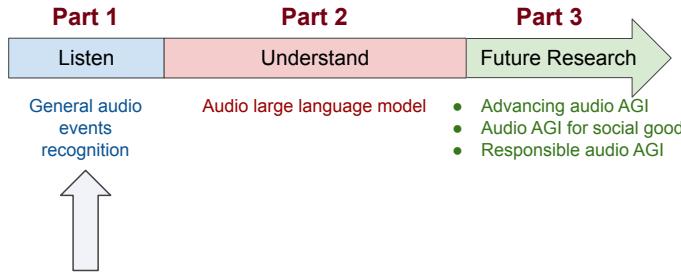


From Audio Perception to Understanding: A Path Towards Audio AGI

Yuan Gong
Research Scientist
MIT Spoken Language System Group

4/1/2024

Outline



2

Motivation

Improve Human-Computer Interaction



Enhance Human Communication



Boost Accessibility



3

History of machine listening



4

History of machine listening

But speech is just a small fraction of general sounds



5

Sample

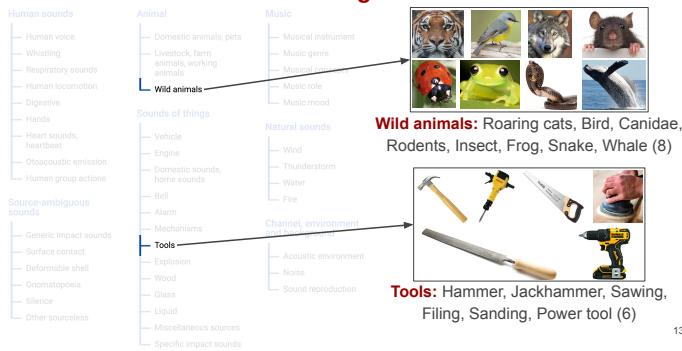


What did you hear?

You **hear** the sound of an emergency vehicle before **seeing** it

**Crucial Information but
not in speech**

There are unlimited number of general sounds



13

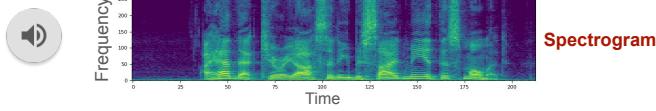
	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet

General audio classification has a larger search space

14

Speech has structured spectrogram patterns

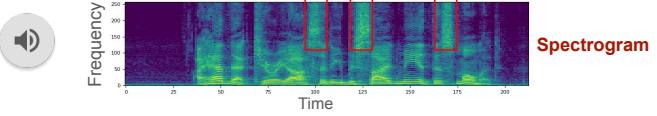
Transcript: I had that curiosity beside me at this moment



15

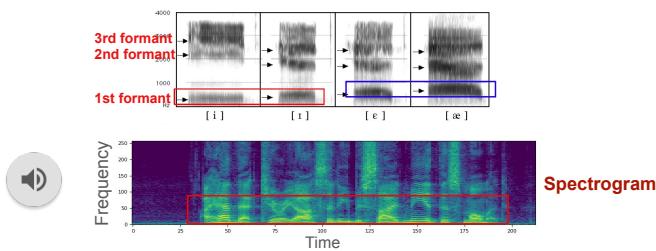
Speech has structured spectrogram patterns

Transcript: I had that curios**i**ty bes**i**de me at thi**s** moment



16

Speech has structured spectrogram patterns

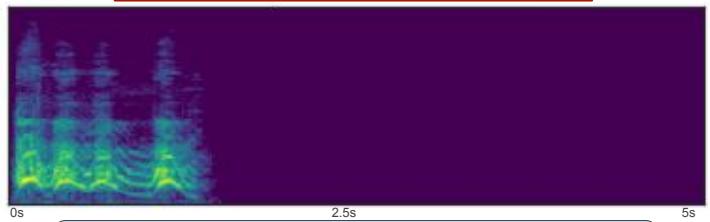


17

- Each vowel has relatively fixed formant patterns
- The duration of each phone is typically < 300ms

General audios have more diverse patterns

Class name: dog - 5-231762-A-0.wav



18

- No elementary units (e.g., phonemes)
- Can be transient-, harmonic-, and noise-like
- Duration can be up to a few seconds

Differences between speech and general audio recognition

	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet
Duration	Typically < 300ms	Up to a few seconds
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification

Data-driven method

19

Differences between speech and general audio recognition

	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet
Duration	Typically < 300ms	Up to a few seconds
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification

Capture long-range context

20

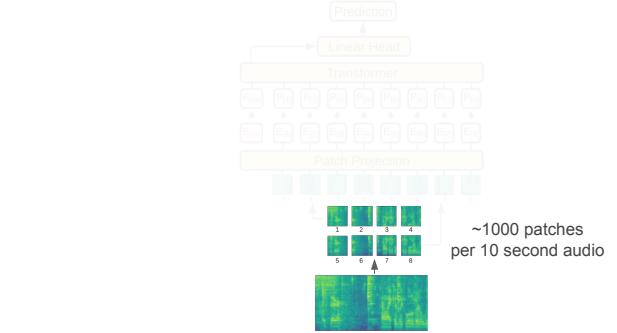
Differences between speech and general audio recognition

	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet
Duration	Typically < 300ms	Up to a few seconds
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification

Attend to multiple sounds

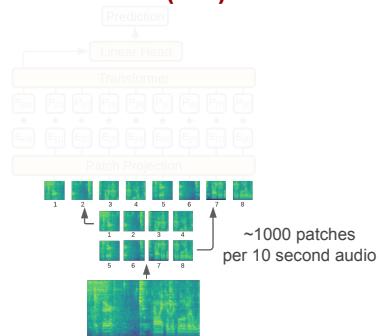
21

Audio spectrogram transformer (AST)



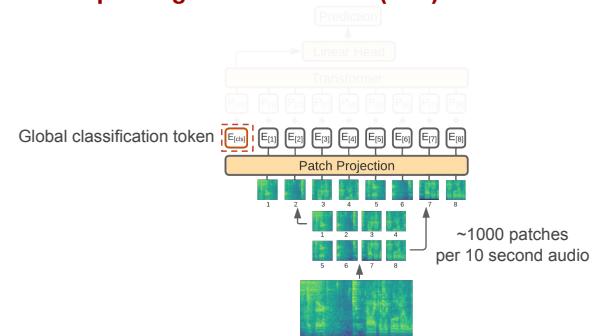
22

Audio spectrogram transformer (AST)



23

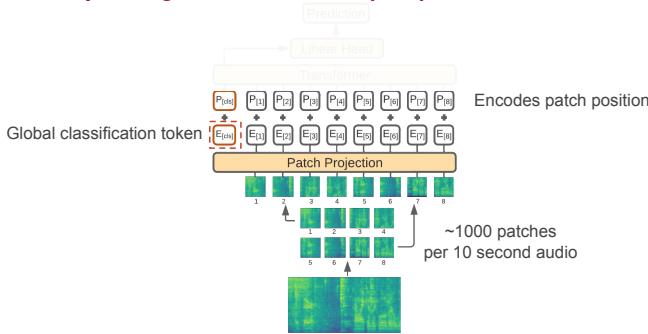
Audio spectrogram transformer (AST)



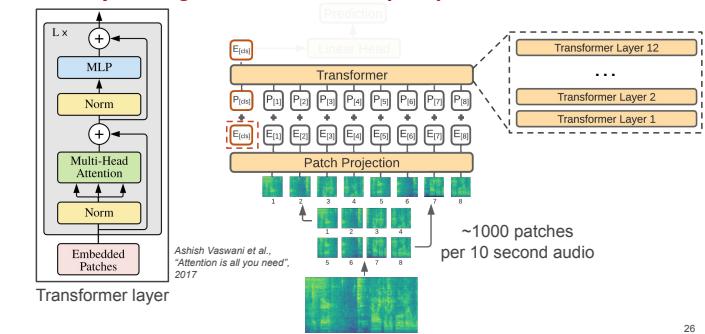
24

Yuan Gong et al., "Audio Spectrogram Transformers", Interspeech 2021.

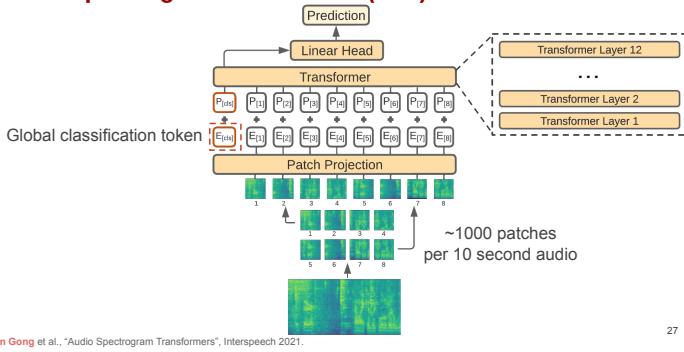
Audio spectrogram transformer (AST)



Audio spectrogram transformer (AST)



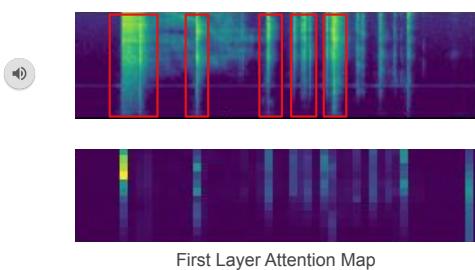
Audio spectrogram transformer (AST)



	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet
Duration	Typically < 300ms	Up to a few seconds
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification

Capture long-range context

Advantage of AST: long-range dependencies



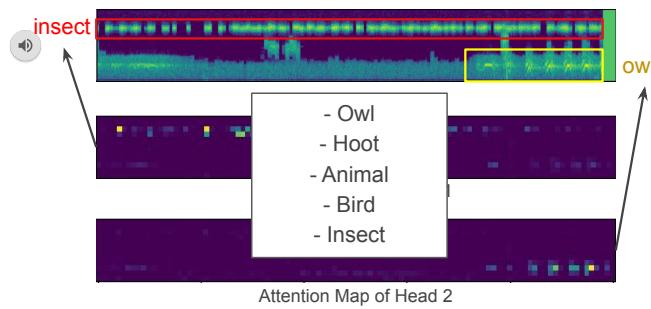
29

	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet
Duration	Typically < 300ms	Up to a few seconds
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification

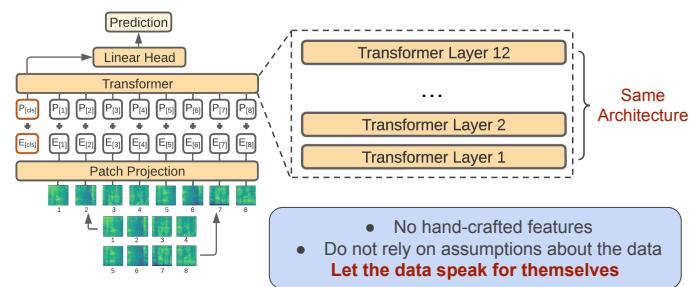
Attend to multiple sounds

30

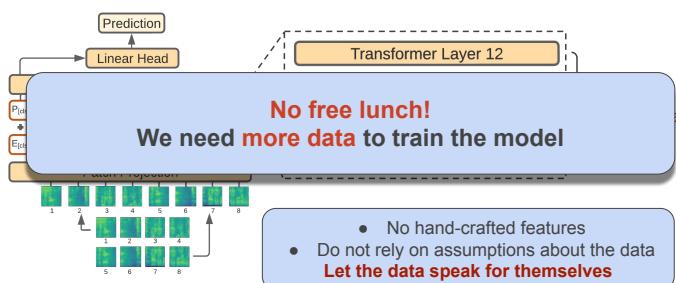
Advantage of AST: dynamic attention



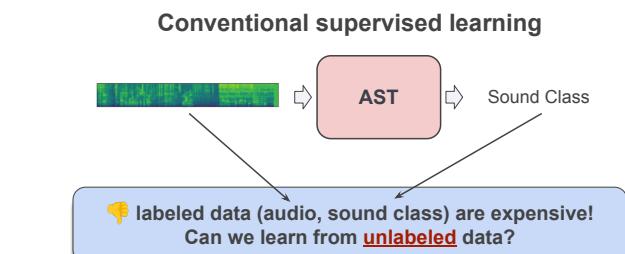
AST is a data-driven model



AST is a data-driven model



The data issue



Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data



Random masking, no human annotation required

Audio self-supervised learning

Stage 2:

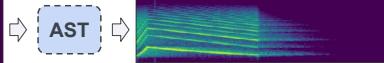
Why this work?

If the model can perfectly complete this task, what does it mean?



Self-supervised learning task (proxy task):
Predict the original spectrogram from the masked spectrogram

Audio self-supervised learning

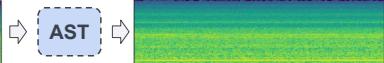


This sound features rise and fall in pitch

The model learns the **spectrogram pattern**

The model does not associate the pattern with sound class

This sound features repetitive, high-frequency tones



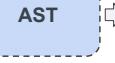
This sound features wide-band noise

37

Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data



AST

At this point, the model cannot do classification

38

Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data

AST



Weight Inheritance

Stage 2: Supervised finetuning

Small-scale labeled data

AST

Sound Class

39

Audio self-supervised learning



This sound features rise and fall in pitch



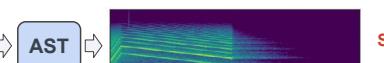
This sound features repetitive, high-frequency tones



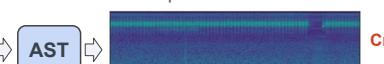
This sound features wide-band noise

40

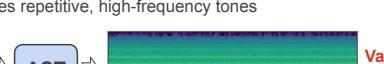
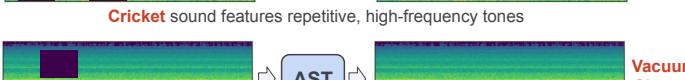
Audio self-supervised learning



Siren



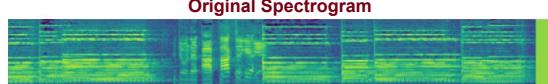
Cricket



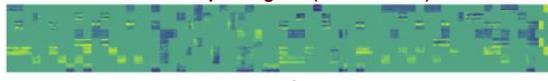
Vacuum Cleaner

41

Surprisingly strong reconstruction ability



Masked Spectrogram (75% Masked)



Reconstructed Spectrogram

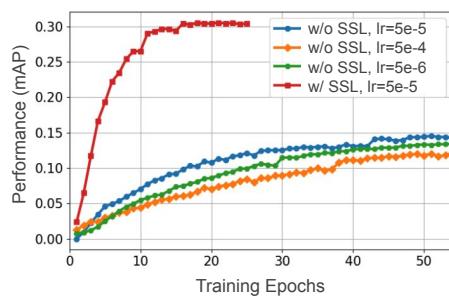


Reconstruction ability serves merely as a sanity check
Ultimately, we care about the performance in sound classification

42

Improvement led by self-supervised learning (SSL)

SSL pretrain data:
2M unlabeled audios
Fine-tuning data:
20K (1%) labeled audios



Yuan Gong et al., "Self-supervised Audio Spectrogram Transformers", AAAI 2022.

43

Unified speech and non-speech recognition



Yuan Gong et al., "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong Audio Event Taggers", Interspeech 2023.

44

Impact of AST

Task	Dataset	Model	Metric Name	Metric Value	Global Rank	Uses Extra Training Data	Result	Benchmark
Audio Tagging	AudioSet	Audio Spectrogram Transformer	mean average precision	0.485	# 1	✓	Compare	
Audio Classification	AudioSet	Audio Spectrogram Transformer	Test mAP	0.485	# 1	✓	Compare	
Audio Classification	ESC-50	Audio Spectrogram Transformer	Top-1 Accuracy	95.7	# 1	✓	Compare	
			PRE-TRAINING DATASET	AudioSet, ImageNet	# 1	✓	Compare	
			Accuracy (5-fold)	95.7	# 1	✓	Compare	
Keyword Spotting	Google Speech Commands	Audio Spectrogram Transformer	Google Speech Commands V2 35	98.11	# 1	✓	Compare	

<https://paperswithcode.com/paper/ast-audio-spectrogram-transformer>, as of 7/14/2021

45

[MIT/ast-finetuned-audioiset-10-10-0.4593](https://huggingface.co/MIT/ast-finetuned-audioiset-10-10-0.4593)

Audio Classification | Transformers | PyTorch | Safetensors

audio-spectrogram-transformer | Inference Endpoints | arxiv:2104.01778 | Li

Train Deploy Use in Transformers | Model card | Files | Community 10 | Edit model card

Downloads last month 2,226,298

Audio Spectrogram Transformer (fine-tuned on AudioSet)

>2 million model downloads per month when AST is 3 years old

<https://huggingface.co/MIT/ast-finetuned-audioiset-10-10-0.4593>, as of 3/31/2024

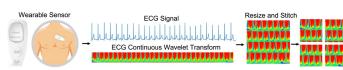
46

Impact of AST

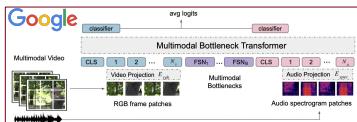
Wildlife Conservation



Bioelectrical Signal Classification



Multi-modal Systems



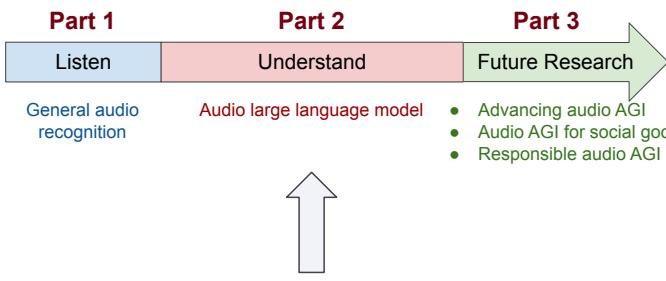
Part 1 Summary

- From speech recognition to general sound recognition
- A new model for general sound recognition - AST
- Leverage unlabeled data to enhance AST performance

47

48

Outline



49

There is one thing missing...



No **thinking** and **understanding** ability!

We only modeled **listening** ability

50

Human cognitive ability



- [Listen]** High-frequency crackling noises, intermittent low-frequency pops. perceive
- [Think]** Could be some burning or frying or arcing sound.
- [Think]** Combined with the speech, it is a cooking sizzling sound.
- [Understand]** A hands-on cooking class.
- [Emotion, Action]** Hungry, what should I eat for lunch? understand

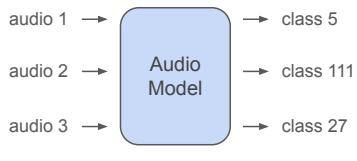
51

Human cognitive ability

- [Listen]** High-frequency crackling noises, intermittent perceive
- We do more than just listen; we also comprehend sounds.
 - Our comprehension integrates speech and non-speech sounds.
- Can we make AI having a similar understanding ability?**
- [Emotion, Action]** Hungry, what should I eat for lunch? understand

52

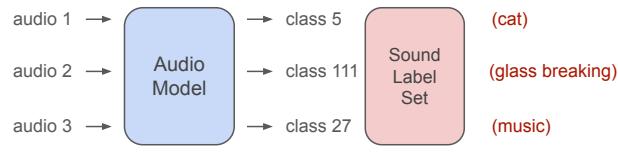
Conventional training won't work



"Listening" model only knows class indices, not sounds

53

Conventional training won't work



"Listening" model only knows class indices, not sounds

54

Conventional training won't work

"cat's meow is soft" knowledge about sound



(cat)

(glass breaking)

(music)

what we need to model
additionally

"Listening" model only knows class indices, not sounds

Conventional training won't work

"cat's meow is soft" knowledge about sound



common sense

(cat)

(glass breaking)

(music)

what we need to model
additionally

"Listening" model only knows class indices, not sounds

55

56

Conventional training won't work

"cat's meow is soft"

"cat may cause glass to break"



knowledge about sound

common sense

(cat)

(glass breaking)

(music)

what we need to model
additionally

"Listening" model only knows class indices, not sounds

Conventional training won't work

Text-based large language models can model these

audio 3 --> class 27

knowledge about sound

common sense

(cat)

(glass breaking)

(music)

what we need to model
additionally

"Listening" model only knows class indices, not sounds

57

58

Pure language LLMs already have knowledge about sounds

YU What is the acoustic feature of dog barking?



The acoustic features of dog barking vary by breed, individual dog, context, and emotion conveyed. However, several common acoustic features can be identified in dog barks:

1. **Frequency:** Dog barks can range widely in frequency. Small breeds tend to have higher-pitched barks, while larger breeds generally produce lower-pitched barks. The fundamental frequency (or pitch) of a dog's bark can range from as low as 160 Hz for large dogs to as high as 2632 Hz for small dogs.
2. **Duration:** The duration of individual barks can vary, but a typical single bark lasts between 0.1 to 1 second.
3. **Harmonic Structure:** Like human voices, dog barks contain harmonics or overtones. The relationship and amplitude of these harmonics compared to the fundamental frequency can contribute to the unique sound of each bark.

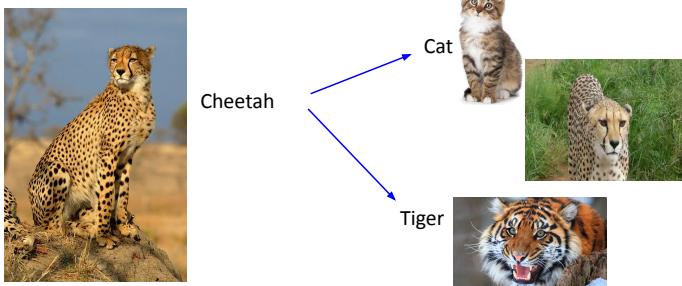
Pure language LLMs already have knowledge about sounds

4. **Temporal Pattern:** Dogs can vary the rate and pattern of their barks. Rapid barking might indicate alarm or urgency, while sporadic barking might be more investigatory or attention-seeking.
5. **Amplitude (Loudness):** The loudness of a bark can convey different emotions or intents. A loud bark might be used for warnings or threats, while softer barks might be more inquisitive or playful.
6. **Tonal Quality:** Some barks are more tonal (having a clear pitch), while others might be more noisy or harsh. This can also contribute to the emotional content of the bark.
7. **Modulation:** Barks can have variations in pitch and volume. A bark that rises in pitch might convey a question or be more attention-seeking, while one that falls might be more commanding or assertive.
8. **Context-specific Characteristics:** Dogs can produce different barks in different contexts. For instance, a bark at a stranger might be different in quality and pattern from a bark when the dog is left alone or hears a peculiar noise.

59

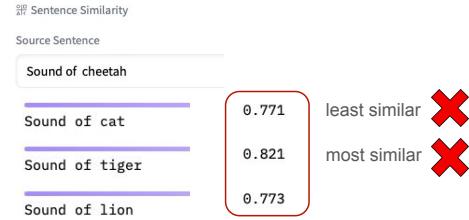
60

Pure language LLMs already have knowledge about sounds



61

"Small" language model (LaBSE) gives wrong answer



Understandable, this needs complex reasoning:

"Roaring cats (lions, tigers, jaguars, and leopards) have an incompletely ossified hyoid, which in theory allows them to roar but not to purr. Cheetahs belong to the purring cats subfamily and as such do not roar."

62

Pure language LLMs already have knowledge about sounds

YU Is Cheetah sound more closer to cat sound or tiger sound? Why?

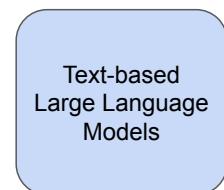
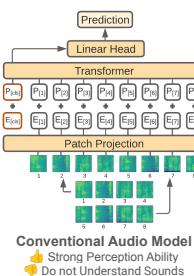
When comparing the sounds of cheetahs to those of domestic cats and tigers, the vocalizations of cheetahs are more akin to domestic cats. Here's a breakdown:

1. Purring:

- Cheetah:** Cheetahs are one of the few big cats that purr in a manner similar to domestic cats. Their purring is continuous, occurring during both inhalation and exhalation.
- Domestic Cat:** Domestic cats purr in much the same way, during both inhalation and exhalation.
- Tiger:** While tigers can produce a purring sound, it's different from the continuous purr of cheetahs and domestic cats. It's generally not as sustained and tends to occur mainly during the exhalation.

63

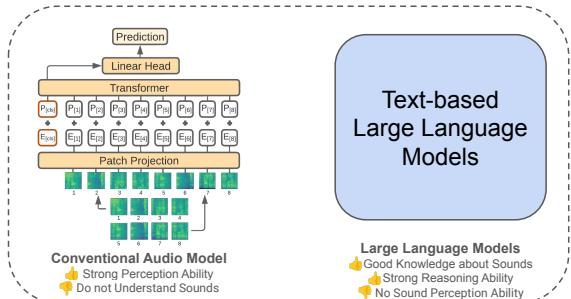
Bridging sound perception and understanding



Large Language Models
Good Knowledge about Sounds
Strong Perception Ability
Strong Reasoning Ability
No Sound Perception Ability

64

Bridging sound perception and understanding



65

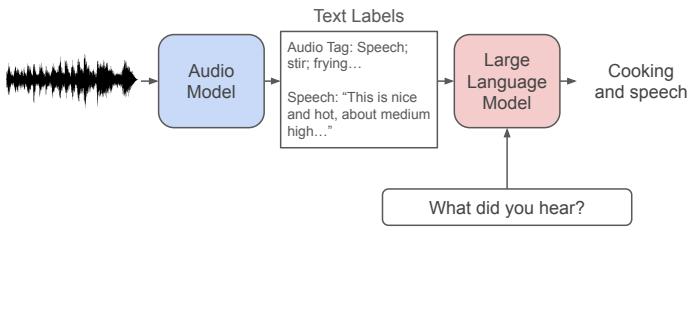
Bridging sound perception and understanding

Listen, Think, and Understand (LTU)

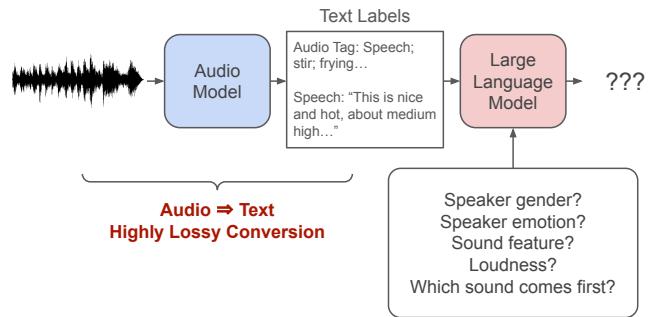
Yuan Gong et al., "Listen, Think, and Understand", ICLR 2024
Yuan Gong et al., "Joint Audio and Speech Understanding", ASRU 2023 (Best Paper Finalist)

66

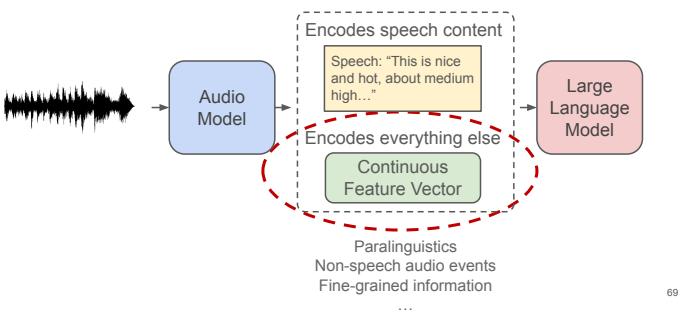
Can we just concatenate an audio model and an LLM?



Can we just concatenate an audio model and an LLM?



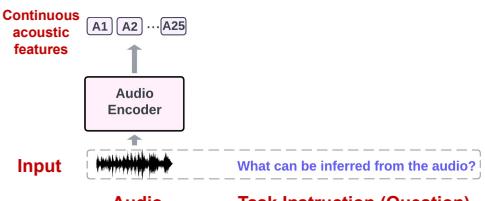
Continuous acoustic features are needed



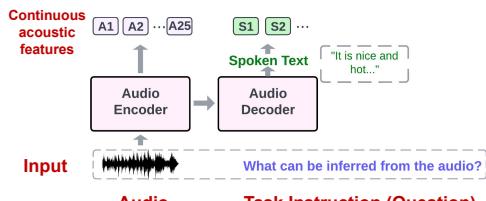
LTU model architecture



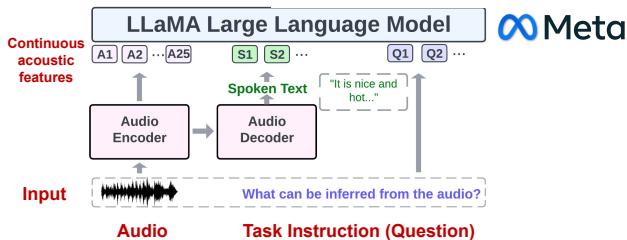
LTU model architecture



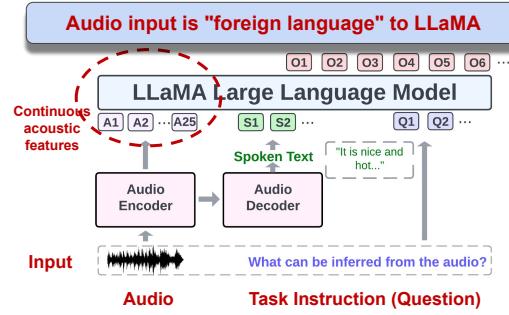
LTU model architecture



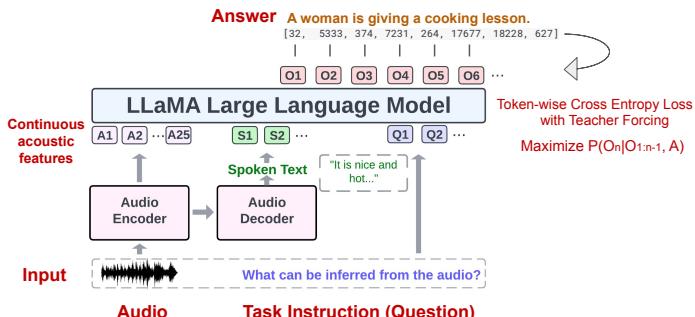
LTU model architecture



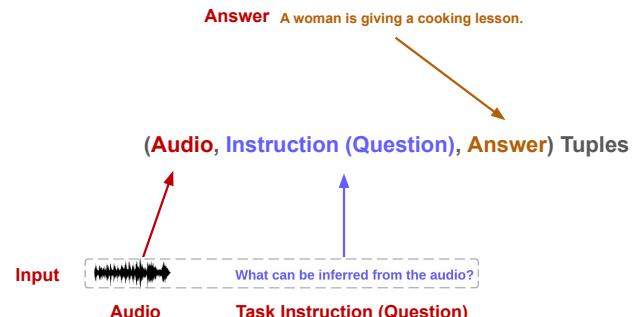
LTU model architecture



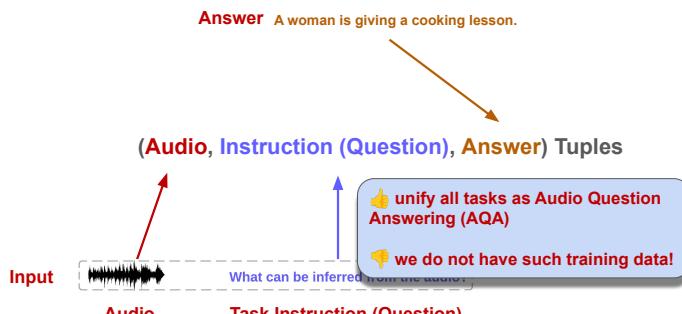
LTU training objective



LTU training data



LTU training data



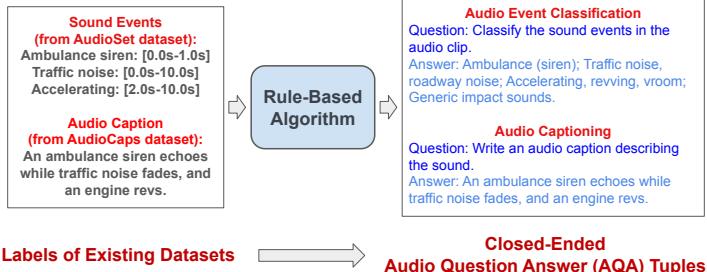
Closed-ended AQAs are easy to generate

Sound Events
(from AudioSet dataset):
Ambulance siren: [0.0s-1.0s]
Traffic noise: [0.0s-10.0s]
Accelerating: [2.0s-10.0s]

Audio Caption
(from AudioCaps dataset):
An ambulance siren echoes while traffic noise fades, and an engine revs.

Labels of Existing Datasets

Closed-ended AQAs are easy to generate



79

If we train LTU with only closed-ended AQAs



On Seen Tasks:

Audio Event Classification

Question: Classify the sound events in the audio clip.

Answer: Female speech, woman speaking; Stir; Cutlery; silverware; Dishes, pots, and pans; Breathing.

Audio Captioning
Question: Write an audio caption.
Answer: A woman speaks while frying food in a pan with background noise from a kitchen or dining room.

Looks Good!

80

If we train LTU with only closed-ended AQAs



On Seen Tasks:

Audio Event Classification
Question: Classify the sound events in the audio clip.
Answer: Female speech, woman speaking; Stir; Cutlery, silverware; Dishes, pots, and pans; Breathing

Audio Captioning
Question: Write an audio caption.
Answer: A woman speaks while frying food in a pan with background noise from a kitchen or dining room.

Looks Good!

But on unseen tasks:

Question: Describe the acoustic features you heard.

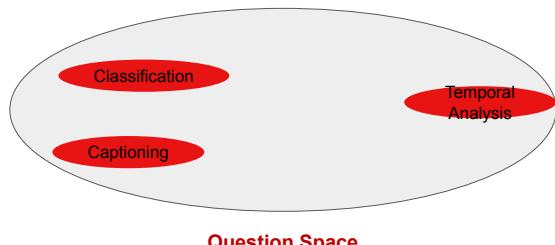
Answer: Female speech, woman speaking; Stir; Cutlery, silverware; Dishes, pots, and pans; Breathing

Question: What is the gender of the speaker?
Answer: A woman speaks while frying food in a pan with background noise from a kitchen or dining room.

... The model cannot generalize!

81

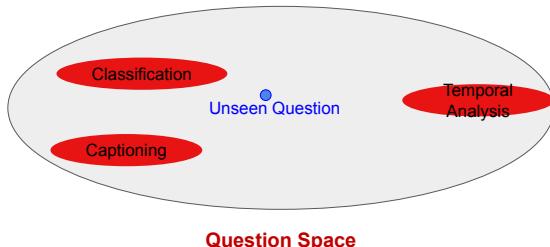
Closed-ended AQAs cannot train a generalized model



Question Space

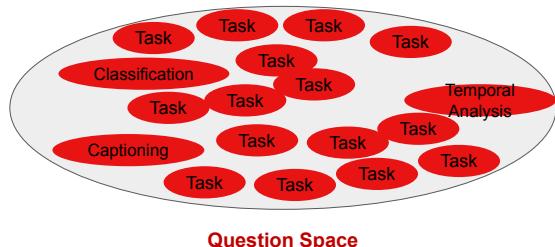
82

Closed-ended AQAs cannot train a generalized model



83

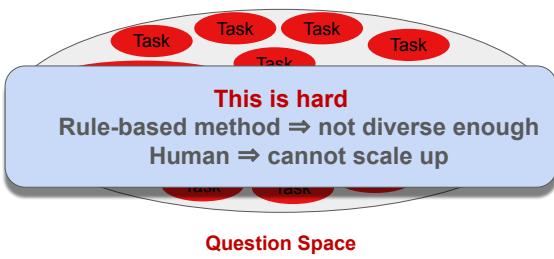
Ideally...



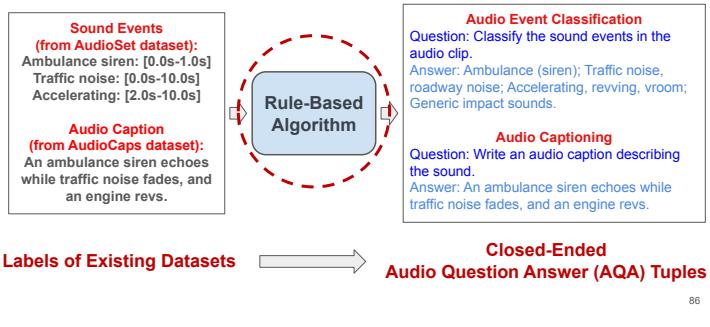
Question Space

84

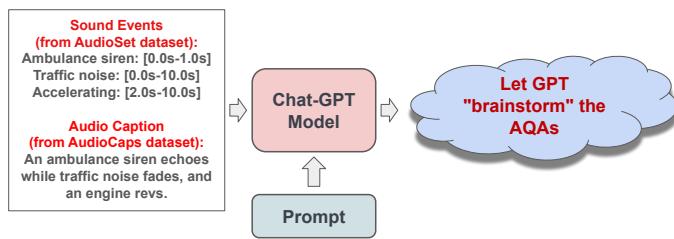
Ideally...



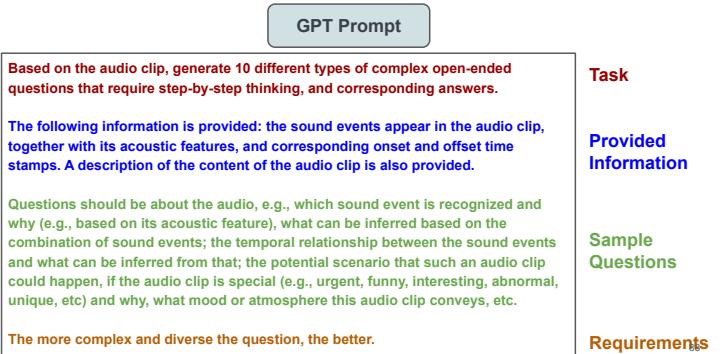
Rule-based algorithm cannot generate open-ended AQAs



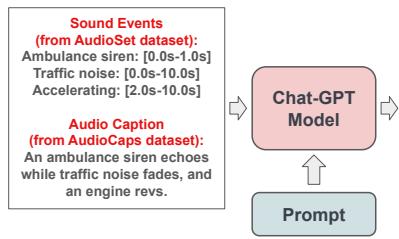
Audio instruction generation with GPT assistant



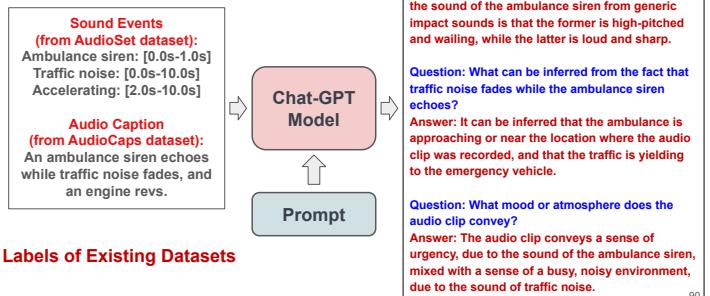
A closer look at the GPT prompt



Audio instruction generation



Audio instruction generation



89

90

Question: What is the acoustic feature that distinguishes the sound of the ambulance siren from the generic impact sounds?

Answer: The acoustic feature that distinguishes the sound of the ambulance siren from generic impact sounds is that the former is high-pitched and wailing, while the latter is loud and sharp.

Sound

low-level

abstract

91

Question: What is the acoustic feature that distinguishes the sound of the ambulance siren from the generic impact sounds?

Answer: The acoustic feature that distinguishes the sound of the ambulance siren from generic impact sounds is that the former is **high-pitched and wailing**, while the latter is **loud and sharp**.

Not in original label
Generated based on ChatGPT's knowledge

Sound

low-level

abstract

92

Question: What is the acoustic feature that distinguishes the sound of the ambulance siren from the generic impact sounds?

Answer: The acoustic feature that distinguishes the sound of the ambulance siren from generic impact sounds is that the former is high-pitched and wailing, while the latter is loud and sharp.

Question: What can be inferred from the fact that traffic noise fades while the ambulance siren echoes?

Answer: It can be inferred that the ambulance is approaching or near the location where the audio clip was recorded, and that the **traffic is yielding to the emergency vehicle**.

Sound

low-level

Understand

abstract

93

Question: What is the acoustic feature that distinguishes the sound of the ambulance siren from the generic impact sounds?

Answer: The acoustic feature that distinguishes the sound of the ambulance siren from generic impact sounds is that the former is high-pitched and wailing, while the latter is loud and sharp.

Question: What can be inferred from the fact that traffic noise fades while the ambulance siren echoes?

Answer: It can be inferred that the ambulance is approaching or near the location where the audio clip was recorded, and that the traffic is yielding to the emergency vehicle.

Question: What mood or atmosphere does the audio clip convey?

Answer: The audio clip conveys a sense of urgency, due to the sound of the ambulance siren, mixed with a sense of a busy, noisy environment, due to the sound of traffic noise.

Sound

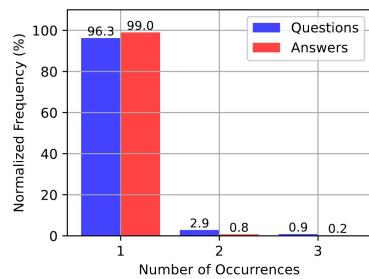
low-level

Understand

abstract

94

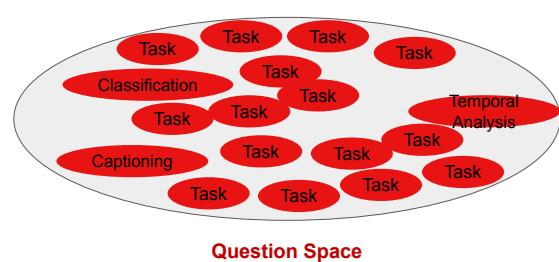
Generated open-ended questions are diverse



96.3% questions and 99.0% answers appears only once

95

Ideally...



96

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 (OpenQA) [5]</i>											
AS-Strong [30]	x	x	x	x	x	x	x	x	102k	683k	901k
AudioSet [31]	x	x	x	x	x	x	x	x	506k	538k	184k
VGGSound											307k
FSD50k											403k
AudioCaps											478k
FordCE											791k
CelebE											89k
SoundB											10k
Sum											3,763k
<i>Spoken</i>											
IEMOCAP											83k
LibriTTS											418k
VoxCeleb											926k
MOSI											355k
Sum											1,784k
<i>Joint</i>											
AudioSet [31]	-	-	x	x	-	-	-	x	85k	747k	97
FMA [29]	-	-	-	x	-	-	-	x	93k	93k	396k
Sum	-	-	-	-	-	-	-	x	175k	93k	1,143k
Total	(9,641k Question Answer Pairs)								1,089k	2,951k	6,690k

9.6 million AQAs in total

- **3.0 million** closed-ended AQAs generated by rule-based method
- **6.6 million** open-ended AQAs generated by GPT-assisted method
- Using **13** existing datasets
- **1 million** audio clips

Revisit the cooking demo

Question: What is the gender of the speaker? Why?

Answer: The speaker is female because her voice has a higher pitch than most male voices.



Question: Describe the acoustic features you heard.

Answer: The audio clip contains the sound of frying food in a small room with high reverberation and low frequency sounds.

Question: What do you feel when you hear this?

Answer: The sound of frying food can evoke hunger or nostalgia for some people, while others may find it unpleasant or annoying.

98

Demo of environment understanding



Question: What is the weather?

Answer: It is likely to be windy or rainy due to the sound of the wind and rain falling.

Question: Where could this audio be recorded?

Answer: This audio could be recorded in a busy urban area.

Question: What should I do when I hear this, why?

Answer: When you hear this, you should move to the side of the road and stay there until the emergency vehicle has passed. This is because the sound of an emergency vehicle is a warning signal that something is happening and you should stay out of the way to avoid accidents or injuries.

99

Demo of emotion recognition

Question: What is the emotion of the speaker? Why?

Answer: 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.



100

Demo of emotion recognition



Question: What is the emotion of the speaker? Why?

Answer: 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.

Speech content

Acoustic feature

101

Demo of emotion recognition



Question: What is the emotion of the speaker? Why?

Answer: 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.

Question: Provide the speaker 5 suggestions.

Answer: 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.

102

Demo of music understanding

Question: What did you hear? Answer: Music, punk rock.

Question: Why the music is rock? Consider both lyrics and sounds.

Answer: 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.

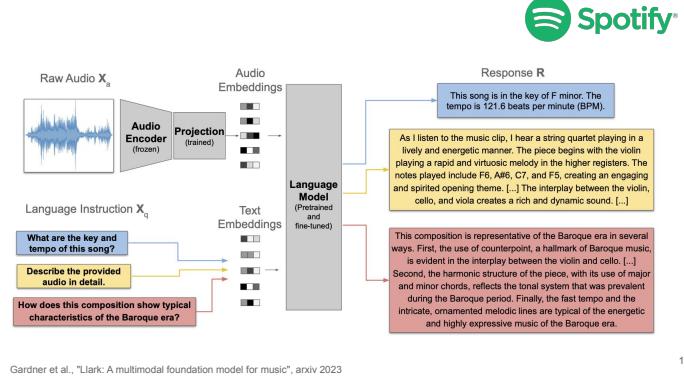
Question: Is this music appropriate to play in a primary school?

Answer: No, the lyrics and overall tone are not suitable for children.

Question: What mood does this music convey?

Answer: 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.

103



104

Demo of information extraction

Question: What did you hear?

Answer: Subway, metro, underground.

Question: What spoken text is heard?

Answer: The spoken text is "This is the transfer point for the Metro Purple Line to Wilshire Western. Boarding platform is located downstairs..."

Question: Which line does this station transfer to?

Answer: The transfer point is to the Metro Purple Line.

Question: What should I do if I hear this?

Answer: You should board the next train and proceed to your destination.

105

Demo of unseen language

Question: 你听到了什么？(What did you hear?) [Instruction in unseen language]
Answer: A train is moving and a man is speaking.

Question: Can you translate the spoken text into Chinese? [Unseen task]

Answer: The spoken text is 'Transfer point for the Metropolitan Purple Line to Busselton West. Boarding platform is located downstairs. This train's destination is Union Station. Please proceed to the boarding platform and present your ticket at the fare gate when boarding.' The translation would be '大都会绿色线到布塞尔东。车站位于下层。这次列车的目的地是联邦站。请在购票门前准备出票并进入车站。'

Question: 这班车的终点站在哪里？用中文作答。(What is the destination?)

Answer in Chinese) [Unseen task and language]
Answer: This train's final destination station is in Union Station. 这班车的终点站在 Union Station。

106

Quick Demo of Listen, Think, and Understand (LTU)



LTU is a new audio model that bridges audio perception and advanced reasoning, it can answer any open-ended question about the given audio. [Paper] [Github]

Note LTU is not an ASR and has limited ability to recognize the speech content, it focuses on general audio perception and understanding.

Input an audio and ask questions! Audio will be converted to 16kHz and padded or trim to 10 seconds. Don't have an audio sample on hand? Try some samples from AudioSet evaluation set: [sample audios from AudioSet evaluation set]

Drop Audio Here
- or -
Click to Upload

question
Why this audio is fun?

Original input audio length 10.03 seconds, number of channels: 1, sampling rate: 16000.

LTU Output

Instruction:
Why this audio is fun?

Response:
The audio is fun because it contains laughter and splashing sounds, which are associated with joy, happiness, and relaxation.

Flag

107

Formal evaluation

Closed-Ended Tasks

Evaluated on 15 major benchmarks

- **23.6%** improvement over previous SOTA
- **First** generative model

Please listen to the given 10-second audio and answer the questions. All answers need to be in English.

Answer 1: A car is driving on a road. (Q: What sound events are in the audio clip?)

Answer 2: The sound events in the audio clip consist of a horn and potentially dangerous atmosphere, with an explosion and gunfire suggesting potential conflict. (Q: What sound events are in the audio clip?)

Did this answer directly address the question? (Regardless of whether the answer is correct or not) Yes No

If asked "Is this answer correct?", please explain which part is wrong.

Please answer the following two questions.

Question: What atmosphere is most in the sound events in the audio clip convey? Correct Partially Correct Wrong

Answer 1: The sound events in the audio clip convey a sense of potentially dangerous atmosphere, with an explosion and gunfire suggesting potential conflict. (Q: What atmosphere is most in the sound events in the audio clip convey?)

Answer 2: The audio clip conveys a sense of danger, noise, speech, explosive sounds, and artillery fire. (Q: What atmosphere is most in the sound events in the audio clip convey?)

Which answer is better? Answer 1 is better Answer 2 is better

Please answer the question based on the audio you heard (at least 5 words).

Ask a few questions based on the audio and provide the corresponding answer.

Question (at least 5 words):

Answer (at least 5 words):

108

Open-Ended Tasks

Subjective Evaluation (Amazon Mechanical Turk)

- **476** independent human evaluators
- **82.9%** correct rate
- **74.9%** of human evaluators rate LTU answers are better than human-crafted answers.

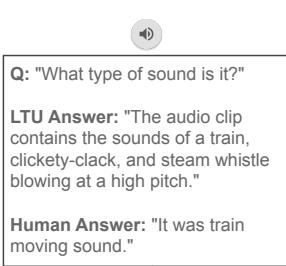
Formal evaluation

Closed-Ended Tasks
Evaluated on 15 major benchmarks

- 23.6% improvement over previous SOTA
- First generative model

Open-Ended Tasks
Subjective Evaluation (Amazon Mechanical Turk)

- 476 independent human evaluators
- 82.9% correct rate
- **74.9%** of human evaluators rate LTU answers are better than human-crafted answers.



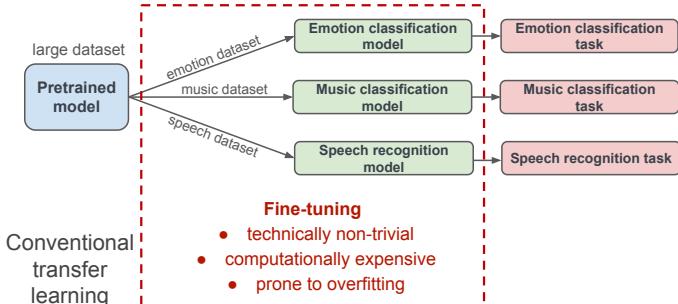
109

Advantage of LTU - 1/ understanding

LTU jointly understands
spoken text
speech paralinguistics (e.g., emotion)
non-speech audio events

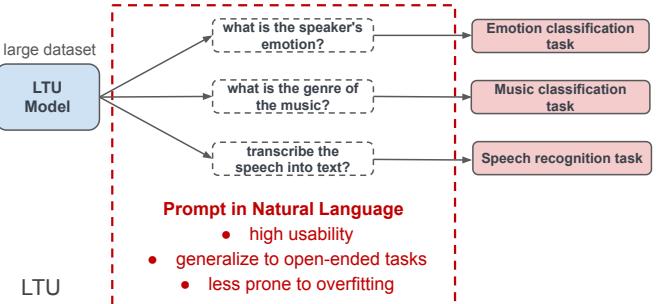
almost **everything** in audio

Advantage of LTU - 2/ broader capacity



111

Advantage of LTU - 2/ broader capacity



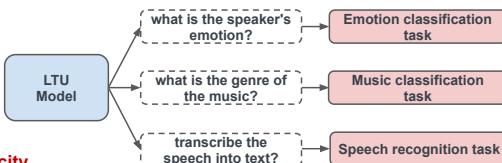
112

LTU is a step closer to artificial general intelligence (AGI)

1/ Understanding

LTU jointly understands
spoken text
speech paralinguistics (e.g., emotion)
non-speech audio events

2/ Broader capacity



113

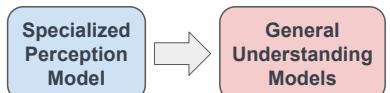
From a broader perspective...

There are many signals:

- **Audio signals**
- Electrical signals
- Environmental signals
- Biosignals
- Vision signals

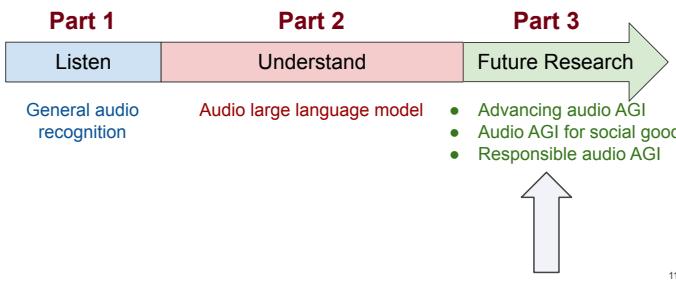
and many others ...

For many natural signals, we desire:



114

Outline



115

Future Research # 1 Advancing Audio AGI

116

Further unify audio generation



We modeled *listening and understanding*

We still left **sound generation** ability

117

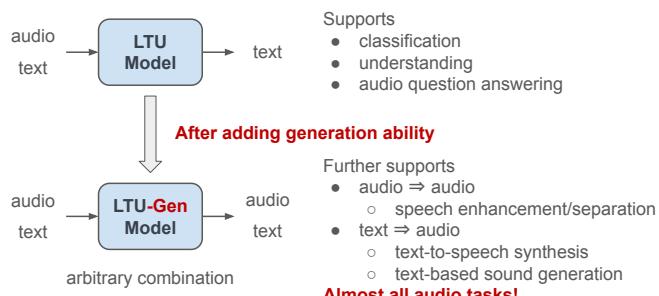
1/ Supports all audio tasks



Potential Funding Sources: NSF-IIS / Industry Sponsorship and Award

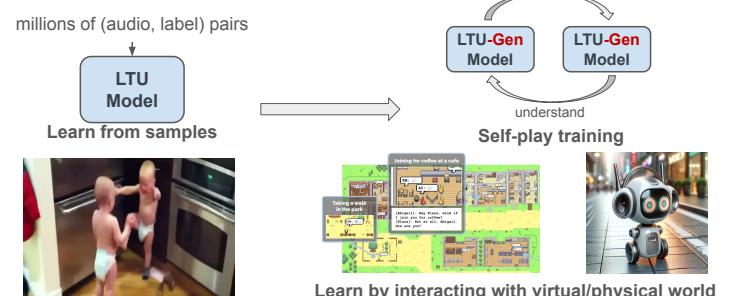
118

1/ Supports all audio tasks



119

2/ Reforming audio AI training



120

Future Research # 2 Audio AGI for Health

121

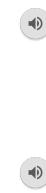
Audio AI for Health



- Speech is usually an **early sign** of Autism Spectrum Disorder, Parkinson's Disease, Frontotemporal Dementia, and many other diseases
- Speech can serve as a **non-invasive, convenient, and low-cost biomarker**

122

Audio AI for Health



**Which is abnormal?
Why?**

The second recording is abnormal

- Slurred speech
- Irregular rhythm

123

Amyotrophic Lateral Sclerosis (ALS)



2018 November

2019 June

7 months

124

Amyotrophic Lateral Sclerosis (ALS)

motor neurons gradually degenerate
⇒
loss of voluntary muscle movement
⇒
difficult to articulate words
⇒
speech becomes slurred and slow

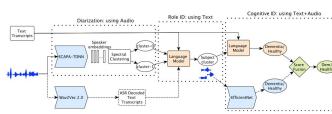


125

Previous work



Child Autism (ASD) Detection



Dementia Detection



Emotional Disorder Detection

Yuan Gong et al., "Top Model Based Multi-Modal Emotion Detection", ACM MM: Audio-Visual Emotion Challenge and Workshop, 2017. (Challenge Winner)

Yuan Gong et al., "Automatic Autism Spectrum Disorder Detection Using Everyday Vocalizations Captured by Smart Devices", ACM-BBCB 2016.

Yuan Gong et al., "Continuous Assessment of Children's Emotional States Using Acoustic Analysis", ICHI 2017.

Nauman Dewalabadi, Yuan Gong et al., "Detecting Dementia from Long Neuropsychological Interviews", Findings of EMNLP 2022.

126

Conventional AI

Prediction

- Small speech health datasets ⇒ **Bad generalization**
- Only prediction is given ⇒ **Not transparent**

"ALS speech is slurred"

Potential Funding Sources: NIH / Pharmaceutical Companies

127

Conventional AI

Prediction

- Small speech health datasets ⇒ **Bad generalization**
- Only prediction is given ⇒ **Not transparent**

We input less (tiny dataset) to the model
and expect more (reliability) from it!

"ALS speech is slurred"

Potential Funding Sources: NIH / Pharmaceutical Companies

128

Do we have to fully rely on speech health datasets?

Recall that:

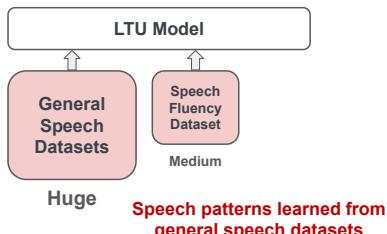
- We can identify which recording is abnormal **without** knowledge of ALS
- After learning about ALS, we can **associate** speech patterns with ALS

Humans not need learn from a large speech health dataset

Potential Funding Sources: NIH / Pharmaceutical Companies

129

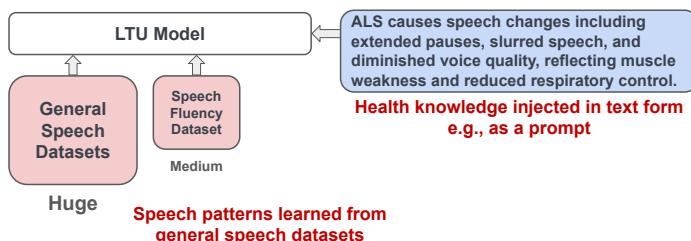
Audio AGI for health



Potential Funding Sources: NIH / Pharmaceutical Companies

130

Audio AGI for health



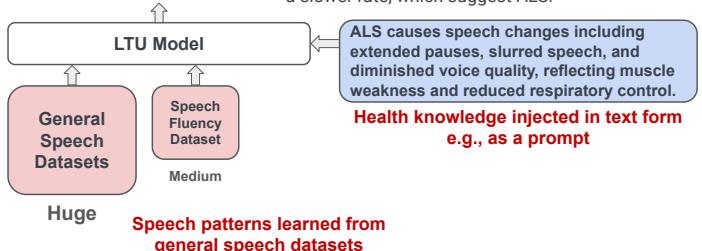
Potential Funding Sources: NIH / Pharmaceutical Companies

131

Audio AGI for health

Prediction and Explanation

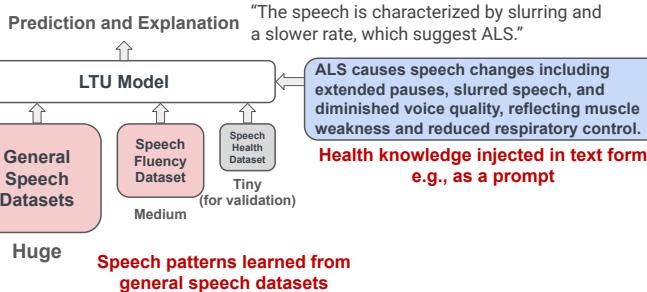
"The speech is characterized by slurring and a slower rate, which suggest ALS."



Potential Funding Sources: NIH / Pharmaceutical Companies

132

Audio AGI for health



Potential Funding Sources: NIH / Pharmaceutical Companies

133

Future Research # 3 Responsible Audio AGI

Responsible audio AGI - security

AP AP News
<https://apnews.com/article/new-hampshire-primary-bl...>

Fake Biden robocall being investigated in New Hampshire

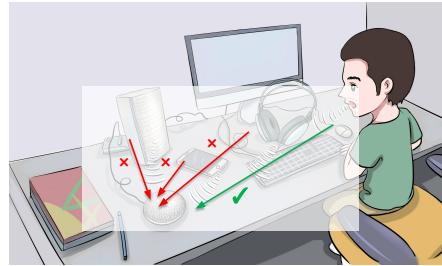
Jan 22, 2024 — New Hampshire officials are investigating reports of an apparent robocall that used AI to mimic President Biden's voice before the primary ...



Potential Funding Sources: NSF-SaTC / DoD / DoJ

135

Responsible audio AGI - security



Only speech from a **live speaker** should be accepted as **legitimate**

Yuan Gong et al., "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems", Interspeech 2019. (Best Student Paper Nomination)

136

Responsible audio AGI - security

AP AP News
<https://apnews.com/article/new-hampshire-primary-bl...>

Fake Biden robocall being investigated in New Hampshire

Jan 22, 2024 — New Hampshire officials are investigating reports of an apparent robocall that used AI to mimic President Biden's voice before the primary ...



- Audio watermarking
- Deepfake detection
- Adversarial attacks and data poisoning
- Policies for responsible use of audio AI

Potential Funding Sources: NSF-SaTC / DoD / DoJ

137

Responsible audio AGI - fairness and trustworthiness

Fairness

- Existing models are mostly English-centric
- Models may bias against minority groups

Trustworthiness

- Hallucination
- Not good at tasks requiring high reliability

Potential Funding Sources: NSF-SaTC / DoD / DoJ

138

Revisit this sample

Question: 你听到了什么? (What did you hear?) [Instruction in unseen language]
Answer: A train is moving and a man is speaking.

Question: Can you translate the spoken text into Chinese? [Unseen task]
Answer: The spoken text is 'Transfer point for the Metropolitan Purple Line to Busselton West. Boarding platform is located downstairs. This train's destination is Union Station. Please proceed to the boarding platform and present your ticket at the fare gate when boarding.' The translation would be '替换站点为大都市绿色线到布塞尔东。车站位于下层。这次列车的目的地是联邦站。请在购票门前准备出票并进入车站。'

Question: 这班车的终点站在哪里? 用中文作答。 (What is the destination?)
Answer in Chinese) [Unseen task and language]
Answer: This train's final destination station is in Union Station. 这班车的终点站在 Union Station.

139

Revisit this sample

Question: 你听到了什么? (What did you hear?) [Instruction in unseen language]
Answer: A train is moving and a man is speaking.

Question: Can you translate the spoken text into Chinese? [Unseen task]
Answer: The spoken text is 'Transfer point for the Metropolitan Purple Line to Busselton West. Boarding platform is located downstairs. This train's destination

Problem:

The model is a strong preference to speak English

Question: 这班车的终点站在哪里? 用中文作答。 (What is the destination?)
Answer in Chinese) [Unseen task and language]
Answer: This train's final destination station is in Union Station. 这班车的终点站在 Union Station.

140

Responsible audio AGI - fairness and trustworthiness

Fairness

- Existing models are mostly English-centric

→ Model may have original biases by nature

- Mainly due to training data, but hard to be solved by collecting multi-lingual data.
- Need a deep integration with translation systems

→ Not good at tasks requiring high reliability

Potential Funding Sources: NSF-SaTC / DoD / DoJ

141

Revisit this sample

Question: 你听到了什么? (What did you hear?) [Instruction in unseen language]
Answer: A train is moving and a man is speaking.

Question: Can you translate the spoken text into Chinese? [Unseen task]
Answer: The spoken text is 'Transfer point for the Metropolitan Purple Line to Busselton West. Boarding platform is located downstairs. This train's destination is Union Station. Please proceed to the boarding platform and present your ticket at the fare gate when boarding.' The translation would be '替换站点为大都市绿色线到布塞尔东。车站位于下层。这次列车的目的地是联邦站。请在购票门前准备出票并进入车站。'

Problem:

Purple Line is mistranslated as "绿色线" (Green Line)

Responsible audio AGI - fairness and trustworthiness

- Hard to be completely solved by pure neural-based methods
 - Require neural-symbolic methods

Trustworthiness

- Hallucination
- Not good at tasks requiring high reliability

Potential Funding Sources: NSF-SaTC / DoD / DoJ

143

Responsible audio AGI

Responsible Speech Foundation Models



Speech foundation models are emerging as a universal solution to various speech tasks. Indeed, their superior performance has extended beyond ASR. For instance, Whisper has proven to be a wider robust audio event tagger [1], showcasing its potential beyond its original training objectives. Despite the advancements, the limitations and risks associated with speech foundation models have not been thoroughly studied. For example, it has been found that wav2vec 2.0 exhibits biases in different paralinguistic features, emotions [2], and accents [3], while HuBERT lacks noise robustness in certain downstream tasks [4]. Besides this, foundation models present challenges in terms of ethical concerns, including privacy, sustainability, fairness, and safety [5]. Furthermore, risks and biases of one model may propagate in usage alongside other models, especially in a unified framework, such as Seamless [6].

Thus, it is necessary to investigate speech foundation models for de-biasing (e.g., consistent accuracy for different languages, different genders, and ages), enhancing factuality (not making mistakes in critical applications), preventing malicious applications (e.g., using a TTS to attack speaker verification systems, not to use for surveillance), and addressing various other aspects.

In this special session, we specialize in responsible aspects of speech foundation models, which are not adequately covered by regular sessions. We aim to facilitate knowledge sharing from diverse speech areas and pioneer discussions on both tech and non-tech issues. Furthermore, in line with the IS 2024 Speech and Beyond theme, we aim to foster connections with other communities such as NLP and ML, which have long been investigating responsible and trustworthy models [7]. Theoretical and position papers from those communities with views, directions, ideas, or solutions on a particular topic for bridging the gap between speech and NLP/ML are also welcome (e.g., integrating speech foundation models with LLMs in dialog systems).

Interspeech 2024 Special Session

Potential Funding Sources: NSF-SaTC / DoD / DoJ

144

My research vision

Develop **Audio and Speech AGI** that
not only has **stronger perception, understanding, and generation capabilities**
but also to be **safer and more socially beneficial**

Audio AGI and other ECE areas



145

146

Acknowledgements



Alexander H. Liu
MIT



Hongyin Luo
MIT



Leonid Karlinsky
MIT-IBM Lab



Yu-An Chung
Meta



Jeff Cheng-I Lai
MIT



James Glass
MIT



147