JOINT AUDIO AND SPEECH UNDERSTANDING

Yuan Gong[✉1]*, Alexander H. Liu*[1]*, Hongyin Luo*[1]*, Leonid Karlinsky*[2]*, James Glass*[1]

[1]MIT CSAIL, USA   [2]MIT-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 layer-wise 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]. We
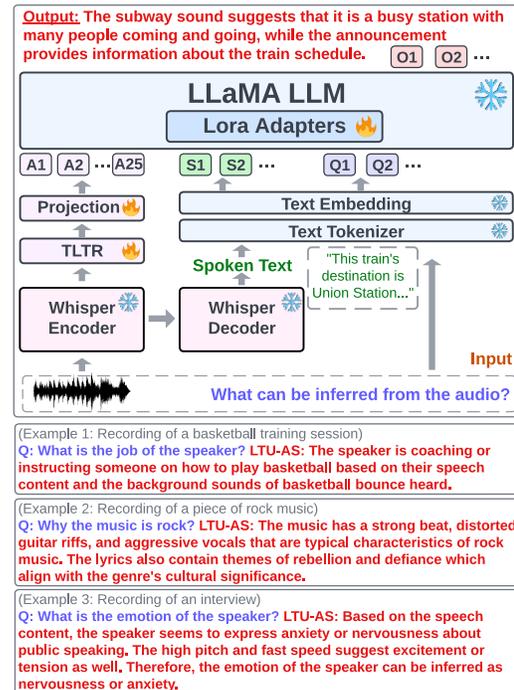
Fig. 1. Illustration of the `LTU-AS` model and real samples showing its *joint* audio and speech understanding ability.

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 recogntion [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 AQAs. 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 AQAs (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 AQA 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 AQA Generation

Generating diverse open-ended AQA 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,

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.

Closed-Ended QA Samples

Q: Identify the sound from the clip? A: Dog, Animal.
Q: Write an audio caption describing the sound?
A: An engine revving and then tires squealing.
Q: What emotion does the speaker convey?
A: Happy, with a sentiment score of 2 (positive).
Q: Can you determine the music genre? A: Folk, Pop.

Open-Ended QA Samples

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)

GPT-Input: In the recording, background sound of domestic animals, dog and speech of "Get him going, get him going with it." is heard.
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.

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.
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.

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.
GPT-Output: Q: What can we infer from the speech content and emotion? A: The speaker seems happy to introduce himself and his company, and he is delivering his words in a slow and clear manner to make sure everyone understands what he is saying.

GPT-Input: Music genre: Folk; Music Lyrics: "The man of his dreams is all that he seems Flowers in sunshine..."
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.

---

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 AQA 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 2. The statistics of the 9.6-million `Open-ASQA` dataset. "x" denotes the corresponding label is used.

Dataset	Audio Event	Audio Caption	Spoken Text*	Speaker Gender	Speaker Age	Speech Style	Speaker Emotion	Music Genre	# Audio Clips	# Closed-Ended QAs	# Open-Ended QAs
Audio Datasets (OpenAQA) [5]											
AS-Strong [30]	x	x	x	x	-	-	-	-	102k	683k	901k
AudioSet [31]	x	-	x	x	-	-	-	x	500k	538k	184k
VGGSound [32]	x	-	x	x	-	-	-	x	184k	367k	907k
FSD50K [33]	x	-	x	x	-	-	-	x	41k	82k	403k
AudioCaps [34]	x	x	x	x	-	-	-	x	46k	97k	478k
FreeSound [35]	-	x	x	-	-	-	-	-	91k	91k	791k
Clotho [36]	-	x	x	-	-	-	-	-	5k	48k	89k
Sound Bible [37]	-	x	x	-	-	-	-	-	1.2k	12k	10k
Sum									845k	1,918k	3,763k
Speech Datasets											
IEMOCAP [23]	-	-	x	x	-	x	x	-	4.3k	26k	83k
LibriTTS [25]	-	-	x	x	-	x	-	-	22k	167k	418k
VoxCeleb2 [26]	-	-	x	x	x	x	-	-	107k	194k	926k
MOSEI [24]	-	-	x	-	-	x	x	-	18k	554k	355k
Sum									151k	941k	1,784k
Joint Audio and Speech Datasets											
AudioSet [31]	x	-	x	x	-	-	-	x	82k	-	747k
FMA [29]	-	-	x	-	-	-	-	x	93k	93k	396k
Sum									175k	93k	1,143k
Total	(9,641k Question Answer Pairs)								1,089k	2,951k	6,690k

Open-Ended Speech AQA: We generate open-ended AQAs about speech using the four commonly used datasets IEMOCAP [23], MOSEI [24], LibriTTS [25], and Vox-Celeb2 [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 AQAs.

Open-Ended Joint Audio and Speech AQA

We use two datasets containing both speech and non-speech audio to generate joint audio and speech AQAs. 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 AQAs with GPT assistance. As shown in Table 1, GPT can generate AQAs 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 AQAs about music understanding with joint lyrics and melody analysis. In total, we generated about 1.1 million open-ended joint audio and speech AQAs.

Table 3. The LTU-AS training curriculum.

Stage	Tr. Params	Tr. Task	Tr. Samples	LR	# Epochs
1	Proj.	Cla.	2.1M	1e-3	2
2	Proj. + TLTR + LoRA	Cla.	2.1M	2e-4	2
3	Proj. + TLTR + LoRA	All	9.6M	2e-4	1

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 AQAs 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.

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.

Model	Audio Classif. ESC-50 [39] (ACC ↑)	Audio Caption AudioCaps (SPICE ↑)	Speech Recognition Librispeech [40] (test-clean WER ↓)	Emotion Recognition IEMOCAP (ACC ↑)	Gender Classif. Voxceleb2 (macro-F1 ↑)	Age Pred. Voxceleb2 (MAE ↓)	Music Genre Classif. GTZAN [41] (ACC ↑)
Best specialized models trained supervisedly on each dataset. Not generalizable to unseen label sets and tasks.							
Best Supervised & Specialized	97.0 [42]	17.7 [43]	1.4 [44]	70.6 [45]	98.3 [27]	9.4 [27]	93.9 [46]
CLIP-like audio-text model. Generalizable to unseen labels, but a pre-defined label set is required for inference							
AudioClip [47]	69.4	-	-	-	-	-	-
CLAP [48]	82.6	-	-	-	-	-	25.2
(Proposed) One single model for all tasks. Directly output label names, no pre-defined label set is needed at inference.							
LTU-AS	80.8 $^{ZS-}$	15.0	**4.9**	65.2	90.8	**7.3**	**50.3**^{ZS}
Ablation Study 1 - Train with only speech or audio data							
LTU (Audio Training Only) [5]	**82.8**	**17.0**	104.2	38.2	77.0	Fail*	29.8
LTU (Speech Training Only)	10.9	0.5	12.9	**69.8**	90.1	7.9	23.5
Ablation Study 2 - Inference with missing modality							
LTU-AS (Audio Input Only)	81.9	14.9	97.2	58.6	**95.6**	8.2	48.2
LTU-AS (Spoken Text Input Only)	7.7	3.5	20.0	45.4	42.0	11.9*	21.5

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.

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.

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.

Model	Audio Question	Speech Question
LTU-Audio Training Only	96%	69%
LTU-Speech Training Only	65%	93%
LTU-AS	**96%**	**94%**

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 (GTZAN)
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: What did you hear? A: Music, punk rock.   Q: Why the music is 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 did you hear? A: Subway, metro, underground.   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 militate 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

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust Speech Recognition via Large-scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *Interspeech*, 2023.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language Models are Few-shot Learners," *NeurIPS*, 2020.

[5] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass, "Listen, Think, and Understand," *arXiv preprint arXiv:2305.10790*, 2023.

[6] Yuan Gong, Yu-An Chung, and James Glass, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021.

[7] Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang, "Unified Keyword Spotting and Audio Tagging on Mobile Devices with Transformers," in *ICASSP*, 2023.

[8] Niko Moritz, Gordon Wichern, Takaaki Hori, and Jonathan Le Roux, "All-in-One Transformer: Unifying Speech Recognition, Audio Tagging, and Event Detection," in *Interspeech*, 2020.

[9] Chaitanya Narisetty, Emiru Tsunoo, Xuankai Chang, Yosuke Kashiwagi, Michael Hentschel, and Shinji Watanabe, "Joint Speech Recognition and Audio Captioning," in *ICASSP*, 2022.

[10] Joseph Turian, Jordie Shier, Humair Raj Khan, et al., "HEAR: Holistic Evaluation of Audio Representations," in *NeurIPS Competitions and Demonstrations Track*, 2021.

[11] Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson, "WavPrompt: Towards Few-Shot Spoken Language Understanding with Frozen Language Models," in *Interspeech*, 2022.

[12] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al., "AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head," *arXiv preprint arXiv:2304.12995*, 2023.

[13] Kai-Wei Chang, Yu-Kai Wang, Hua Shen, et al., "SpeechPrompt v2: Prompt Tuning for Speech Classification Tasks," *arXiv preprint arXiv:2303.00733*, 2023.

[14] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu, "SpeechGPT: Empowering Large Language Models with Intrinsic Cross-modal Conversational Abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[15] Tiantian Feng and Shrikanth Narayanan, "PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches For Speech Emotion Recognition Using Pre-trained Speech Models," *arXiv preprint arXiv:2306.05350*, 2023.

[16] Anfeng Xu, Rajat Hebbar, Rimita Lahiri, Tiantian Feng, Lindsay Butler, Lue Shen, Helen Tager-Flusberg, and Shrikanth Narayanan, "Understanding Spoken Language Development of Children with ASD Using Pre-trained Speech Embeddings," in *Interspeech*, 2023.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is All You Need," in *NeurIPS*, 2017.

[18] Wei-Lin Chiang, Zhuohan Li, Zi Lin, et al., "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality," March 2023.

[19] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio, "An Empirical Investigation of Catastrophic Forgetting in Gradient-based Neural Networks," *arXiv preprint arXiv:1312.6211*, 2013.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[21] Angela Fan, Mike Lewis, and Yann Dauphin, "Hierarchical Neural Story Generation," in *ACL*, 2018.

[22] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher, "Ctrl: A Conditional Transformer Language Model for Controllable Generation," *arXiv preprint arXiv:1909.05858*, 2019.

[23] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language resources and evaluation*, 2008.

[24] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *ACL*, 2018.

[25] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *arXiv preprint arXiv:1904.02882*, 2019.

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[27] Khaled Hechmi, Trung Ngo Trong, Ville Hautamäki, and Tomi Kinnunen, "VoxCeleb Enrichment for Age and Gender Recognition," in *ASRU*, 2021.

[28] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and Music Signal Analysis in Python," in *The 14th Python in Science Conference*, 2015.

[29] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "FMA: A Dataset For Music Analysis," in *18th International Society for Music Information Retrieval Conference*, 2017.

[30] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal, "The Benefit Of Temporally-Strong Labels In Audio Event Classification," in *ICASSP*, 2021.

[31] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An Ontology and Human-labeled Dataset for Audio Events," in *ICASSP*, 2017.

[32] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A Large-scale Audio-Visual Dataset," in *ICASSP*, 2020.

[33] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE Transactions on Audio, Speech, and Language Processing*, 2021.

[34] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating Captions for Audios in The Wild," in *NAACL*, 2019.

[35] "FreeSound Project," https://freesound.org/.

[36] Samuel Lipping, Konstantinos Drossos, and Tuoams Virtanen, "Crowdsourcing a Dataset of Audio Captions," in *DCASE*, 2019.

[37] "Sound Bible," https://soundbible.com/.

[38] Yuan Gong, Yu-An Chung, and James Glass, "PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2021.

[39] Karol J Piczak, "ESC: Dataset for Environmental Sound Classification," in *ACM-MM*, 2015.

[40] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *ICASSP*, 2015.

[41] Bob L Sturm, "The GTZAN Dataset: Its Contents, Its Faults, Their Effects on Evaluation, and Its Future Use," *arXiv preprint arXiv:1306.1461*, 2013.

[42] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *ICASSP*, 2022.

[43] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, Jaeheon Sim, Jinwoo Lee, and Kyogu Lee, "Improving Audio-Language Learning with MixGen and Multi-Level Test-Time Augmentation," *arXiv preprint arXiv:2210.17143*, 2022.

[44] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al., "BigSSL: Exploring the Frontier of Large-scale Semi-supervised Learning for Automatic Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[45] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale Self-supervised Pre-training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[46] Caifeng Liu, Lin Feng, Guochao Liu, Huibing Wang, and Shenglan Liu, "Bottom-up Broadcast Neural Network for Music Genre Classification," *Multimedia Tools and Applications*, 2021.

[47] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "AudioCLIP: Extending CLIP to Image, Text and Audio," in *ICASSP*, 2022.

[48] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP: Learning Audio Concepts from Natural Language Supervision," in *ICASSP*, 2023.