



Joint Audio and Speech Understanding



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

Interactive demo, try it!

Introduction

Background

- Conventional Audio Recognition and ASR model can perceive sounds but **cannot understand sounds**.
- LLMs already encode knowledge about audio and speech understanding but **cannot perceive sounds**.
- Audio events, speech paralinguistics, and speech content should to be understood **jointly**.

Our Goal

- Develop Audio and Speech LLM that can **perceive and understand** audio events, speech paralinguistics, and speech content **jointly**.

Key Methods

- Combine pretrained audio model and large language model
- A new GPT-Assisted method to generate open-ended training data

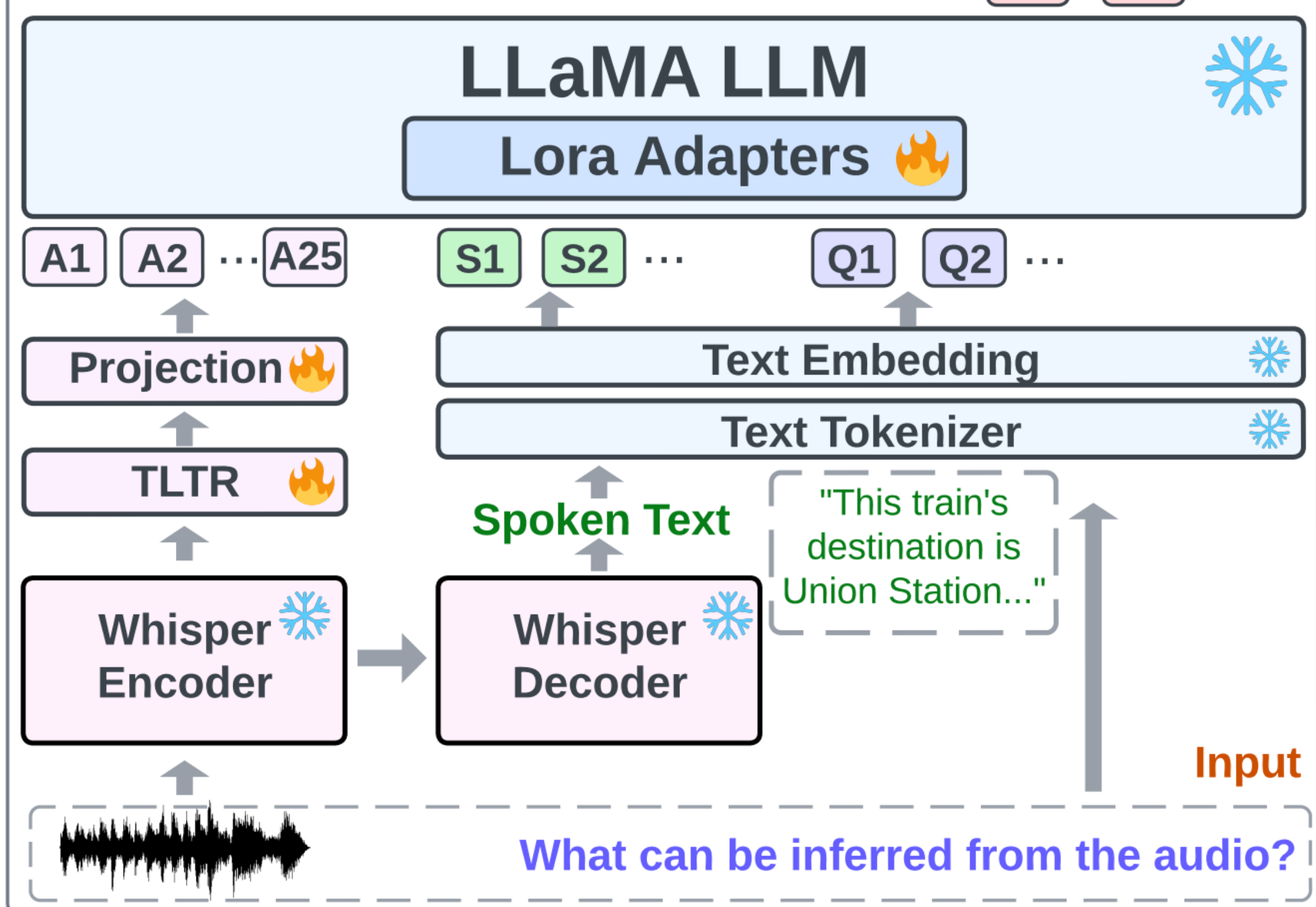
Open-Source Code and Data: <https://github.com/YuanGongND/ltu>

Model - LTU-AS

(Listen to, Think of, and Understand Audio and Speech)

- LLM** - fine-tune LLaMA 7B with (Audio, Q, A) data
- Encoding audio** - Whisper encoder encodes both audio and speech paralinguistic information [1].
- Encoding speech** - Whisper decoder transcribed speech to text, and input to LLM.

Output: 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. O1 O2 ...



- Training** - standard next token prediction; only audio encoder, projection layer, and LoRA Adapters are trained.

- Percept-to-Understand Learning curriculum**

Guide the model to attend to the audio input in early stages to mitigate hallucination.

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

[1] Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers, Interspeech 2023

Dataset – Open-ASQA

(Open-Ended Joint Audio and Speech AQA)

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
<i>Audio Datasets (OpenAQA)</i>											
AS-Strong	x	x	x	x	-	-	-	-	102k	683k	901k
AudioSet	x	-	x	x	-	-	-	x	500k	538k	184k
VGGSound	x	-	x	x	-	-	-	x	184k	367k	907k
FSD50K	x	-	x	x	-	-	-	x	41k	82k	403k
AudioCaps	x	x	x	x	-	-	-	x	46k	97k	478k
FreeSound	-	x	x	-	-	-	-	-	91k	91k	791k
Clotho	-	x	x	-	-	-	-	-	5k	48k	89k
Sound Bible	-	x	x	-	-	-	-	-	1.2k	12k	10k
Sum									845k	1,918k	3,763k
<i>Speech Datasets</i>											
IEMOCAP	-	-	x	x	-	x	x	-	4.3k	26k	83k
LibriTTS	-	-	x	x	-	x	-	-	22k	167k	418k
VoxCeleb2	-	-	x	x	x	x	-	-	107k	194k	926k
MOSEI	-	-	x	-	-	x	x	-	18k	554k	355k
Sum									151k	941k	1,784k
<i>Joint Audio and Speech Datasets</i>											
AudioSet	x	-	x	x	-	-	-	x	82k	-	747k
FMA	-	-	x	-	-	-	-	x	93k	93k	396k
Sum									175k	93k	1,143k
Total	(9,641k Question Answer Pairs)								1,089k	2,951k	6,690k

- Scale** - 9.6M Audio-QA pairs of ~1M audio clips from mixture of datasets
- Closed-ended AQAs** - Q: templates generated by GPT-3.5-Turbo; A: metadata
- Open-ended AQAs** - Q&A: GPT-3.5-Turbo generated, with metadata as prompt
- Joint Audio and Speech Understanding**: 1.2M AQAs generated from AudioSet and FMA datasets that have both audio and speech metadata

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.

GPT Prompt for Open-Ended QA Generation:

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 Generated QA Sample:

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.

Results

Closed-ended Audio/Speech Benchmarks

Model	Audio Classif.	Audio Caption	Speech Recognition	Emotion Recognition	Gender Classif.	Age Pred.	Music Genre Classif.
	ESC-50 (ACC ↑)	AudioCaps (SPICE ↑)	Librispeech (test-clean WER ↓)	IEMOCAP (ACC ↑)	Voxceleb2 (macro-F1 ↑)	Voxceleb2 (MAE ↓)	GTZAN (ACC ↑)
Best specialized models trained supervisedly on each dataset. Not generalizable to unseen label sets and tasks.							
Best Supervised & Specialized	97.0	17.7	1.4	70.6	98.3	9.4	93.9
CLIP-like audio-text model. Generalizable to unseen labels, but a pre-defined label set is required for inference							
AudioClip	69.4	-	-	-	-	-	-
CLAP	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 ^{2S}	15.0	4.9	65.2	90.8	7.3	50.3 ^{2S}
Ablation Study 1 - Train with only speech or audio data							
LTU (Audio Training Only)	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

Open-ended Audio QA (Please play with the interactive demo)

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.



Audio sample

Conclusion

We built an audio and speech LLM called LTU-AS - the first model that can **simultaneously** recognize and **jointly** understand spoken text, speech paralinguistics, and non-speech audio events - almost **everything** perceivable from audio signals.