” are the model inputs, and “answer” is the ground truth label. By unifying all training samples in this format, we not only map all labels to a semantic space, but are also able to train LTU–AS with a variety of different tasks easily.

3.1. Closed-Ended QA Generation

For each task and dataset, we paraphrase the question (e.g., “What is the audio event”) with GPT-3.5-Turbo assistance to generate a diverse question set, so LTU–AS won’t overfit to a specific question for a task. However, the answers are generated with a rule-based algorithm based on the original label of the dataset, and thus have a fixed format. We thus call such AQAs closed-ended AQAs. The upper section of Table 1 shows samples of closed-ended AQA pairs.

**Closed-Ended Audio AQA**: Closed-ended audio questions are from the original OpenAQA dataset, which consists of 1.9 million AQAs about the audio event labels, acoustic features, audio captioning, and audio temporal analysis. The audio tracks are from 8 audio datasets. Please refer to Table 2 and [5] for more details.

**Closed-Ended Speech AQA**: We created 941k closed-ended speech AQAs based on 4 commonly used speech datasets. The first category of questions asks the original labels of the datasets, e.g., speaker emotion/gender for IEMOCAP [23], speaker emotion and sentiment score for MOSEI [24], speaker gender for LibriTTS [25], and speaker age and gender for VoxCeleb2 [26, 27]. In addition to these original labels, we further annotate the speaker style of speech speed (computed with Whisper time stamps), pitch, and energy (computed with librosa [28]), and generate AQAs asking the speaker style. Finally, we also mix in about 150k ASR AQAs that have questions asking about the spoken text, and the answers are the transcriptions. Note that since LTU–AS has an internal Whisper model feeding the transcribed text to LLaMA, the ASR task is no more than an identity mapping for LTU–AS, which is fundamentally different from SpeechGPT [14]. We include ASR AQAs just to guide the model following ASR instructions.

**Closed-Ended Joint Audio and Speech AQA**: Most joint audio and speech AQAs in this project are open-ended. The only 93k closed-ended joint audio and speech AQAs are of the music genre prediction task on the FMA [29] dataset, which requires LTU–AS to consider both lyrics (text) and acoustic information to make the prediction.

3.2. Open-Ended QA Generation

Generating diverse open-ended QA pairs at a large scale poses challenges with human-based efforts being impractical, and rule-based methods limiting output diversity. We thus use Audio Instruction Generation (AIG) proposed in [5] to generate open-ended AQAs with GPT-3.5-Turbo assistance. Specifically, since GPT does not take audio or speech as input, we input the meta information of the audio (e.g., audio events, speech style, emotion, and spoken text) to the GPT-3.5-Turbo model in the form of pure text as a surrogate, and then use the prompt shown in Table 1 to let the GPT model generate QA pairs. As shown in Table 1, the generated open-ended QA pairs are diverse and of high quality. Notably, 65.1% of open-ended questions appear only once in the dataset.

Note that AIG is only used for data generation; during model training, only the raw audio and generated QA pairs are input to the LTU–AS model. Thus, the model is forced to learn directly from the raw audio that contains richer and more fine-grained information compared to the extracted meta-information. Similarly, during inference, LTU–AS solely uses raw audio to answer the question.

**Open-Ended Audio AQA**: We use the approximately 3.7 million AQAs about non-speech audio events from the original OpenAQA dataset [5].

Table 1. Sample closed- and open-ended training QA pairs.
Open-ended QA pairs are generated by Audio Instruction Generation (AIG) with the shown GPT prompt and input.

*GPT prompt is shortened due to space limitation.

<table>
<thead>
<tr>
<th>Closed-Ended QA Samples</th>
<th>Open-Ended QA Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: Identify the source? A: Dog, Animal.</td>
<td>GPT Prompt*: Based on the following audio/speech, generate 10 different types of complex open-ended questions that require step-by-step thinking, and corresponding answers. Questions can be e.g., How are speech content and background sounds related? ... etc. Format each QA pair in a single line as a JSON dictionary. (varies for each task/dataset)</td>
</tr>
<tr>
<td>Q: Write a caption for the sound? A: An engine revving and tires squealing.</td>
<td>GPT-Input: In the recording, background sound of domestic animals, dog and speech of “Get him going, get him going with it.” is heard.</td>
</tr>
<tr>
<td>Q: What emotion does the speaker convey? A: Happy, with a sentiment score of 2 (positive).</td>
<td>GPT-Output: Q: What can be inferred from the audio? A: It can be inferred that the speech is addressing the dog in the background.</td>
</tr>
<tr>
<td>Q: Can you determine the music genre? A: Folk, Pop.</td>
<td>GPT-Input: In the recording, background sound of explosion and speech of “Flames engulfed the tank and began to heat the propane inside. Twenty volunteer firefighters...” is heard.</td>
</tr>
<tr>
<td><strong>Open-Ended QA Samples</strong></td>
<td><strong>Closed-Ended QA Samples</strong></td>
</tr>
<tr>
<td>GPT-Input: Speech: “Hi, my name is Jamie and I’m a Master Sales Consultant here at Northline Volkswagen...”; Speaker gender: Male; Pitch: low; Volume: very high; Speed: very slow; Emotion: happy.</td>
<td>GPT-Input: Music genre: Folk; Music Lyrics: “The man of his dreams is all that he seems Flowers in sunshine...”</td>
</tr>
<tr>
<td>GPT-Output: Q: Why do the speech content and explosion sound create a tense atmosphere? A: The speech content describes a tragic event and the explosion sound adds intensity and suspense to the narrative.</td>
<td>GPT-Output: Q: What mood does this music convey? A: The music conveys a cheerful and relaxed mood because of the lively melody and lyrics about flowers and sunshine.</td>
</tr>
</tbody>
</table>
Table 2. The statistics of the 9.6-million Open-ASQA dataset. “x” denotes the corresponding label is used.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Audio Event</th>
<th>Audio Caption</th>
<th>Spoken Text*</th>
<th>Speaker Gender</th>
<th>Speaker Age</th>
<th>Speech Style</th>
<th>Speaker Emotion</th>
<th>Music Genre</th>
<th># Audio Clips</th>
<th># Closed-Ended QAs</th>
<th># Open-Ended QAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS-Strong [30]</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>102k</td>
<td>683k</td>
<td>901k</td>
</tr>
<tr>
<td>AudioSet [31]</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>500k</td>
<td>538k</td>
<td>184k</td>
</tr>
<tr>
<td>VOGSound [32]</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>184k</td>
<td>367k</td>
<td>907k</td>
</tr>
<tr>
<td>FSD50K [33]</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>41k</td>
<td>82k</td>
<td>403k</td>
</tr>
<tr>
<td>AudioCaps [34]</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>46k</td>
<td>97k</td>
<td>478k</td>
</tr>
<tr>
<td>FreeSound [35]</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>91k</td>
<td>91k</td>
<td>791k</td>
</tr>
<tr>
<td>Clotho [36]</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5k</td>
<td>48k</td>
<td>89k</td>
</tr>
<tr>
<td>Sound Bible [37]</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.2k</td>
<td>12k</td>
<td>10k</td>
</tr>
<tr>
<td>Sum</td>
<td>845k</td>
<td>1,918k</td>
<td>3,763k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Speech Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Audio Event</th>
<th>Audio Caption</th>
<th>Spoken Text*</th>
<th>Speaker Gender</th>
<th>Speaker Age</th>
<th>Speech Style</th>
<th>Speaker Emotion</th>
<th>Music Genre</th>
<th># Audio Clips</th>
<th># Closed-Ended QAs</th>
<th># Open-Ended QAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP [23]</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>4.3k</td>
<td>26k</td>
<td>83k</td>
</tr>
<tr>
<td>LibriTTS [25]</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22k</td>
<td>167k</td>
<td>418k</td>
</tr>
<tr>
<td>VoxCeleb2 [26]</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>107k</td>
<td>194k</td>
<td>926k</td>
</tr>
<tr>
<td>MOSEI [24]</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>18k</td>
<td>554k</td>
<td>355k</td>
</tr>
<tr>
<td>Sum</td>
<td>151k</td>
<td>941k</td>
<td>1,784k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Joint Audio and Speech Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Audio Event</th>
<th>Audio Caption</th>
<th>Spoken Text*</th>
<th>Speaker Gender</th>
<th>Speaker Age</th>
<th>Speech Style</th>
<th>Speaker Emotion</th>
<th>Music Genre</th>
<th># Audio Clips</th>
<th># Closed-Ended QAs</th>
<th># Open-Ended QAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AudioSet [31]</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>82k</td>
<td>-</td>
<td>747k</td>
</tr>
<tr>
<td>FMA [29]</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>93k</td>
<td>93k</td>
<td>396k</td>
</tr>
<tr>
<td>Sum</td>
<td>175k</td>
<td>93k</td>
<td>1,143k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (9,641k Question Answer Pairs)</td>
<td>1,089k</td>
<td>2,951k</td>
<td>6,690k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Open-Ended Speech AQA: We generate open-ended AQA about speech using the four commonly used datasets IEMOCAP [23], MOSEI [24], LibriTTS [25], and VoxCeleb2 [26, 27]. We input all speech meta information including the original dataset labels (e.g., speaker emotion, gender, and age), extracted speech style features (e.g., pitch, speed, and volume), and Whisper transcribed spoken text, all in text form, to GPT-3.5-Turbo with the prompt shown in Table 1. For age, pitch, speed, and volume, we also quantize each of them into 5 categories (e.g., very low - very high) to help GPT understand the value. The input meta information to GPT of each dataset is marked as “x” in Table 2. Our intent was to input as much information as possible to GPT to generate high-quality AQA.

Open-Ended Joint Audio and Speech AQA

We use two datasets containing both speech and non-speech audio to generate joint audio and speech AQA. The first dataset we use is AudioSet [31]. Although AudioSet-2M has about 1M samples containing speech, and it has already been used in the original OpenAQA dataset, the spoken text was ignored. Specifically, a single label “speech” rather than the actual spoken text is input to GPT-3.5-Turbo for OpenAQA generation. In this work, we first sample a 500k subset from AudioSet-2M using the sound class balancing algorithm proposed in [38] to guarantee the diversity of non-speech audio events. We then use Whisper to transcribe the 500k AudioSet subset and select samples having no_speech_prob<0.2 and spoken text length over 5. This heuristic made it quite likely that the spoken text was transcribed correctly and had sufficient length to encompass substantive content. This resulted in 82k samples meeting the requirement. They were used to generate joint audio and speech AQA with GPT assistance. As shown in Table 1, GPT can generate AQA for joint audio and speech understanding, e.g., in the first sample, GPT outputs an answer explaining the speech is addressing the dog by understanding the speech content and the dog sound.

The second dataset we use is the FMA [29] music dataset. We input the list of music genres, title (if provided), and Whisper transcribed lyrics of each music clip to GPT and let it generate AQA about music understanding with joint lyrics and melody analysis. In total, we generated about 1.1 million open-ended joint audio and speech AQA.

Table 3. The LTU-AS training curriculum.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Tr. Params</th>
<th>Tr. Task</th>
<th>Tr. Samples</th>
<th>LR</th>
<th># Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Proj.</td>
<td>Cla.</td>
<td>2.1M</td>
<td>1e-3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Proj. + TLTR + LoRA</td>
<td>Cla.</td>
<td>2.1M</td>
<td>2e-4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Proj. + TLTR + LoRA</td>
<td>All</td>
<td>9.6M</td>
<td>2e-4</td>
<td>1</td>
</tr>
</tbody>
</table>

4. TRAINING LTU-AS

As for the LTU model [5], we use a three-stage training curriculum shown in Table 3 to train LTU-AS. In the first stage, only the randomly initialized projection layer is trainable. The TLTR and LoRA adapters are unfrozen in the second and third stages to stabilize training. In addition, in the first and second stages, we only train LTU-AS with AQA of classification tasks where the model gets a high penalty for wrong predictions. The model is thus forced to attend to the audio input rather than using its language ability to hallucinate.
LTU-AS is an audio-specialized model, (for other classification tasks). Label that has the highest similarity score as the prediction. LTU-AS is trained with only audio and only speech datasets. Specifically, we use a fixed prompt (e.g., “write an audio caption describing the sound.” for audio-text model). Generalizable to unseen labels, but a pre-defined label set is required for inference. The results are summarized in Table 4. First, as a foundational model, LTU-AS performs well on both audio and speech tasks. It works particularly well on tasks requiring both audio and speech understanding, which exactly meets our expectations. E.g., the accuracy of LTU-AS is nearly twice that of CLAP [48] on the zero-shot GTZAN music genre classification task; the MAE of speaker age prediction is even lower than the SOTA specialized model that does not require any pre-defined label set and directly outputs predictions in natural language, which makes it a more practical system for real-world applications.

Second, training with both non-speech audio and speech data is crucial for LTU-AS to become a unified sound perception model. In Ablation Study 1, we compare LTU-AS with LTU models trained with only audio and only speech datasets. Though audio- and speech-specialized LTUs perform slightly better on tasks in their respective training domain, they almost fail on tasks in the domain they are not trained on.

Table 4. Closed-ended task performance. ZS: Zero-shot evaluation; ZS-: The dataset is not used in training, but it is sourced from the same project as part of the training data. * Model does not follow instructions on part of or entire of the dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Best supervised &amp; Specialized</td>
<td>97.0 [42]</td>
<td>17.7 [43]</td>
<td>1.4 [44]</td>
<td>70.6 [45]</td>
<td>98.3 [27]</td>
<td>9.4 [27]</td>
<td>93.9 [46]</td>
</tr>
<tr>
<td>CLIP-like audio-text model. Generalizable to unseen labels, but a pre-defined label set is required for inference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AudioClip [47]</td>
<td>69.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CLAP [48]</td>
<td>82.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(Proposed) One single model for all tasks. Directly output label names, no pre-defined label set is needed at inference.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LTU-AS</td>
<td>80.8**</td>
<td>15.0</td>
<td>4.9</td>
<td>65.2</td>
<td>90.8</td>
<td>7.3</td>
<td>50.3**</td>
</tr>
</tbody>
</table>

Ablation Study 1 - Train with only speech or audio data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LTU (Audio Training Only) [5]</td>
<td>82.8</td>
<td>17.0</td>
<td>104.2</td>
<td>38.2</td>
<td>77.0</td>
<td>Fail*</td>
<td>29.8</td>
</tr>
<tr>
<td>LTU (Speech Training Only)</td>
<td>10.9</td>
<td>0.5</td>
<td>12.9</td>
<td>69.8</td>
<td>90.1</td>
<td>7.9</td>
<td>23.5</td>
</tr>
</tbody>
</table>

Ablation Study 2 - Inference with missing modality

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LTU-AS (Audio Input Only)</td>
<td>81.9</td>
<td>14.9</td>
<td>97.2</td>
<td>58.6</td>
<td>95.6</td>
<td>8.2</td>
<td>48.2</td>
</tr>
<tr>
<td>LTU-AS (Spoken Text Input Only)</td>
<td>7.7</td>
<td>3.5</td>
<td>20.0</td>
<td>45.4</td>
<td>42.0</td>
<td>11.9*</td>
<td>21.5</td>
</tr>
</tbody>
</table>

5. EXPERIMENTS

5.1. Closed-Ended Tasks Evaluation

Although the main novelty of LTU-AS is open-ended audio and speech understanding, we first rigorously evaluate its performance on seven standard closed-ended audio and speech tasks because these tasks serve as the foundation for advanced reasoning. Specifically, for each task, we use a fixed prompt (e.g., “write an audio caption describing the sound.”) for audio classification) and either apply a regular expression to the LTU-AS to get the prediction (for ASR, audio captioning, gender classification, and age prediction), or compute the cosine similarity between the text embedding (gpt-text-embedding-ada) of LTU-AS output and each label, and use label that has the highest similarity score as the prediction (for other classification tasks).

The results are summarized in Table 4. First, as a foundational model, LTU-AS performs well on both audio and speech tasks. It works particularly well on tasks requiring both audio and speech understanding, which exactly meets our expectations. E.g., the accuracy of LTU-AS is nearly twice that of CLAP [48] on the zero-shot GTZAN music genre classification task; the MAE of speaker age prediction is even lower than the SOTA specialized model that only works for the task. Compared with CLIP-like models [48, 47], LTU-AS does not require any pre-defined label set and directly outputs predictions in natural language, which makes it a more practical system for real-world applications.

Third, to take a closer look at how LLaMA attends to continuous audio token input {A} and spoken text token input {S} on different tasks, we manually remove one input modality for Ablation Study 2. For most tasks, a missing modality leads to a performance drop, indicating that LLaMA takes both {A} and {S} into its decision-making. Even on audio classification and gender classification tasks where {S} is not useful, including {S} leads to only a slight performance drop, demonstrating that LTU-AS can correctly attend to {A} and {S} based on the input audio and question. Finally, we observe the ASR performance of LTU-AS (4.9% WER) is worse than its internal Whisper model (3.5% WER) due to occasionally not following instructions and changing spelling.

Table 5. LTU instruction following rate rated by GPT-4.

<table>
<thead>
<tr>
<th>Model</th>
<th>Audio Question</th>
<th>Speech Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTU-Audio Training Only</td>
<td>96%</td>
<td>69%</td>
</tr>
<tr>
<td>LTU-Speech Training Only</td>
<td>65%</td>
<td>93%</td>
</tr>
<tr>
<td>LTU-AS</td>
<td>96%</td>
<td>94%</td>
</tr>
</tbody>
</table>

5.2. Open-Ended Audio Question Answering Evaluation

In addition to the good performance on closed-ended tasks, LTU-AS also exhibits superior performance when it comes to answering open-ended questions. We quantitatively measure the instruction following rate of LTU-AS on audio and speech questions and compare it with LTU models trained with only audio data and speech data. Specifically, we use GPT-4 to generate 100 audio and speech questions based on AudioSet and VoxCeleb evaluation sets, respectively, and input the questions and corresponding audios to LTU models and collect its answers. Finally, we use GPT-4 to evaluate if the LTU model output answers the given questions by the prompt “Below is a pair of question and response. Identify if the response directly answers the question and give a clear answer.” As shown in Table 5, LTU-AS has an instruction
Table 6. Real samples of LTU–AS open-ended question answering on evaluation samples. Acoustic and speech information in the answers are marked in green and red, respectively, to highlight how LTU–AS understand them jointly.

Sample 1 Basketball Coaching (AS-Eval)
Transcribed speech: To the left, the second time through they’ll do the same move to the right. We’ll do a series of about five moves. Here we’ve got a 6’9 kid, I expect him to make the same move...
Q: What can be inferred from the audio events and spoken text? A: The audio event is a basketball bounce while the spoken text suggests that someone is giving instructions.
Q: what is the job of the speaker? A: The speaker is coaching or instructing someone on how to play basketball based on their speech content and the background sounds of basketball bounce heard.
Q: Where could this audio be recorded? A: This audio clip could have been recorded in a basketball court or gymnasium where people are playing basketball and speaking.

Sample 2 Interview (VoxCeleb2-Eval)
Transcribed speech: ... I was so anxious about getting up and talking in front of people that I thought I still sort of lose...
Q: What is the emotion of the speaker? Why? A: Based on the speech content, the speaker seems to express anxiety or nervousness about public speaking. The high pitch and fast speed suggest excitement or tension as well. Therefore, the emotion of the speaker can be inferred as nervousness or anxiety.
Q: Is the speaker a good public speaker? A: No, the speaker feels anxious about speaking in front of people and thinks they will still feel lose-lose about it.
Q: Provide the speaker 5 suggestions. A: 1. Take a deep breath before speaking. 2. Practice public speaking regularly. 3. Use humor or lightheartedness to ease nerves. 4. Focus on the audience rather than yourself during speeches. 5. Seek feedback from others to improve your delivery.

Sample 3 Rock Music (OTZAN)
Transcribed Lyrics: Gotta be tonight, my days are blue, babe Mama says you don’t, daddy says you won’t. And I’m falling off her side No way I’m gonna lose out this time. Tie your mother down...
Q: Why is the music rock? Consider both lyrics and sounds. A: The music has a strong beat, distorted guitar riffs, and aggressive vocals that are typical characteristics of rock music. The lyrics also contain themes of rebellion and defiance which align with the genre’s cultural significance.
Q: Can you suggest a title for this music? A: Tie Your Mother Down - A Punk Rock Anthem.
Q: Is this music appropriate to play in a primary school? A: No, the lyrics and overall tone are not suitable for children.
Q: What mood does this music convey? A: The music conveys a rebellious and energetic mood with its fast-paced tempo and aggressive guitar riffs. The lyrics also add to the sense of rebellion.

Sample 4 Subway Station (FSD50K)
Transcribed speech: This is the transfer point for the Metro Purple Line to Wilshire Western. Boarding platform is located downstairs...
Q: What spoken text is heard? A: The spoken text is “This is the transfer point for the Metro Purple Line to Wilshire Western. Boarding platform is located downstairs...”
Q: What can be inferred from the sound and announcement? A: The subway sound suggests that it is a busy station with many people coming and going, while the announcement provides information about the train schedule.
Q: Where is the boarding platform located? A: The boarding platform is located downstairs.
Q: Which line does this station transfer to? A: The transfer point is to the Metro Purple Line.
Q: What should I do if I hear this? A: You should board the next train and proceed to your destination.

following rate over 94% for both audio and speech questions, while LTU trained with only audio/speech dataset does not follow instruction well on questions out of its training domain. Please note that this automatic evaluation protocol may overestimate the actual instruction following rate because 1) the evaluation questions are generated by GPT-4, not humans; and 2) GPT-4 may overrate LTU–AS answers as LTU–AS is trained to generate answers similar to GPT outputs.

As shown in Table 6, LTU–AS understands the world by combing audio and speech information: In example 1, LTU–AS correctly identifies the job of the speaker as a basketball coach because the spoken text is about instructing while bouncing basketballs are heard in the background. Without understanding the spoken text, the speaker could be a basketball player, while without understanding the audio, the speaker could be a football coach. Similarly, in example 2, LTU–AS knows the speaker is anxious because of spoken content and expresses concern about public speaking while the speaker speaks fast with a high pitch. LTU–AS exhibits emerging reasoning ability and connects sounds to actions: In Sample 2, LTU–AS can provide suggestions to the speaker based on his situation; in Sample 3, LTU–AS can suggest a title for the music, and does not recommend to play it in a primary school because the lyrics and music tone are not suitable for children; in Sample 4, LTU–AS not only correctly extracts the information about the boarding platform and transfer line, but also suggests boarding the next train when we hear the announcement. All these demonstrate LTU–AS can comprehend the input audio and speech.

6. CONCLUSIONS

In this paper, we present LTU–AS, a novel joint audio and speech understanding model that can simultaneously recognize and jointly understand spoken text, speech paralinguistics, and non-speech audio events. We identify three key components in successfully building LTU–AS. First, LTU–AS uses a strong audio encoder (Whisper) and a strong reasoning model (LLaMA). The former provides precise perception ability while the latter provides advanced reasoning ability. Second, LTU–AS is trained with the new large-scale dataset Open–ASQA with a wide range of diverse audio and speech tasks. In particular, the open-ended questions generated with GPT assistance are crucial to empower LTU–AS to answer free-form questions. Third, LTU–AS is trained with a multi-stage training curriculum to mitigate hallucination. LTU–AS achieves good performance on all tested closed-ended audio and speech benchmarks, particularly on tasks requiring joint audio and speech understanding. More importantly, when answering free-form, open-ended questions, LTU–AS effectively combines information from audio and speech, and exhibits emerging reasoning abilities.

Ethics Statement: To prevent the potential misuse of the proposed audio LLM, we intentionally do not include speaker identification in our task list. The music used in model training is Creative Commons-licensed [29].
7. REFERENCES


