



Massachusetts Institute of Technology

From Audio Perception to Understanding: A Path Towards Audio AGI

Yuan Gong
Research Scientist
MIT Spoken Language System Group

4/1/2024

Outline

Part 1

Listen

General audio
events
recognition

Part 2

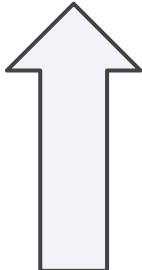
Understand

Audio large language model

Part 3

Future Research

- Advancing audio AGI
- Audio AGI for social good
- Responsible audio AGI



Motivation

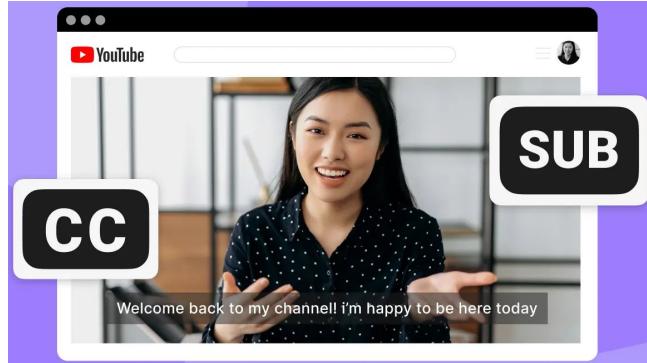
Improve Human-Computer Interaction



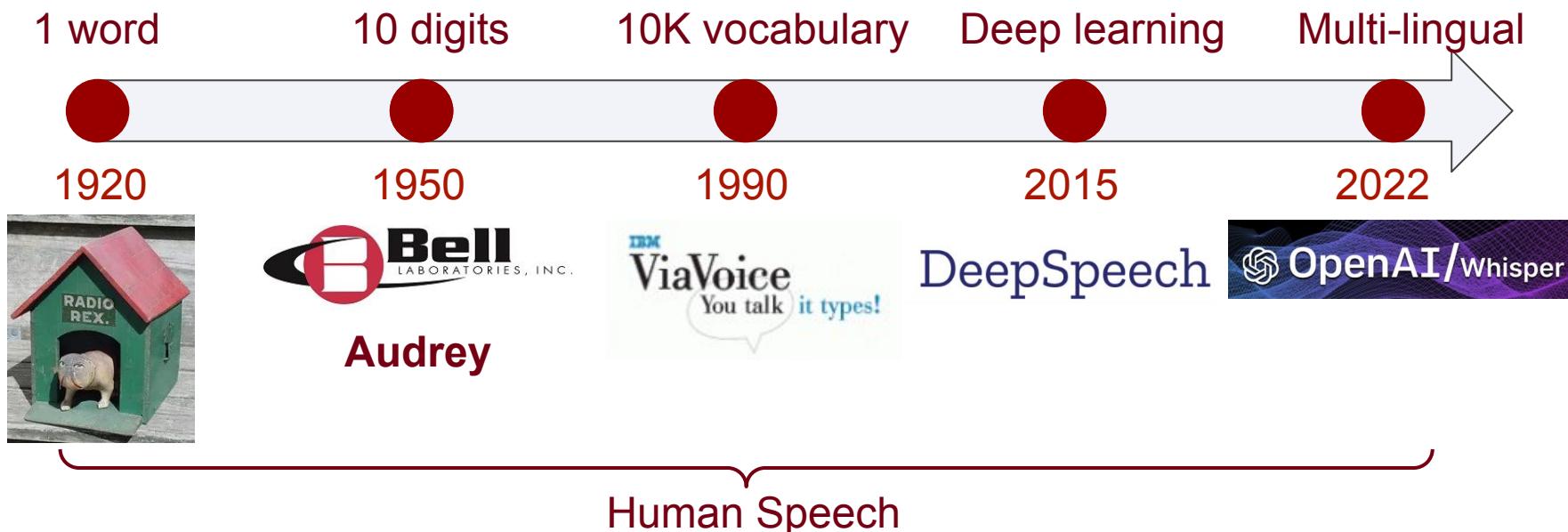
Enhance Human Communication



Boost Accessibility



History of machine listening



History of machine listening

But speech is just a small fraction of general sounds



Audrey



DeepSpeech



Human Speech

Sample



What did you hear?

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

Crucial Information but
not in speech

Applications of general sound recognition

Further Boost Accessibility



Embodied AI with Audio Capability



Music Recognition

Nocturne Op. 9, No. 2

Arranged by Abby Knudsen
Music by Frédéric Chopin

Sound-Based Surveillance



History of machine listening

First workshop
on general audio
recognition



First large
general audio dataset



What is the difference between
speech and general audio signals?



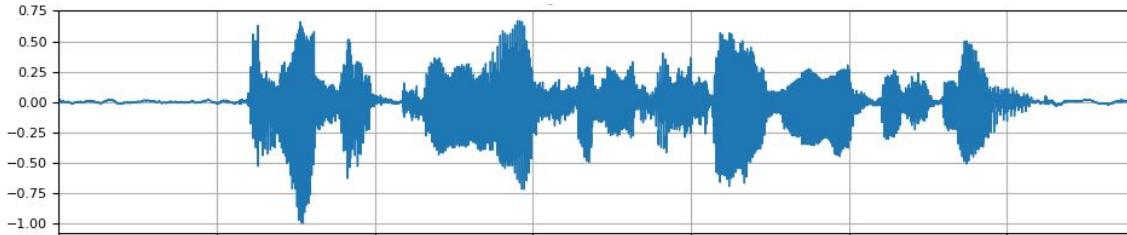
Audrey



DeepSpeech

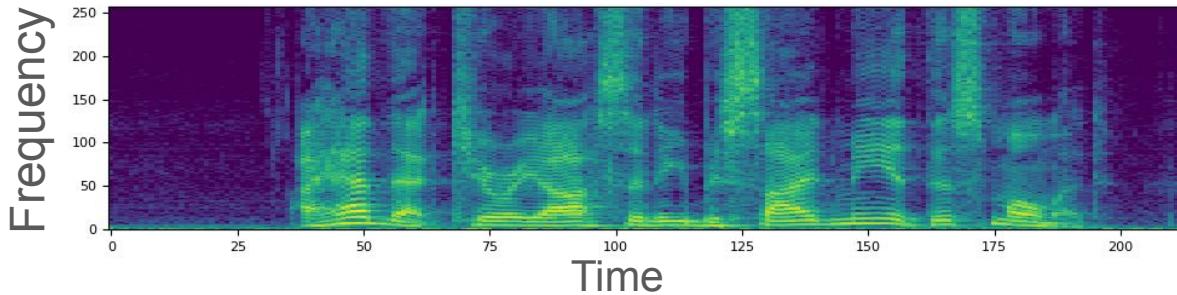


Preliminary: audio spectrogram



Waveform

Short-Time Fourier Transform



Spectrogram

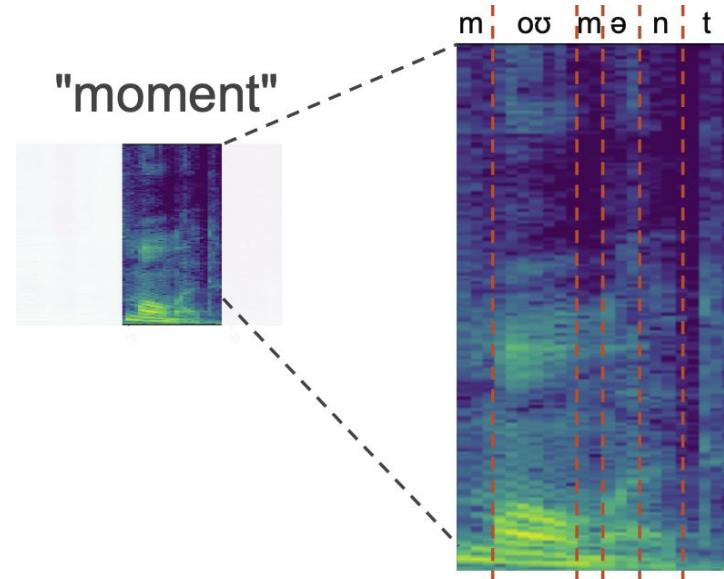
Audio spectrogram is a form of audio representation that is easier for both humans and machines to analyze audio

Speech is composed of a finite number of elementary units

Speech

- composed of elementary units: phonemes
 - **~40** phonemes for English

- **Vowels:** [i], [ə], [u]
- **Consonants:**
 - **Plosives:** [p], [t], [k]
 - **Fricatives:** [f], [s], [ʃ]
 - **Nasals:** [m], [n], [ŋ]
- **Semivowels:** [j], [w]



There are unlimited number of general sounds

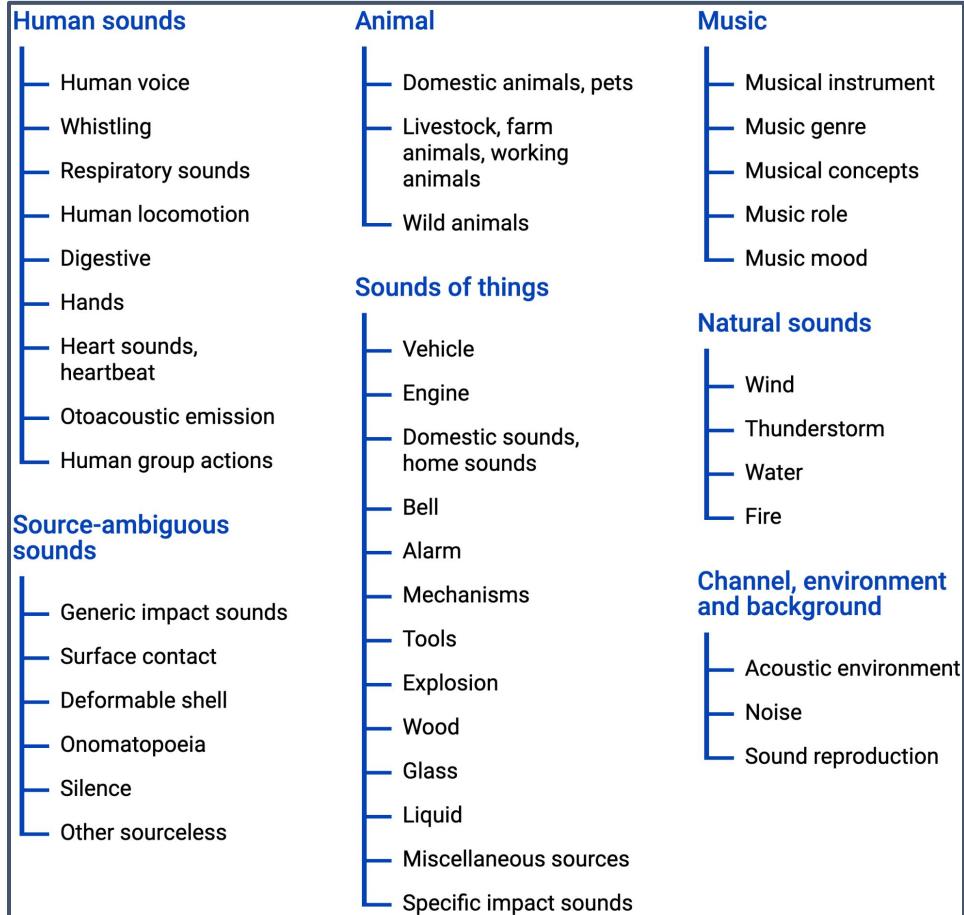
Speech

- composed of elementary units: phonemes
 - **~40** phonemes for English

General audio

- **no** elementary units
- **unlimited number** of sound classes
 - **527** classes in AudioSet

There are unlimited number of general sounds



**AudioSet ontology
(top-level)
527 classes in total**

There are unlimited number of general sounds

Human sounds

- Human voice
- Whistling
- Respiratory sounds
- Human locomotion
- Digestive
- Hands
- Heart sounds, heartbeat
- Otoacoustic emission
- Human group actions

Source-ambiguous sounds

- Generic impact sounds
- Surface contact
- Deformable shell
- Onomatopoeia
- Silence
- Other sourceless

Animal

- Domestic animals, pets
- Livestock, farm animals, working animals
- Wild animals**

Sounds of things

- Vehicle
- Engine
- Domestic sounds, home sounds
- Bell
- Alarm
- Mechanisms

Tools

- Explosion
- Wood
- Glass
- Liquid
- Miscellaneous sources
- Specific impact sounds

Music

- Musical instrument
- Music genre
- Musical concepts
- Music role
- Music mood

Natural sounds

- Wind
- Thunderstorm
- Water
- Fire

Channel, environment and background

- Acoustic environment
- Noise
- Sound reproduction



Wild animals: Roaring cats, Bird, Canidae, Rodents, Insect, Frog, Snake, Whale (8)



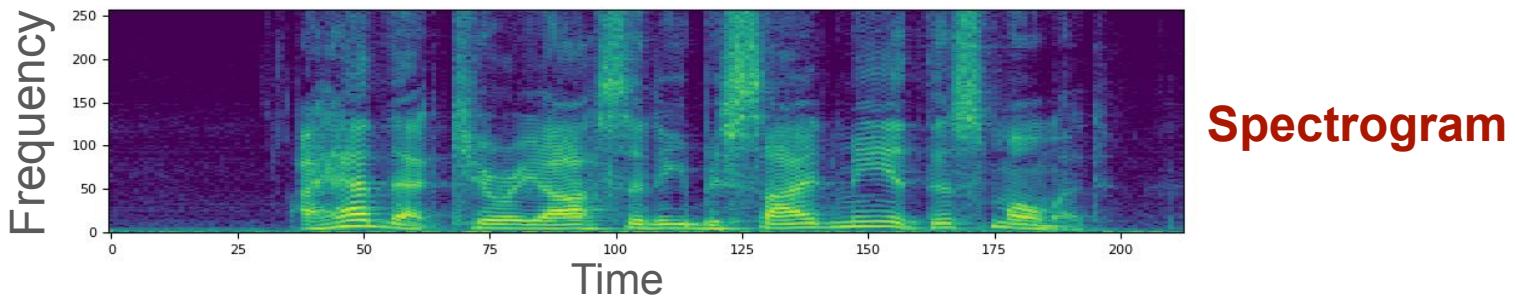
Tools: Hammer, Jackhammer, Sawing, Filing, Sanding, Power tool (6)

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

General audio classification has a larger search space

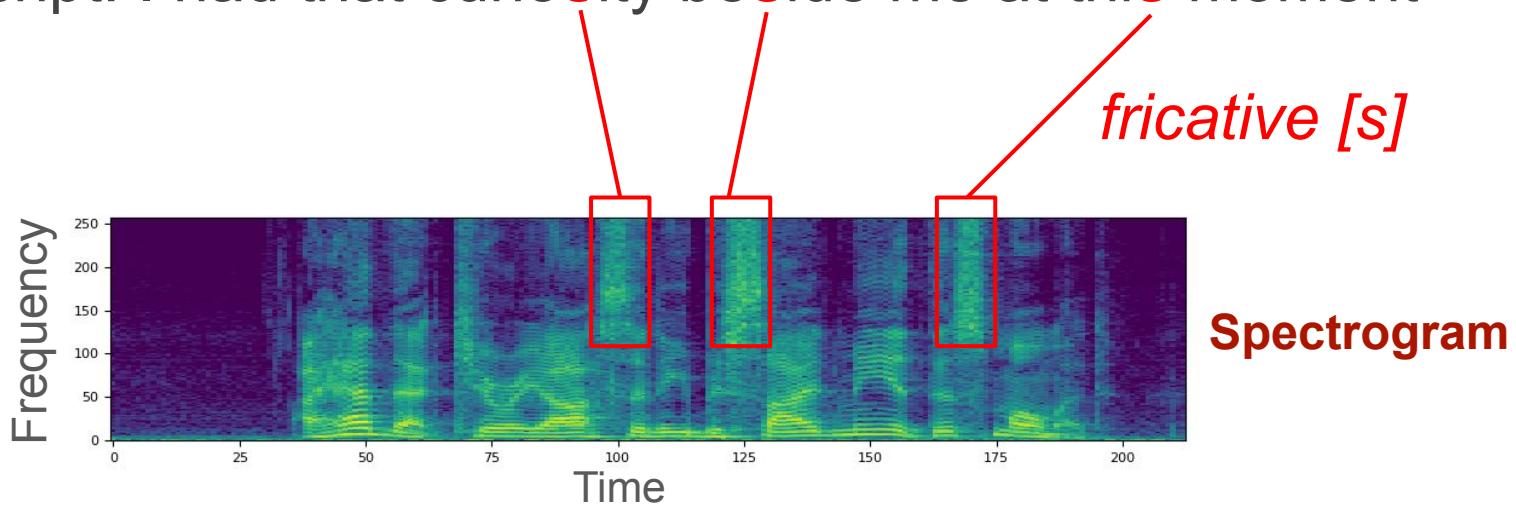
Speech has structured spectrogram patterns

Transcript: I had that curiosity beside me at this moment



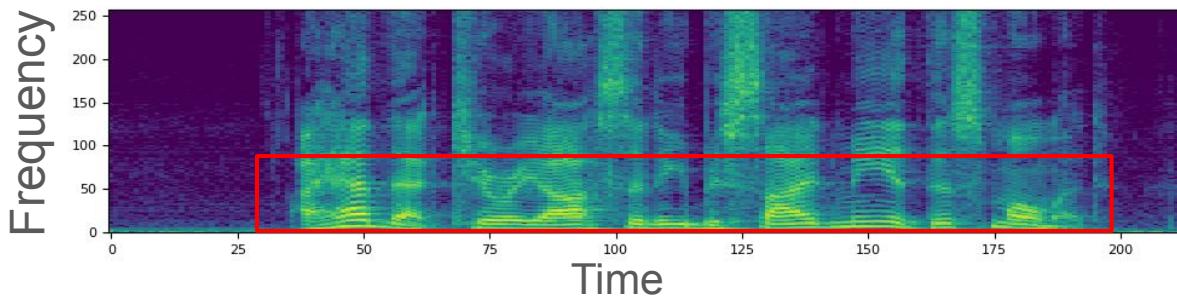
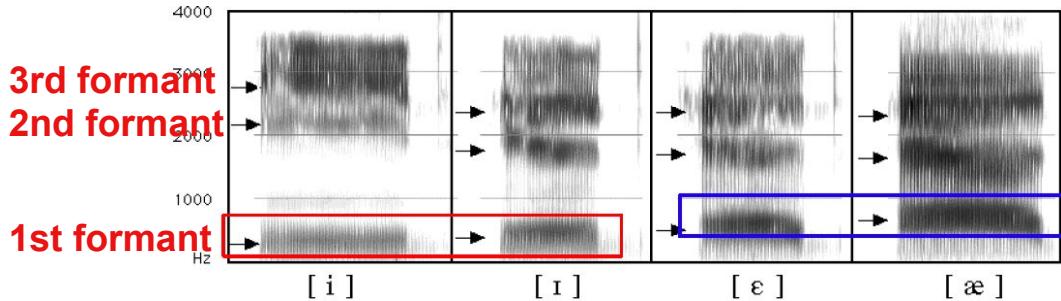
Speech has structured spectrogram patterns

Transcript: I had that curios**s**ity beside me at this moment



Fricatives exhibit high-frequency turbulence

Speech has structured spectrogram patterns

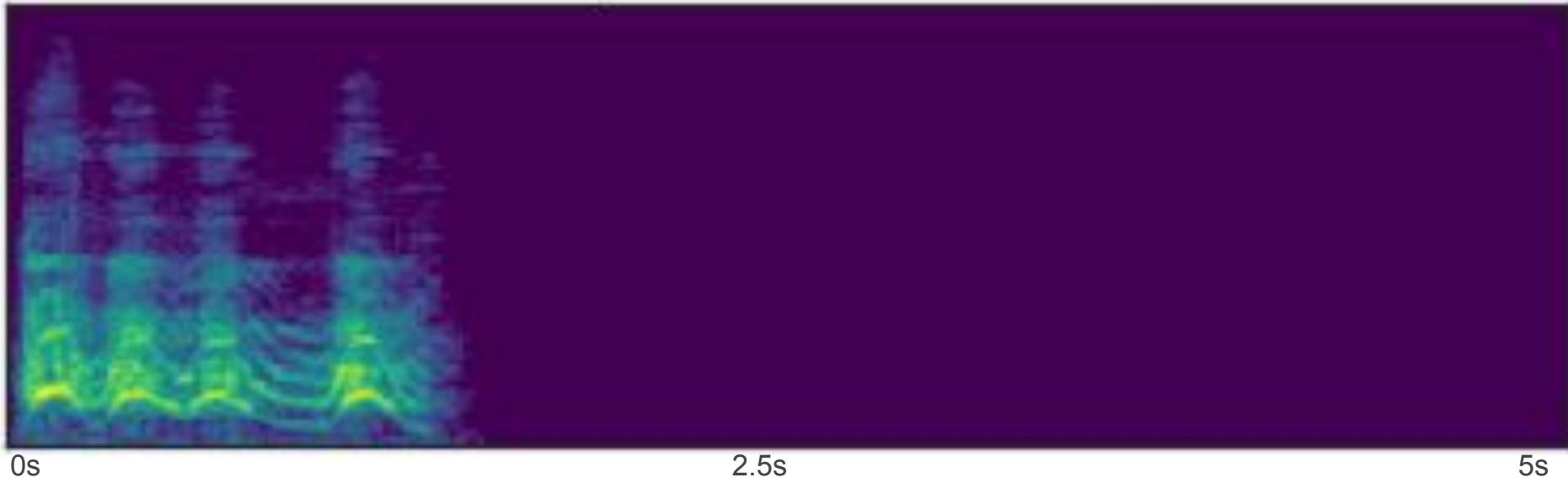


Spectrogram

- 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



- 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

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

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



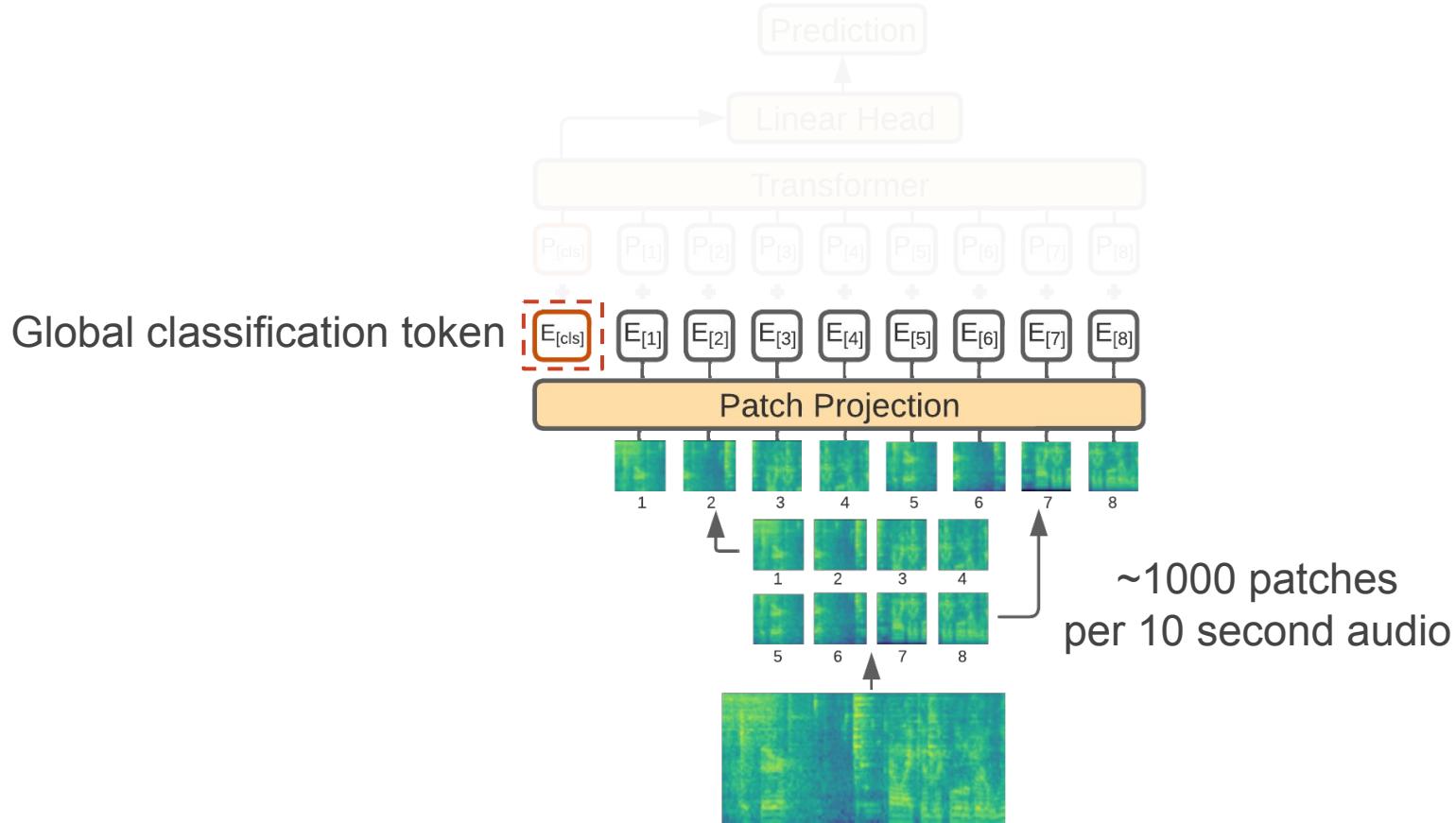
Audio spectrogram transformer (AST)



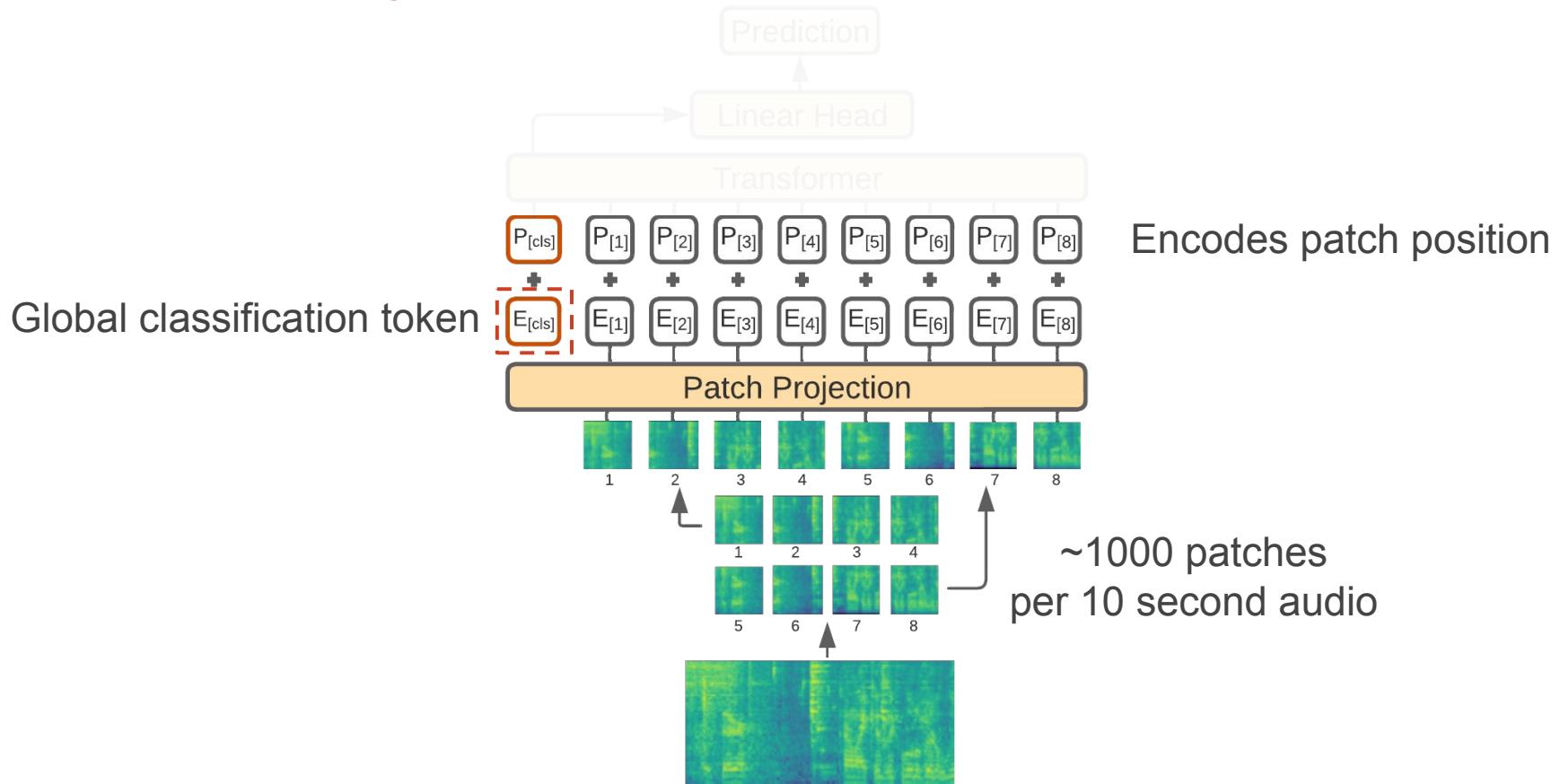
Audio spectrogram transformer (AST)



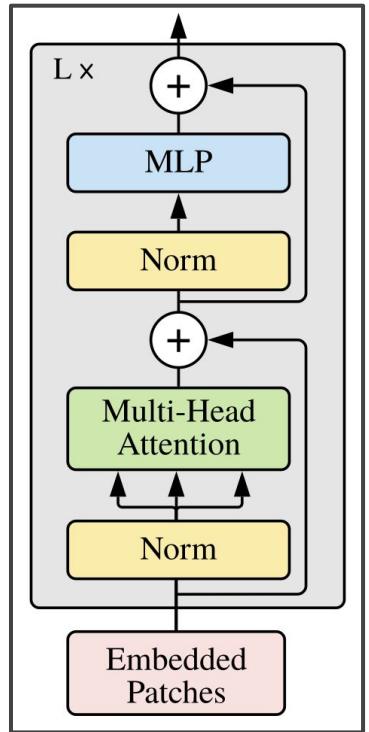
Audio spectrogram transformer (AST)



Audio spectrogram transformer (AST)

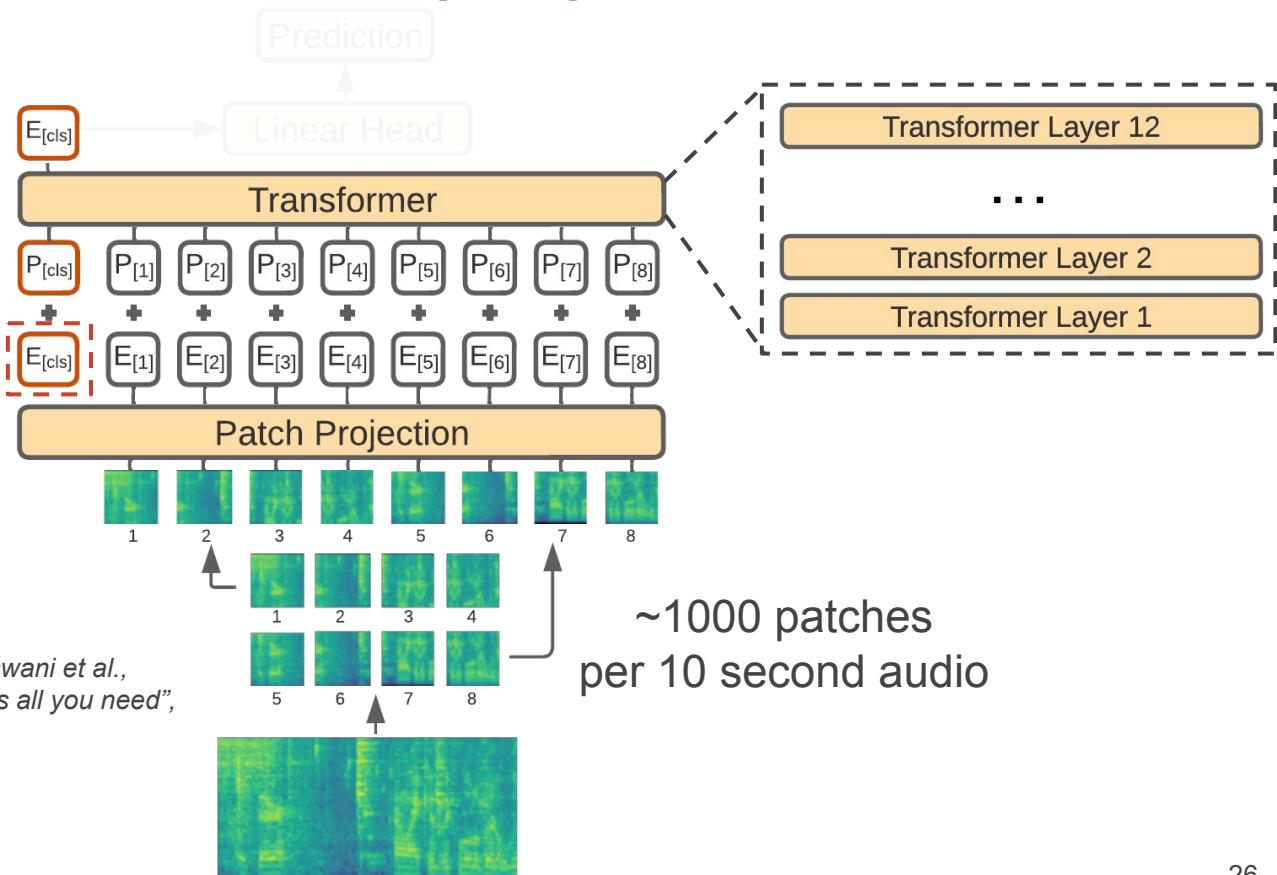


Audio spectrogram transformer (AST)



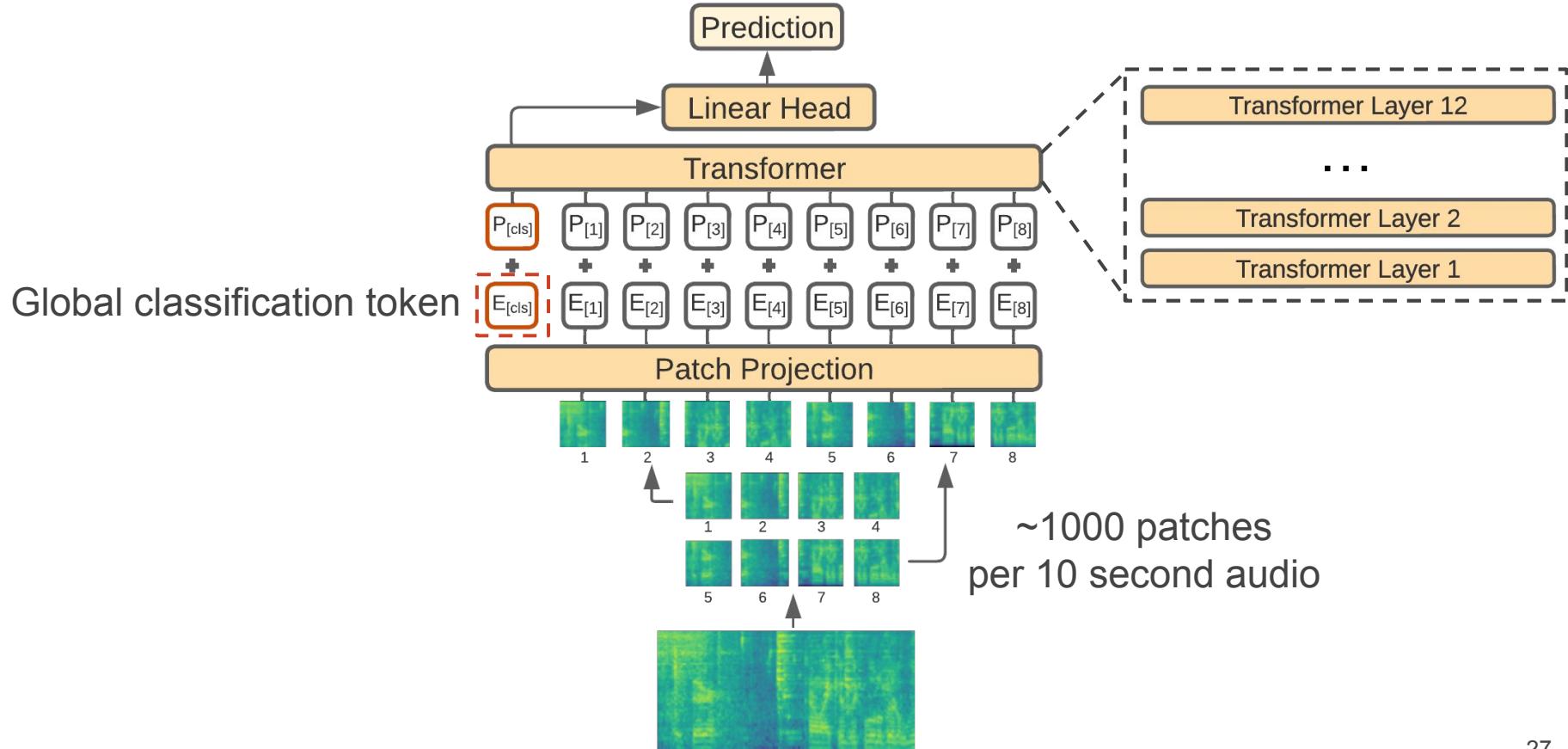
Transformer layer

Ashish Vaswani et al.,
“Attention is all you need”,
2017



~1000 patches
per 10 second audio

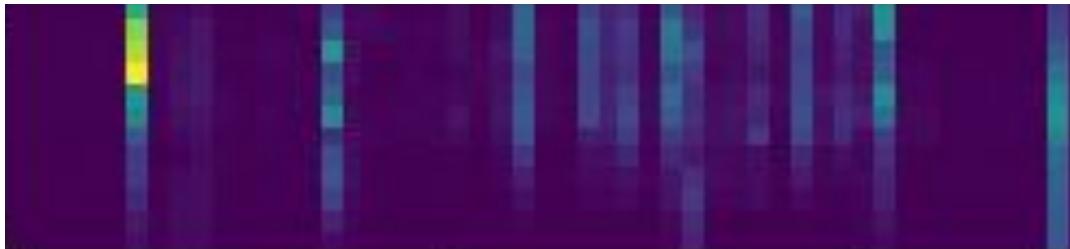
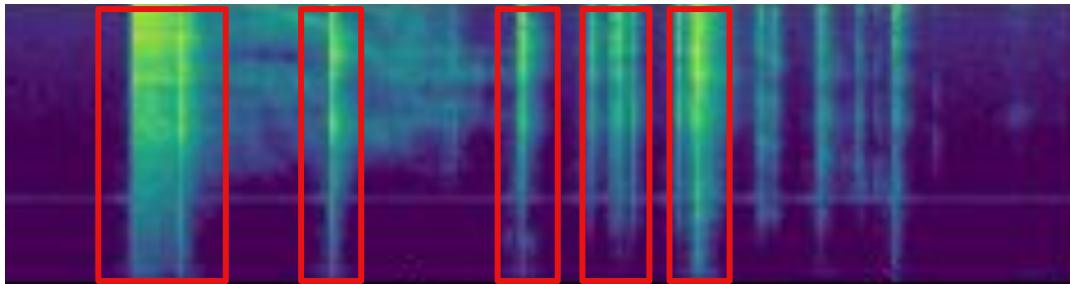
Audio spectrogram transformer (AST)



	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



First Layer Attention Map

	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

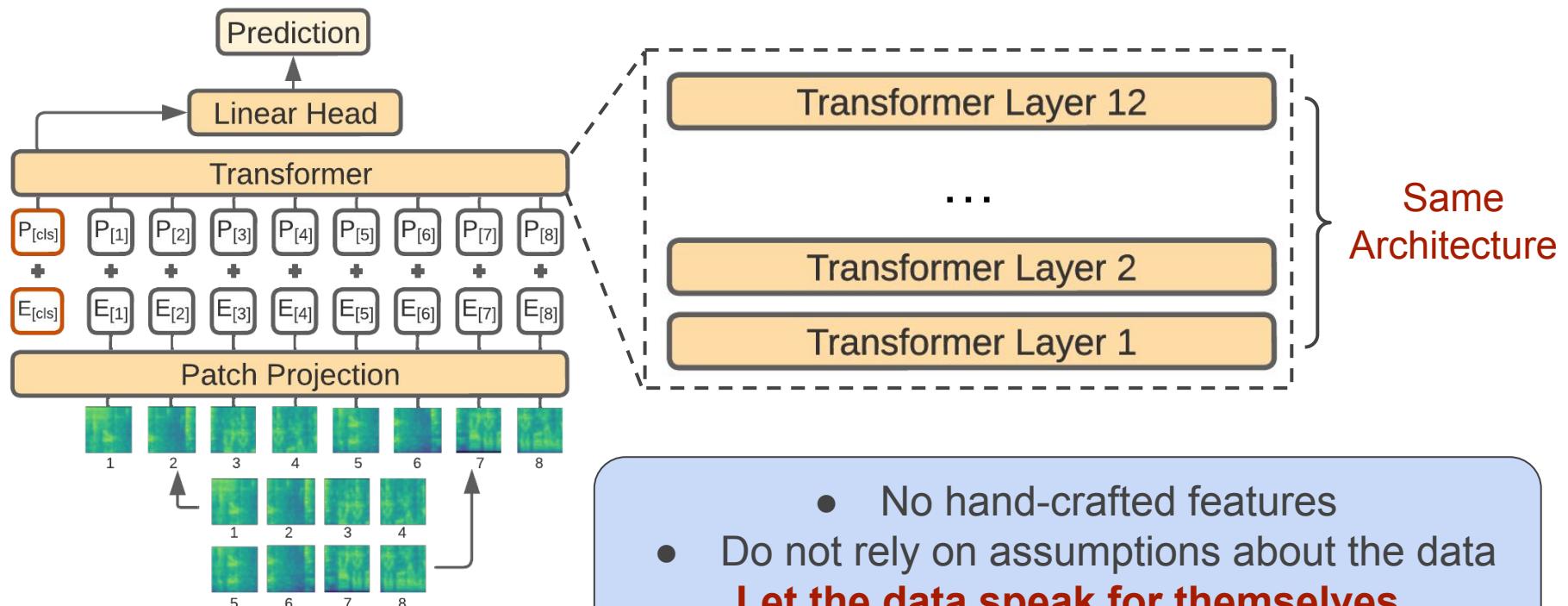
Attend to multiple sounds



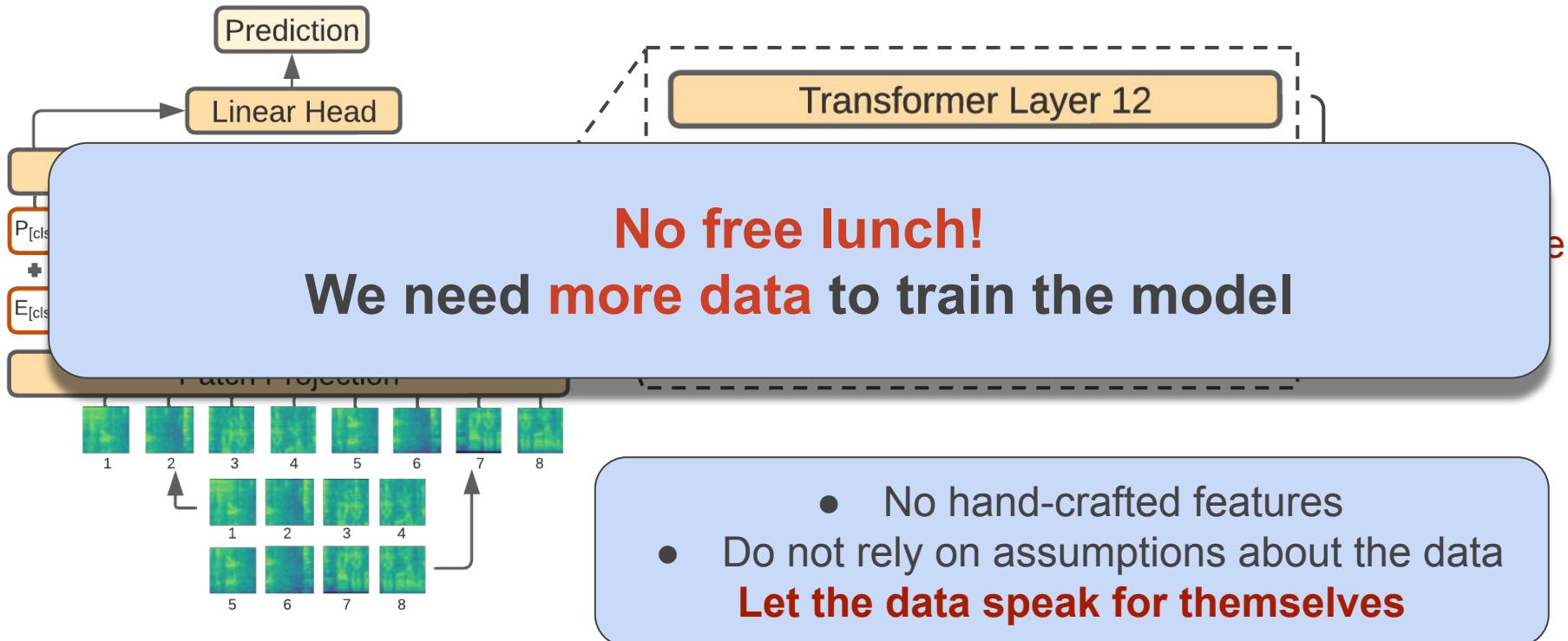
Advantage of AST: dynamic attention



AST is a data-driven model

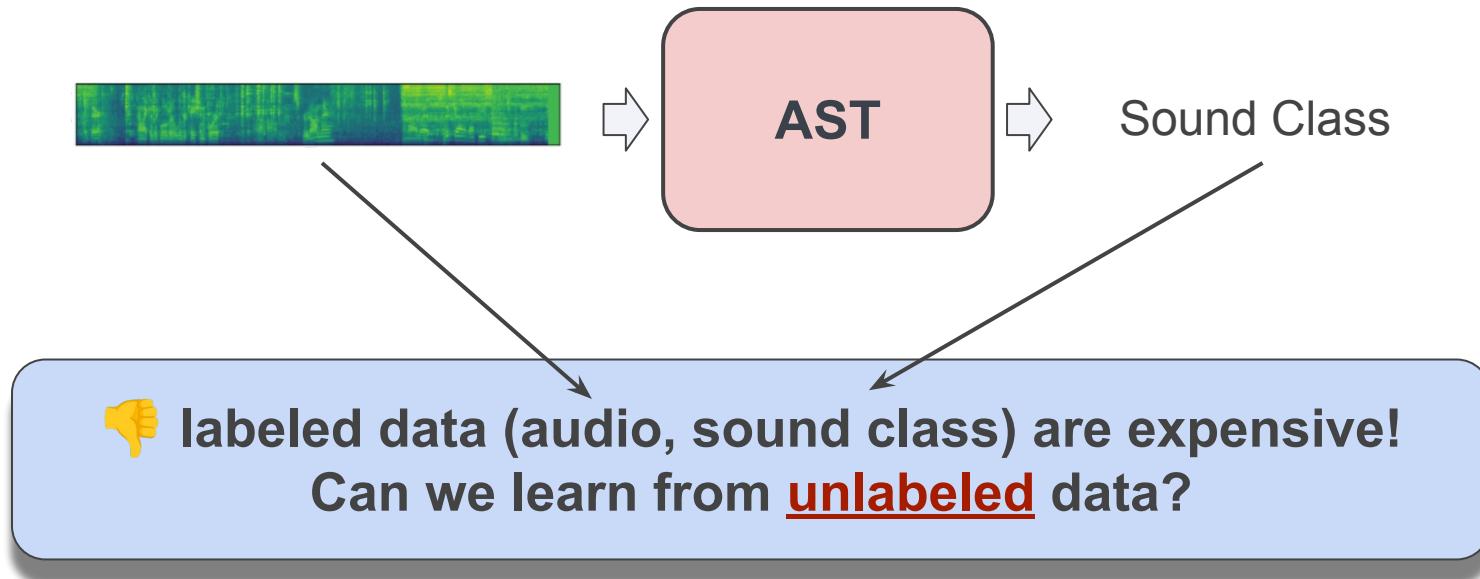


AST is a data-driven model



The data issue

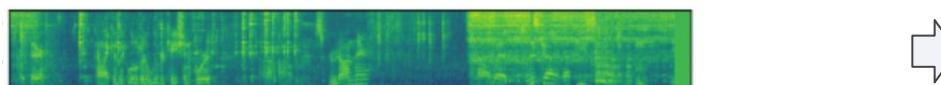
Conventional supervised learning



Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data



Random masking, no human annotation required

Audio self-supervised learning

Sta

Why this work?

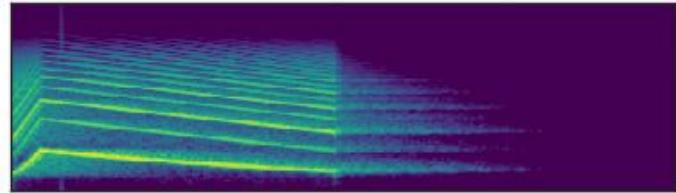
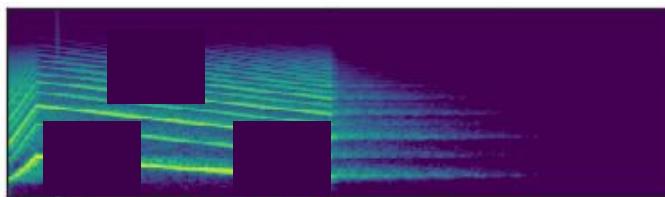
La

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

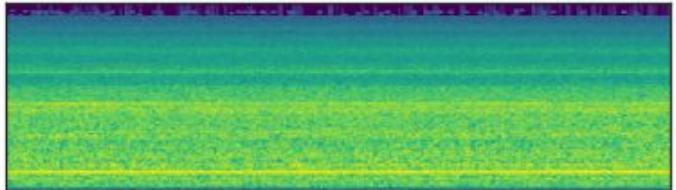
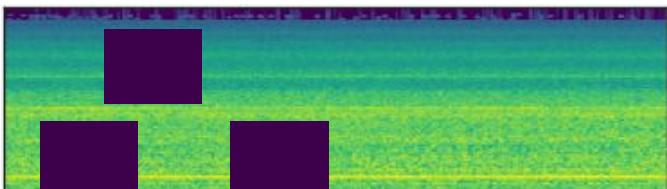


This sound features rise and fall in pitch

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

Audio self-supervised learning

Stage 1: Self-supervised pretraining



At this point, the model cannot do classification

Audio self-supervised learning

Stage 1: Self-supervised pretraining



Large-scale unlabeled data

AST

Weight Inheritance

Stage 2: Supervised finetuning

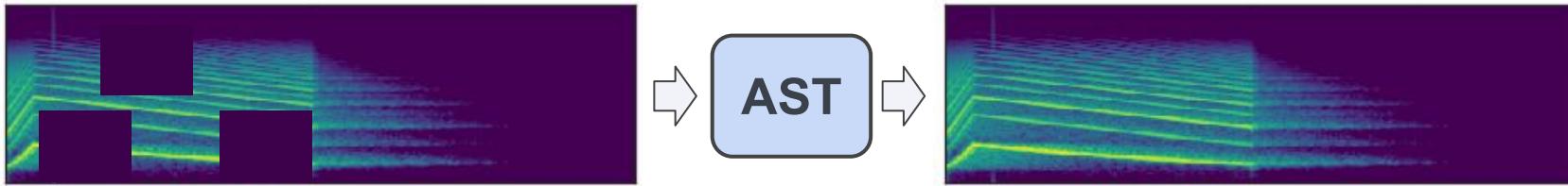


Small-scale labeled data

AST

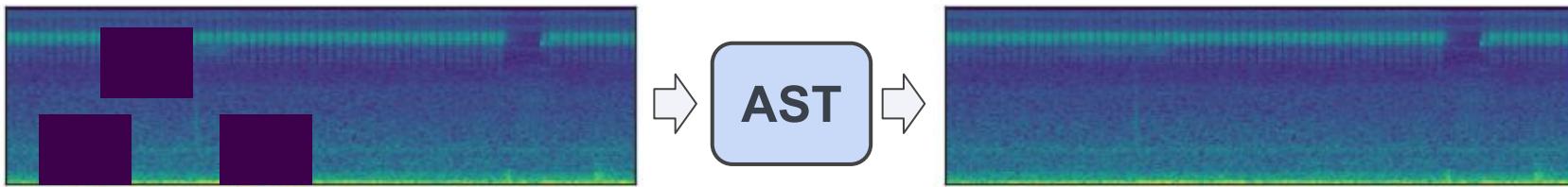
Sound Class

Audio self-supervised learning



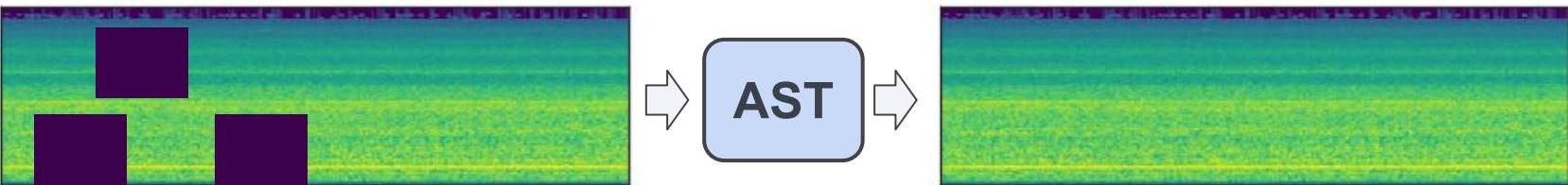
Siren

This sound features rise and fall in pitch



Cricket

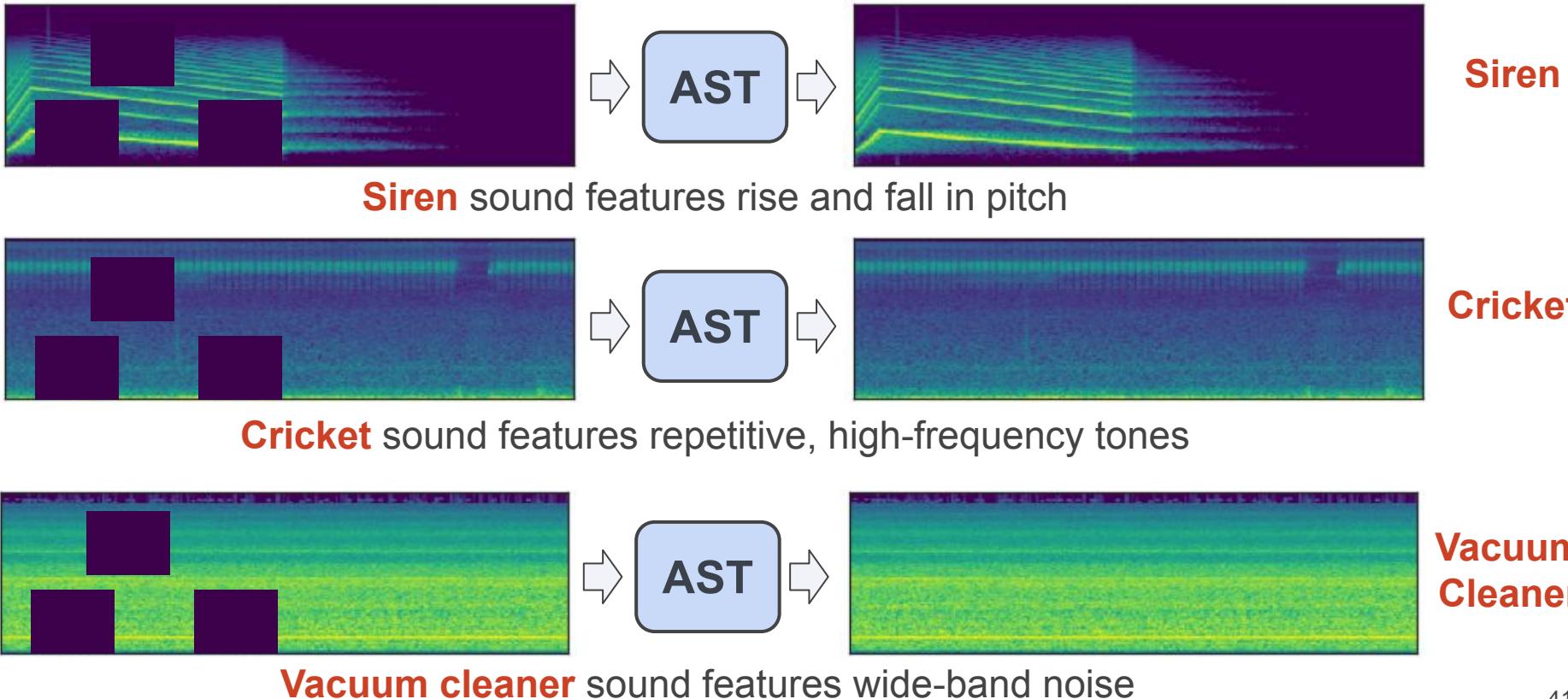
This sound features repetitive, high-frequency tones



Vacuum
Cleaner

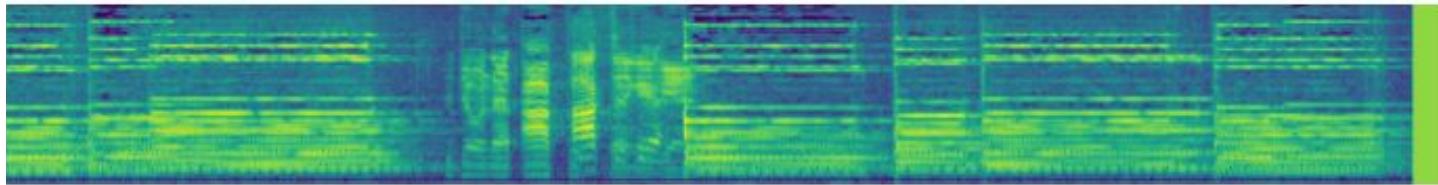
This sound features wide-band noise

Audio self-supervised learning



Surprisingly strong reconstruction ability

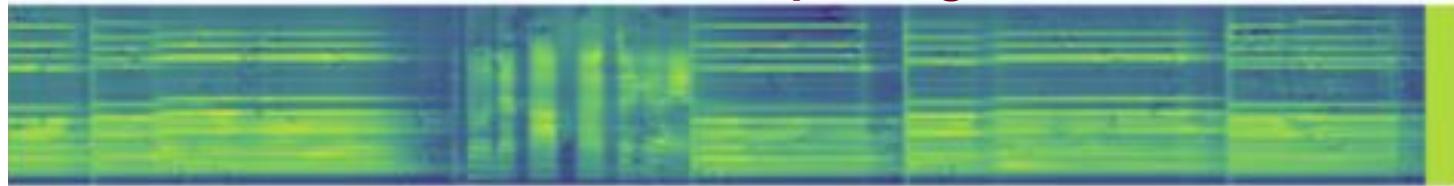
Original Spectrogram



Masked Spectrogram (75% Masked)



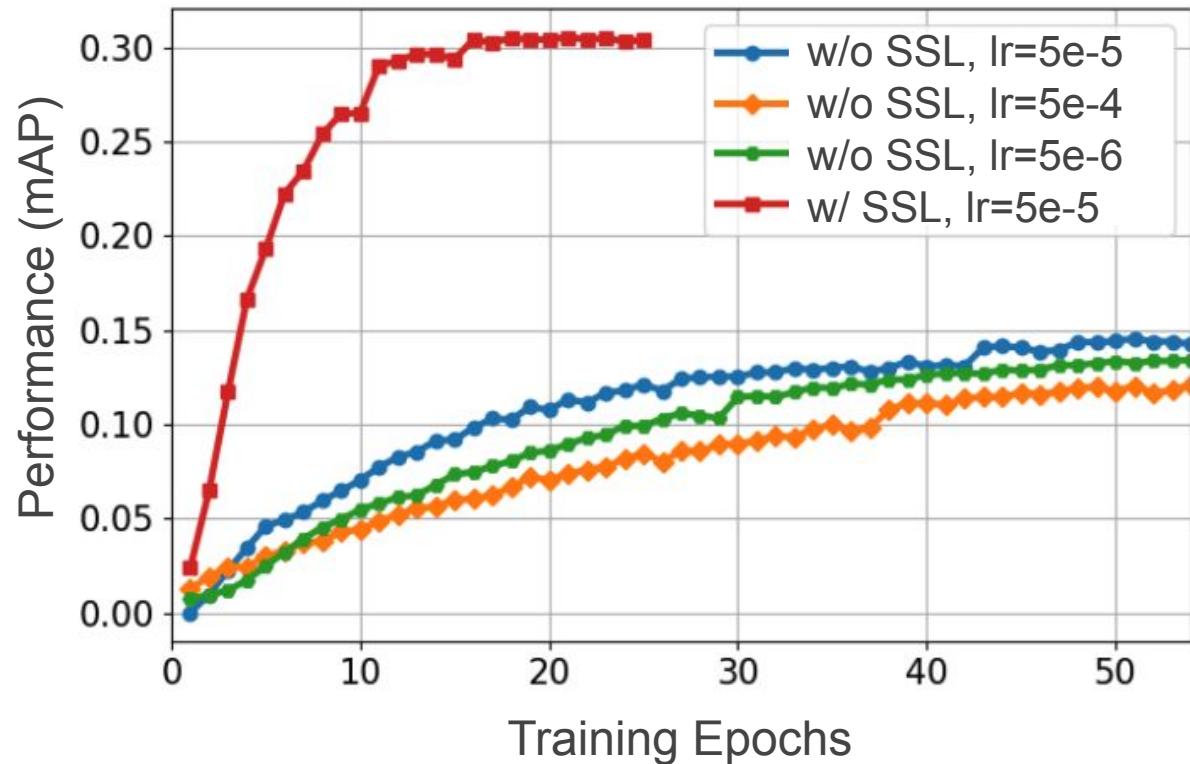
Reconstructed Spectrogram



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

Improvement led by self-supervised learning (SSL)

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



Unified speech and non-speech recognition



Audio Tags: Speech; Dishes, pots, and pans; Stir; Frying (fo

Speech: This is nice and hot, about medium high.

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

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

Audio Spectrogram Transformer (fine-tuned on AudioSet)

Downloads last month
2,226,298

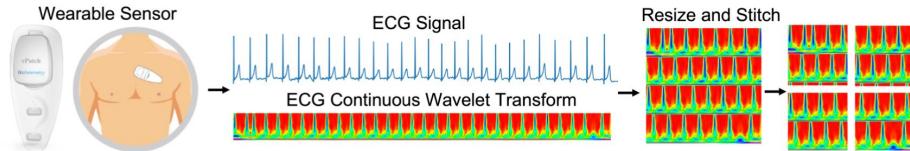
>2 million model downloads per month
when AST is 3 years old

Impact of AST

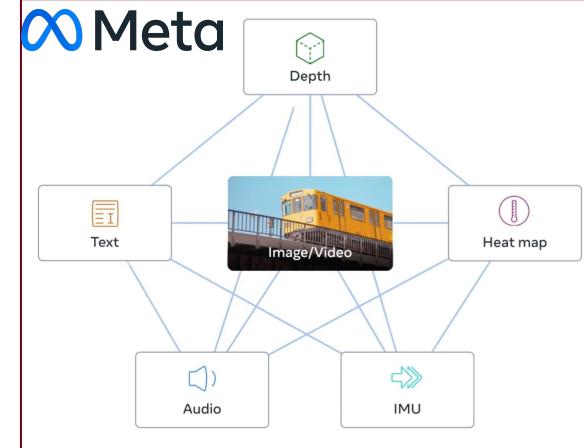
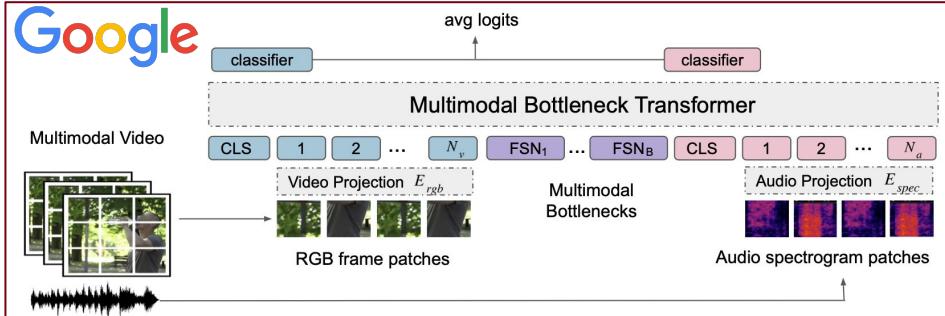
Wildlife Conservation



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

Outline

Part 1

Listen

General audio
recognition

Part 2

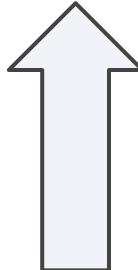
Understand

Audio large language model

Part 3

Future Research

- Advancing audio AGI
- Audio AGI for social good
- Responsible audio AGI



There is one thing missing...



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

We only modeled ***listening*** ability

Human cognitive ability



[Listen] High-frequency crackling noises, intermittent low-frequency pops.

[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?

perceive



understand

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.

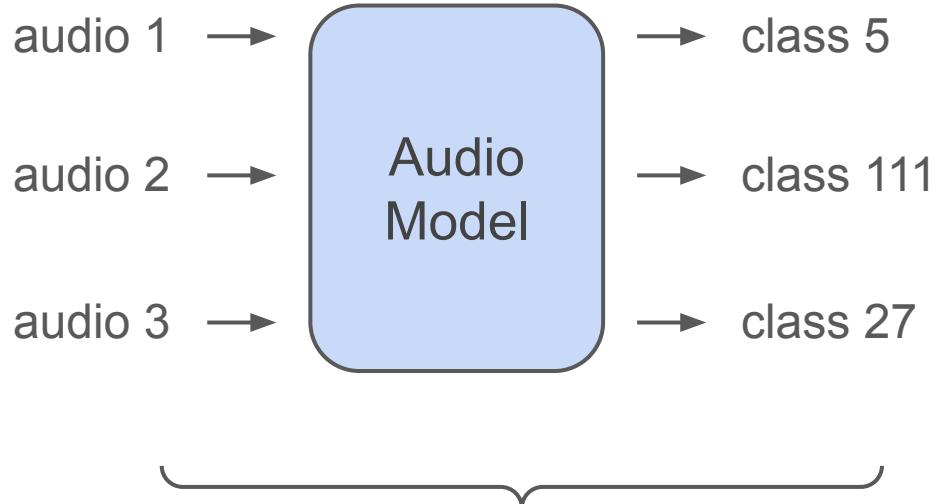
Can we make AI having a similar understanding ability?

[Understanding] A hands-on cooking teaching course.

[Emotion, Action] Hungry, what should I eat for lunch?

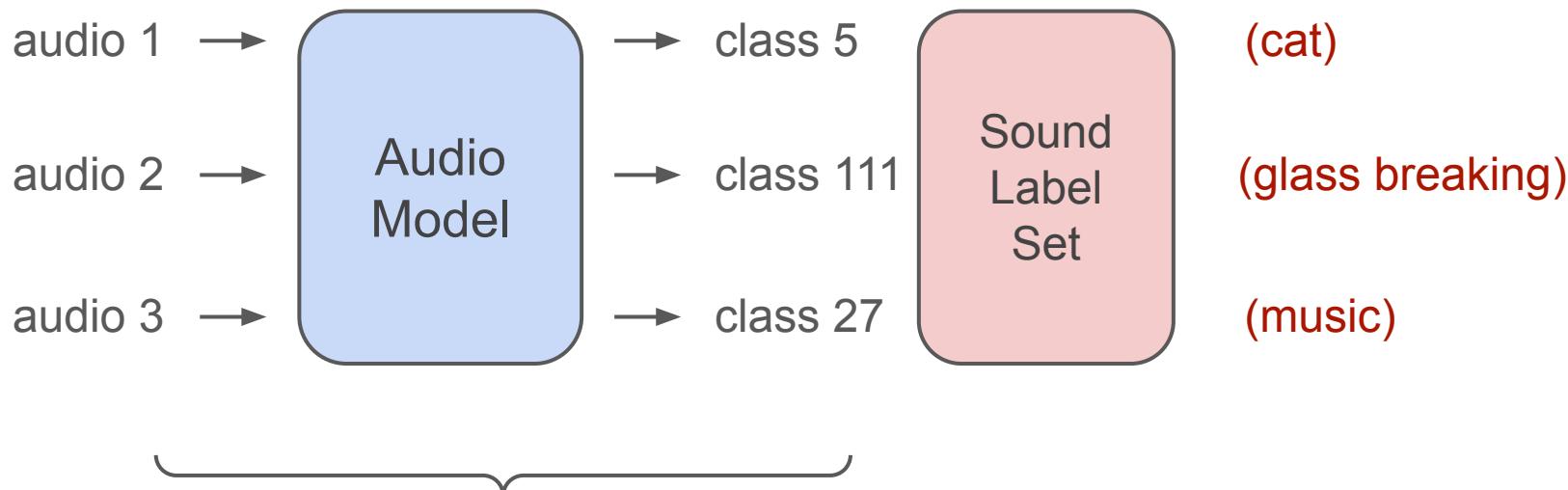
understand

Conventional training won't work



"Listening" model only knows class indices, not sounds

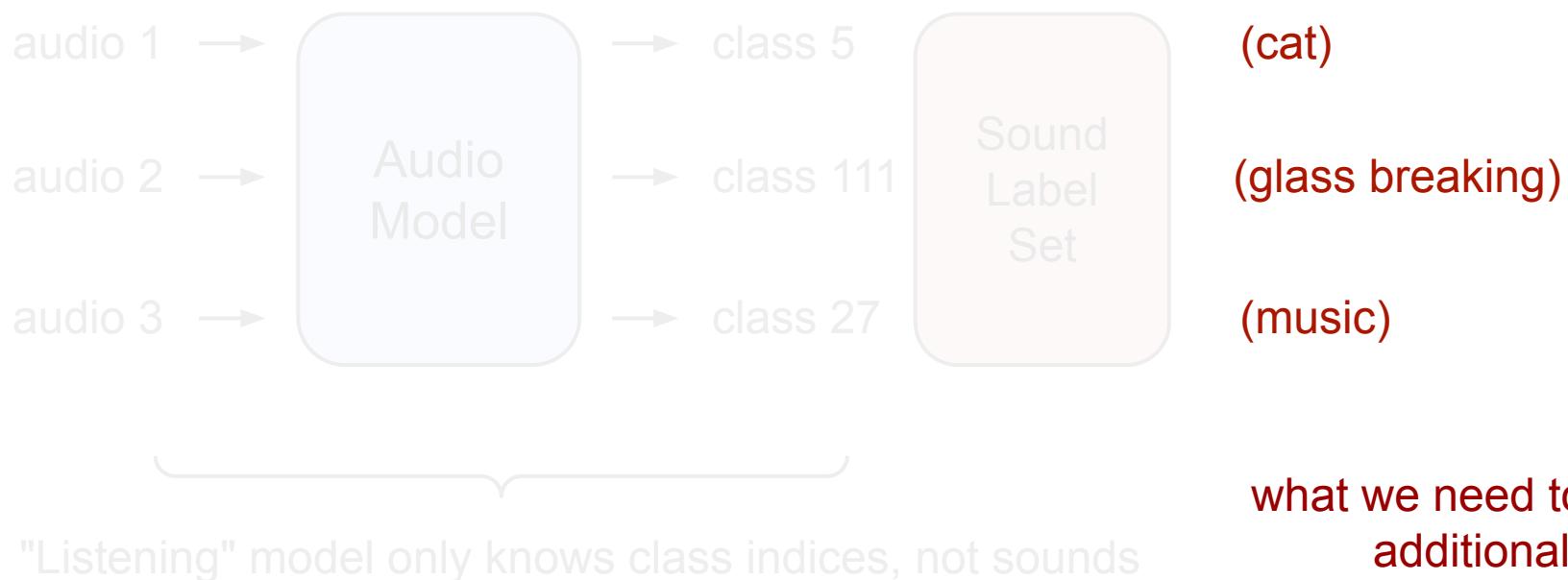
Conventional training won't work



"Listening" model only knows class indices, not sounds

Conventional training won't work

"cat's meow is soft" knowledge about sound



Conventional training won't work

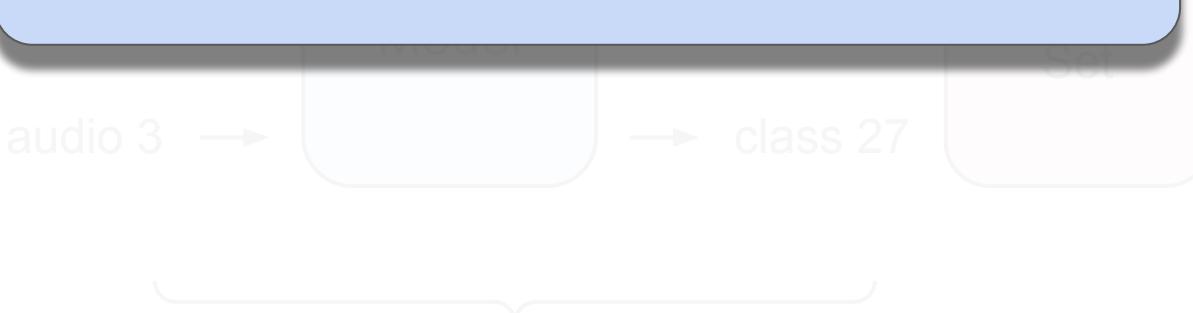


Conventional training won't work



Conventional training won't work

Text-based large language models can model these



knowledge about sound

common sense

(cat)

(glass breaking)

(music)

what we need to model additionally

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.

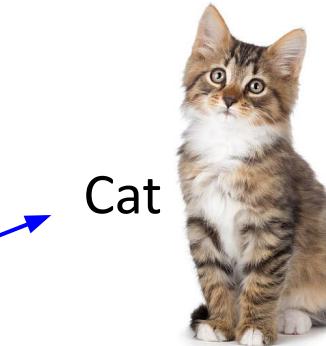
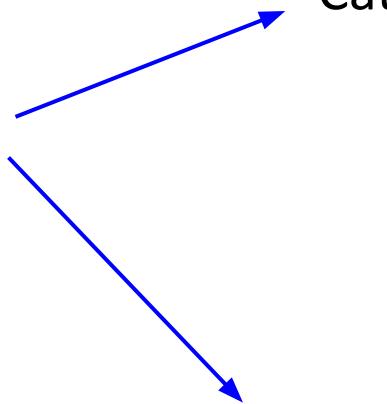
Pure language LLMs already have knowledge about sounds

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.

Pure language LLMs already have knowledge about sounds



Cheetah



Cat



Tiger



“Small” language model (LaBSE) gives wrong answer

한국어
Sentence Similarity

Source Sentence

Sound of cheetah

Sound of cat

Sound of tiger

Sound of lion

0.771

0.821

0.773

least similar



most similar



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."

Pure language LLMs already have knowledge about sounds

YU

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

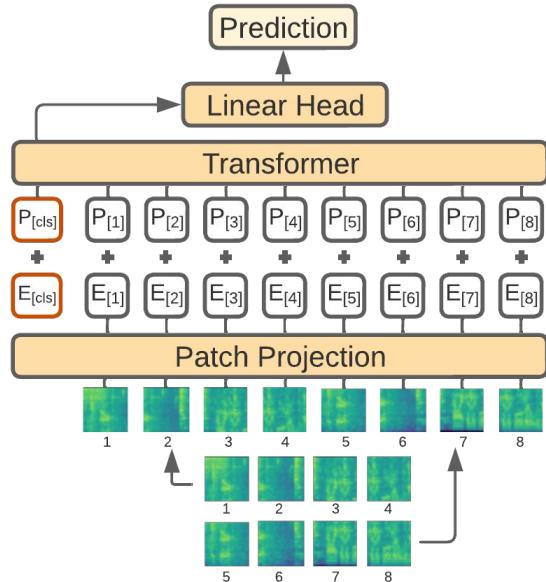


When comparing the sounds of cheetahs to those of domestic cats and tigers, the vocalizations of cheetahs are more akin to domestic cats. Here's a breakdown:

1. Purring:

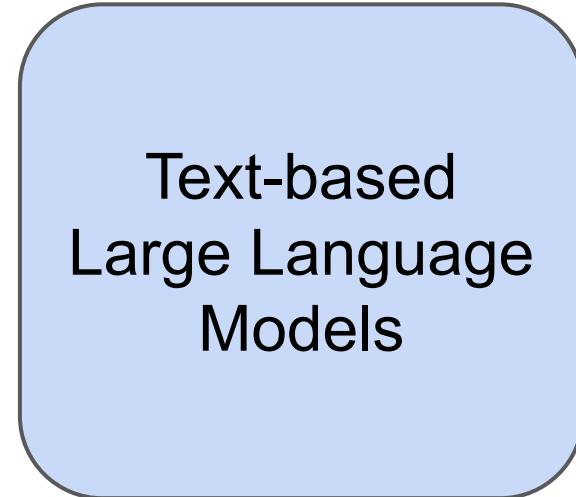
- **Cheetah:** Cheetahs are one of the few big cats that purr in a manner similar to domestic cats. Their purring is continuous, occurring during both inhalation and exhalation.
- **Domestic Cat:** Domestic cats purr in much the same way, during both inhalation and exhalation.
- **Tiger:** While tigers 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.

Bridging sound perception and understanding



Conventional Audio Model

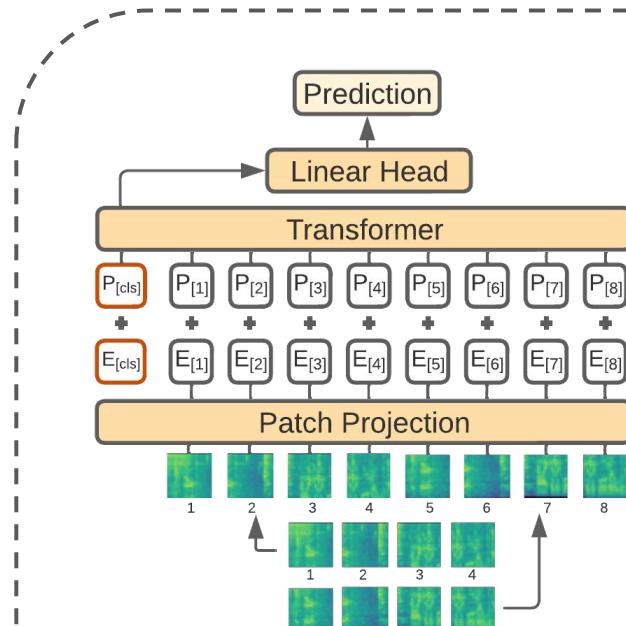
- 👍 Strong Perception Ability
- 👎 Do not Understand Sounds



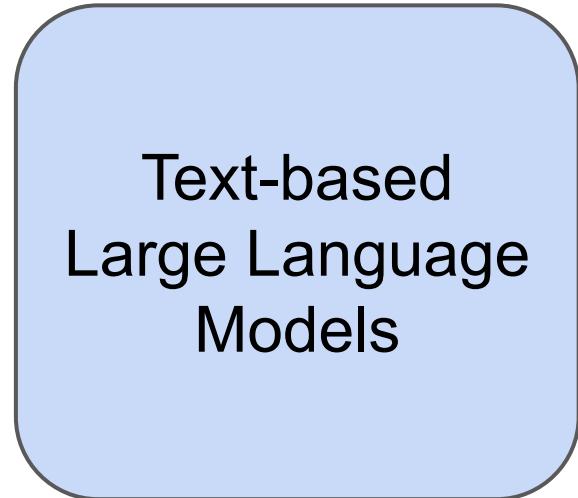
Large Language Models

- 👍 Good Knowledge about Sounds
- 👍 Strong Reasoning Ability
- 👎 No Sound Perception Ability

Bridging sound perception and understanding



Conventional Audio Model
👍 Strong Perception Ability
👎 Do not Understand Sounds

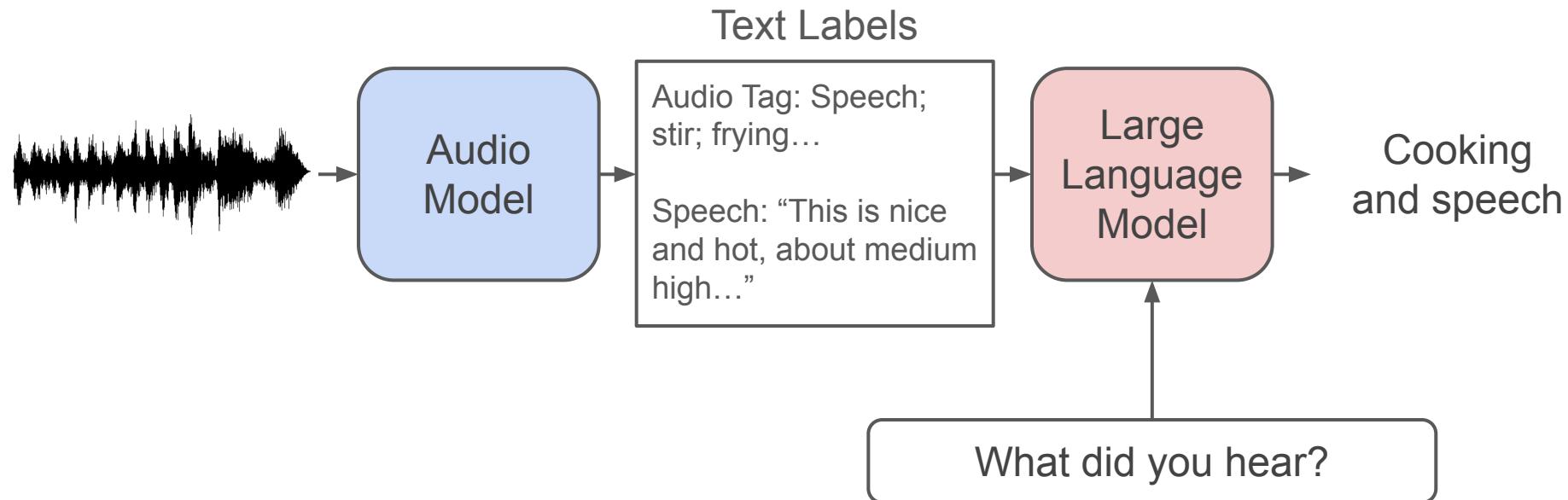


Large Language Models
👍 Good Knowledge about Sounds
👍 Strong Reasoning Ability
👎 No Sound Perception Ability

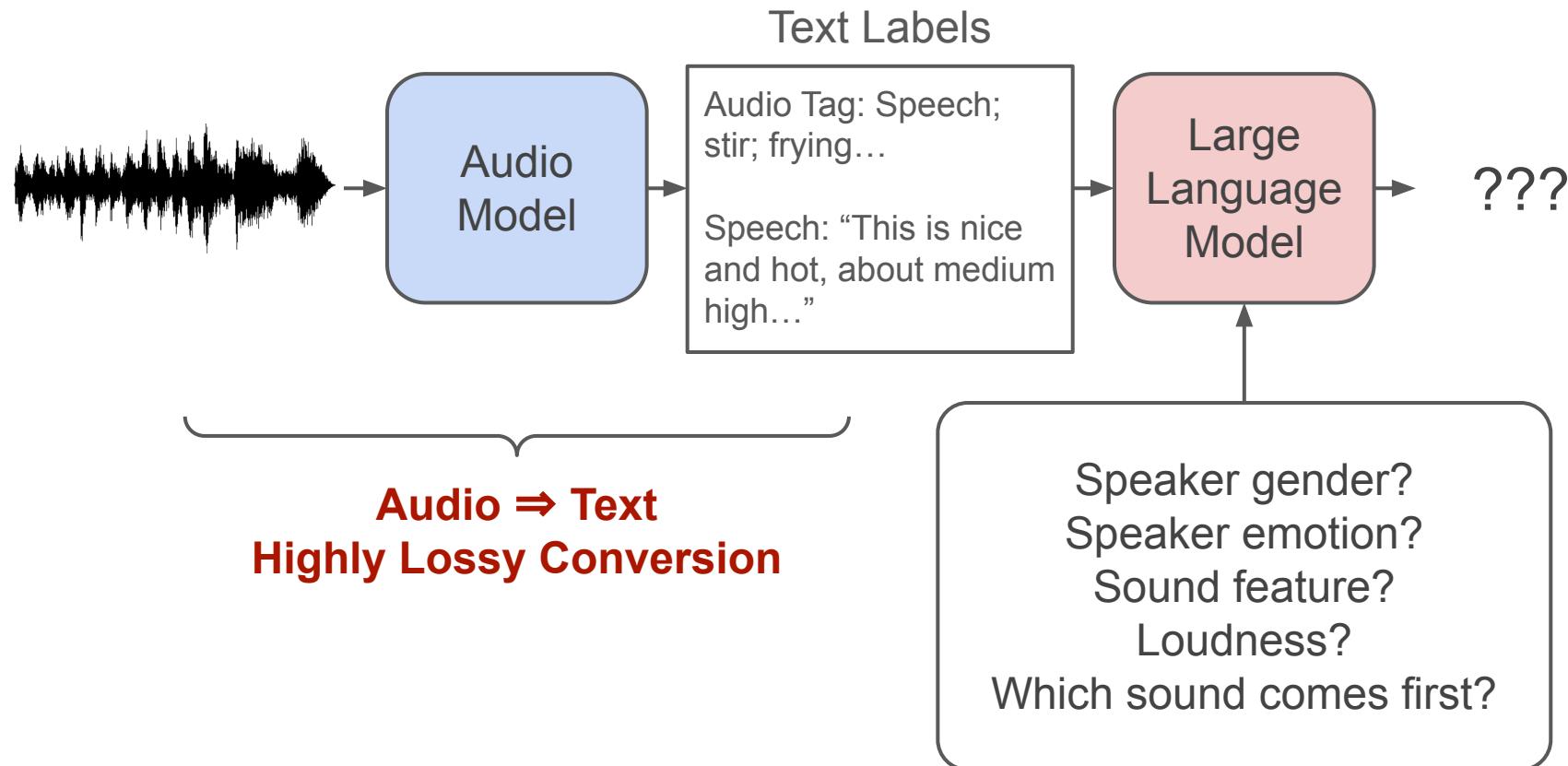
Bridging sound perception and understanding

Listen, Think, and Understand
(LTU)

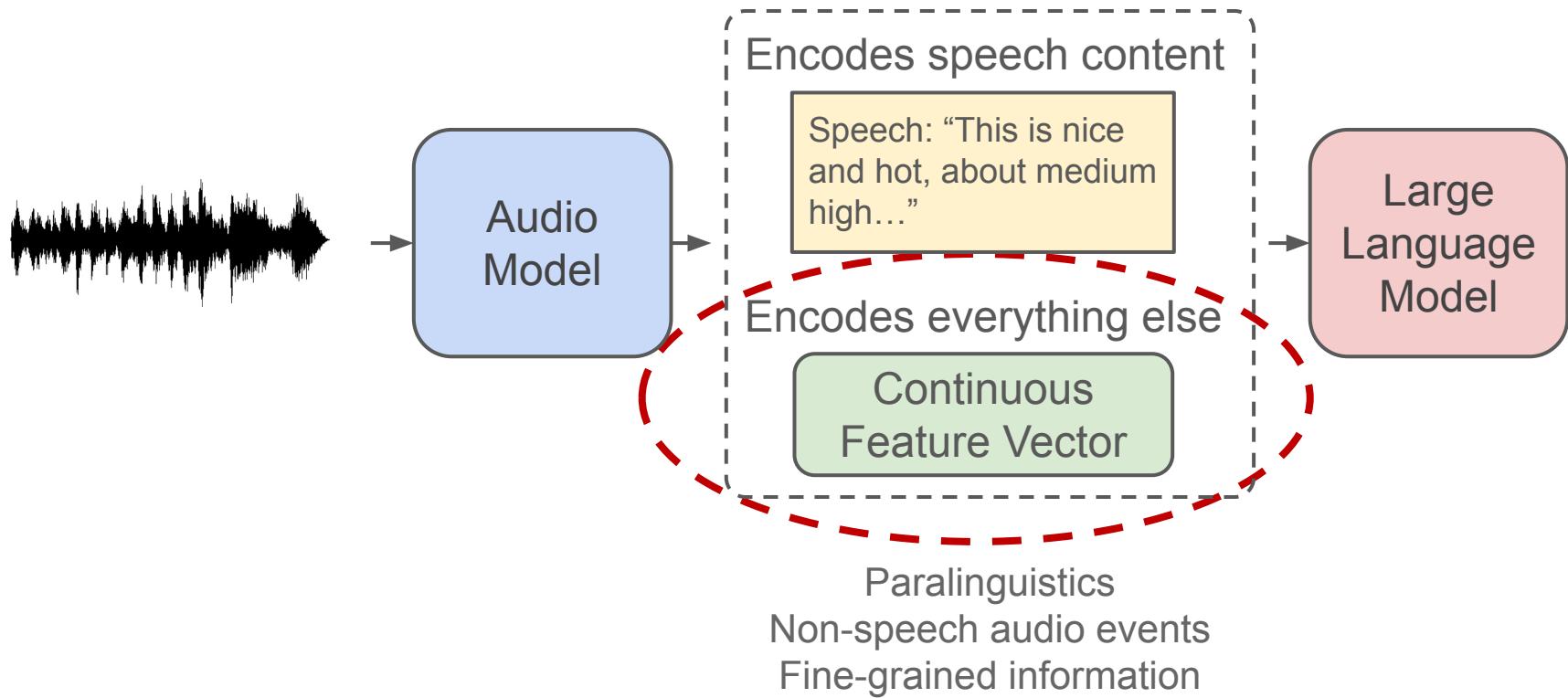
Can we just concatenate an audio model and an LLM?



Can we just concatenate an audio model and an LLM?



Continuous acoustic features are needed



LTU model architecture

Input

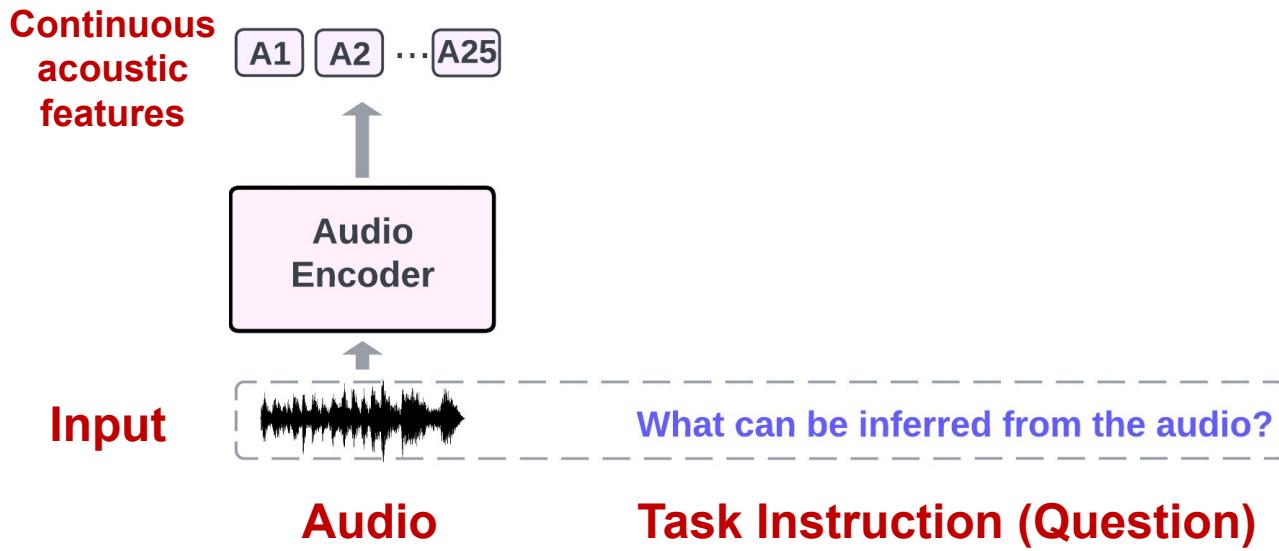


Audio

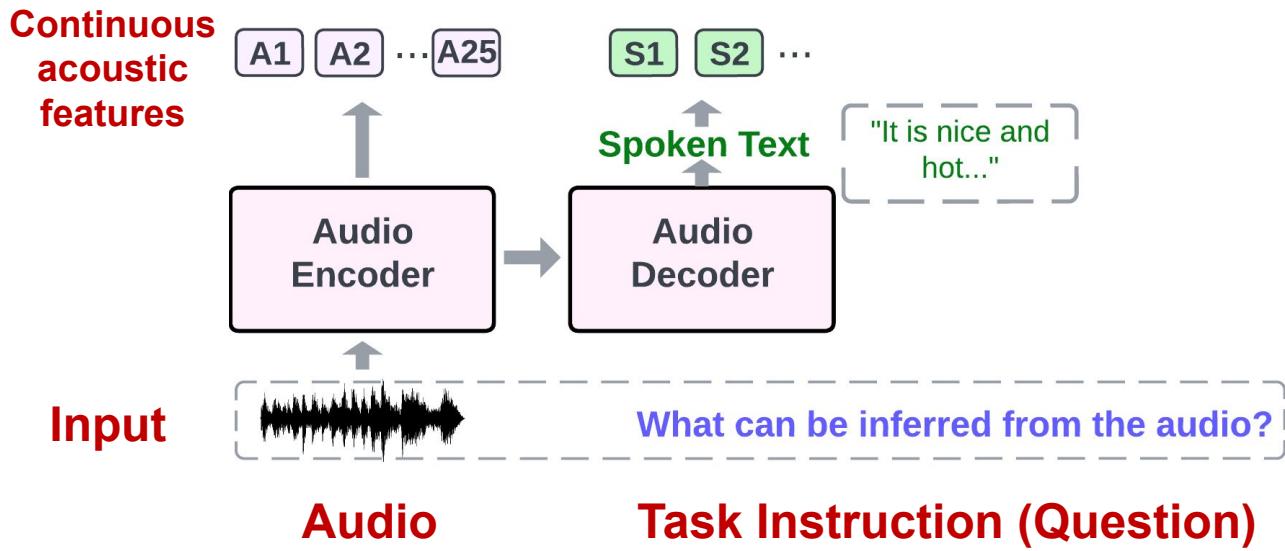
What can be inferred from the audio?

Task Instruction (Question)

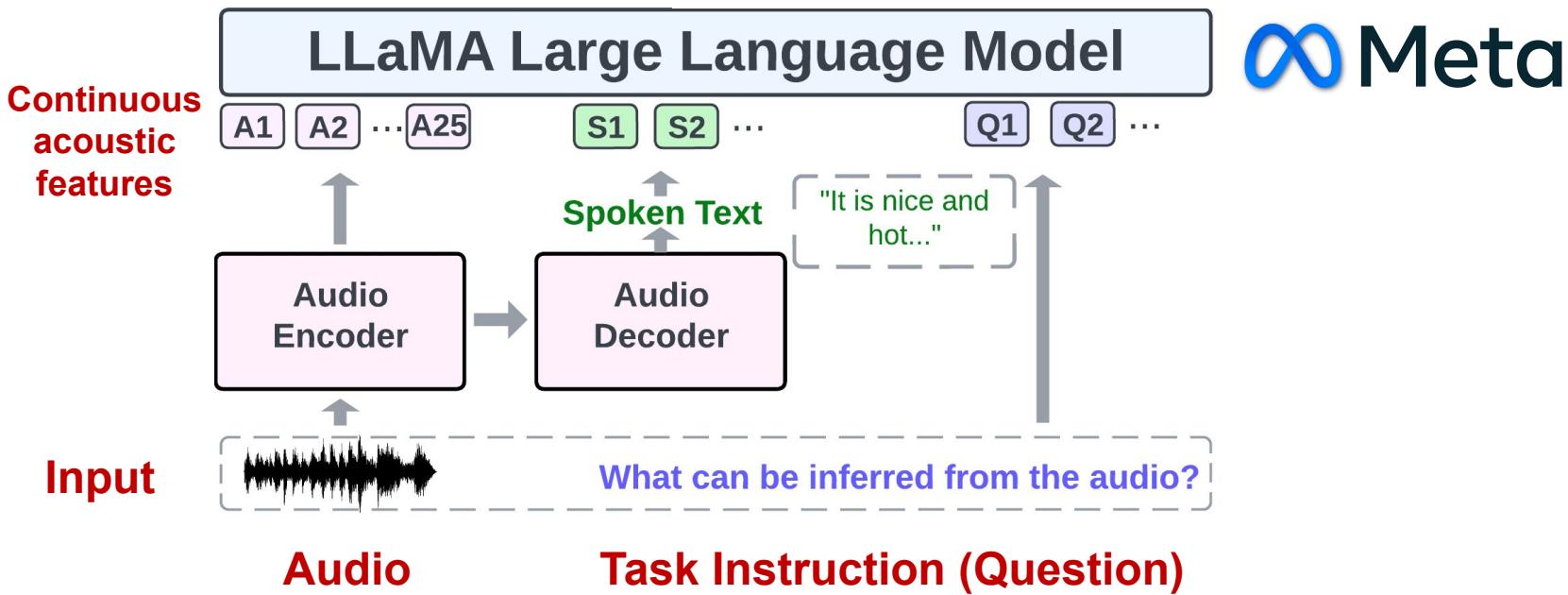
LTU model architecture



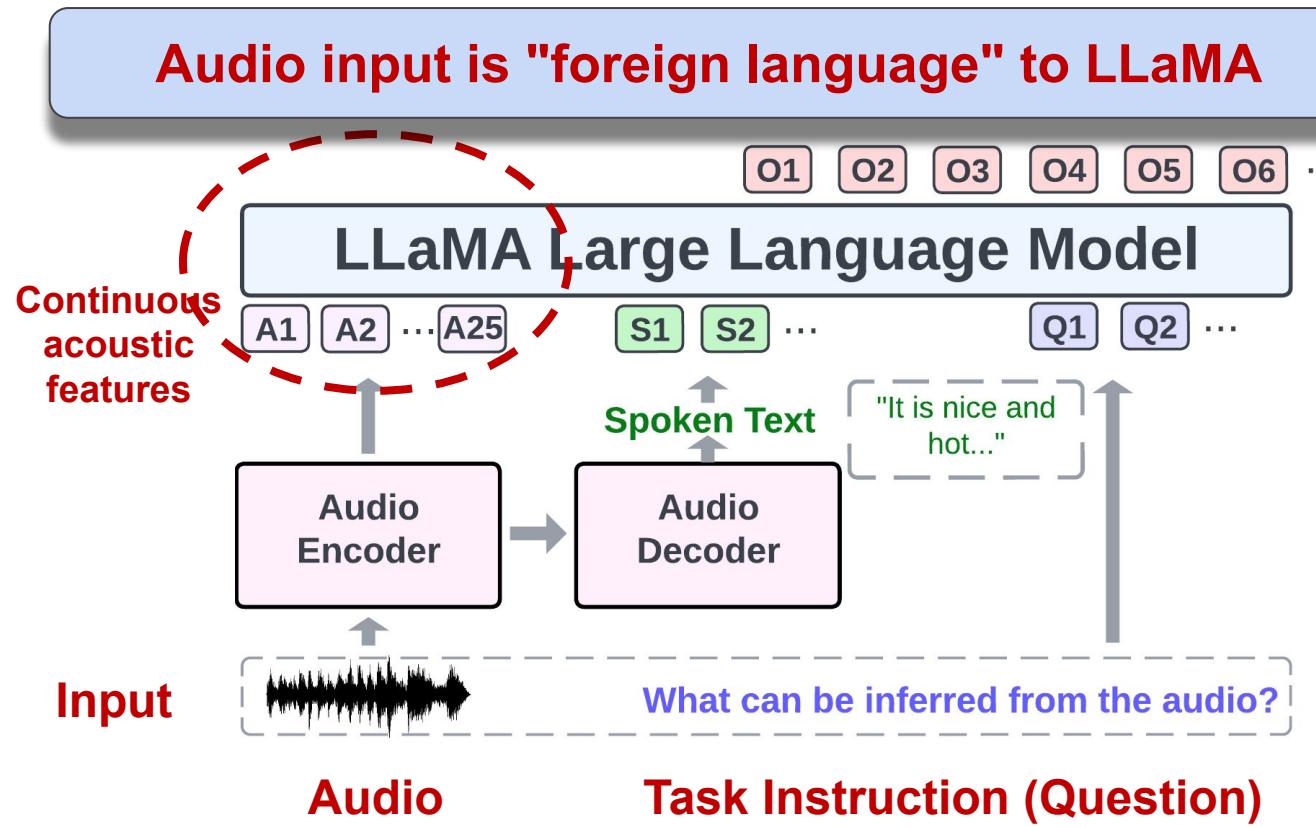
LTU model architecture



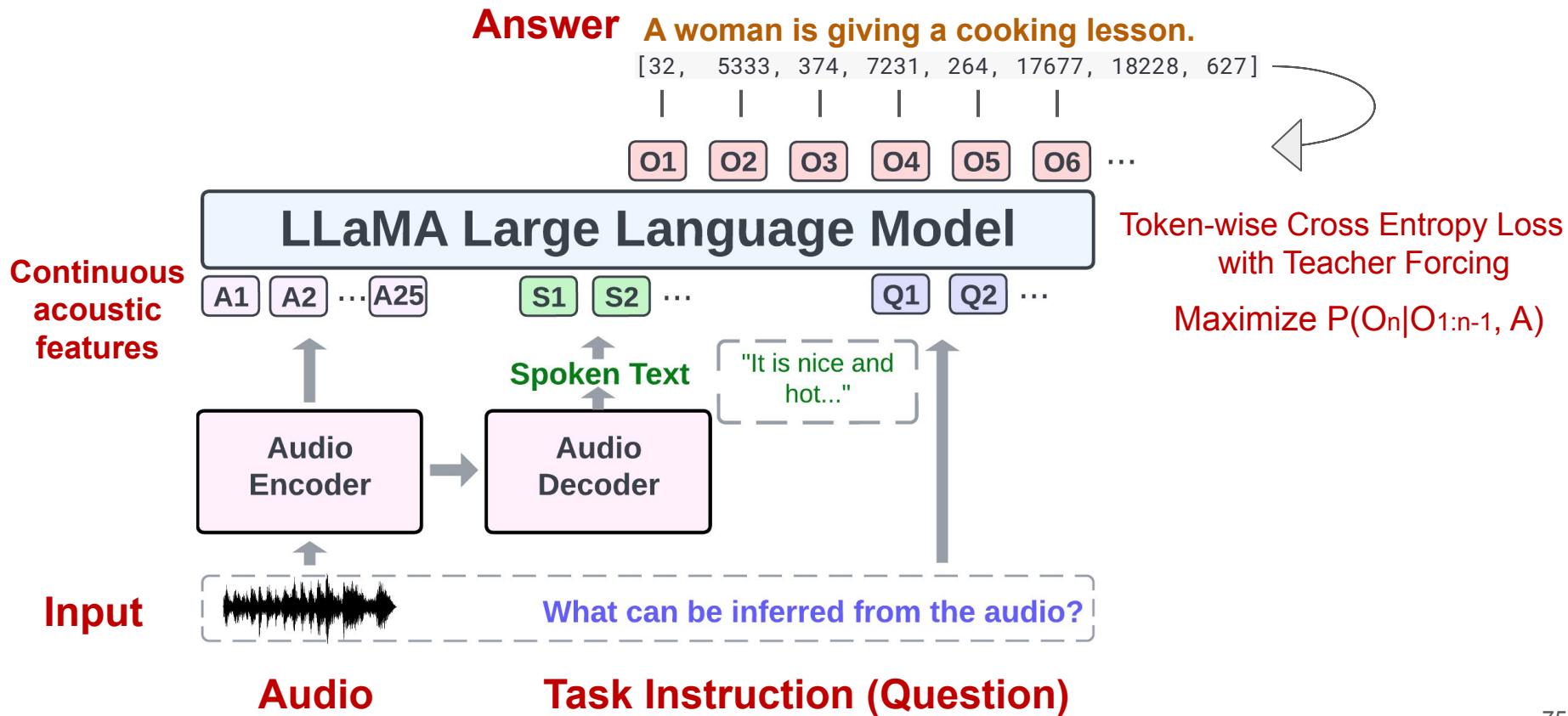
LTU model architecture



LTU model architecture

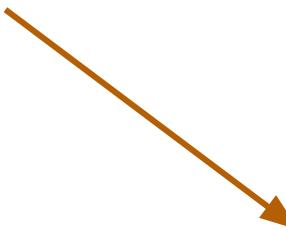


LTU training objective



LTU training data

Answer A woman is giving a cooking lesson.



(Audio, Instruction (Question), Answer) Tuples

Input



Audio

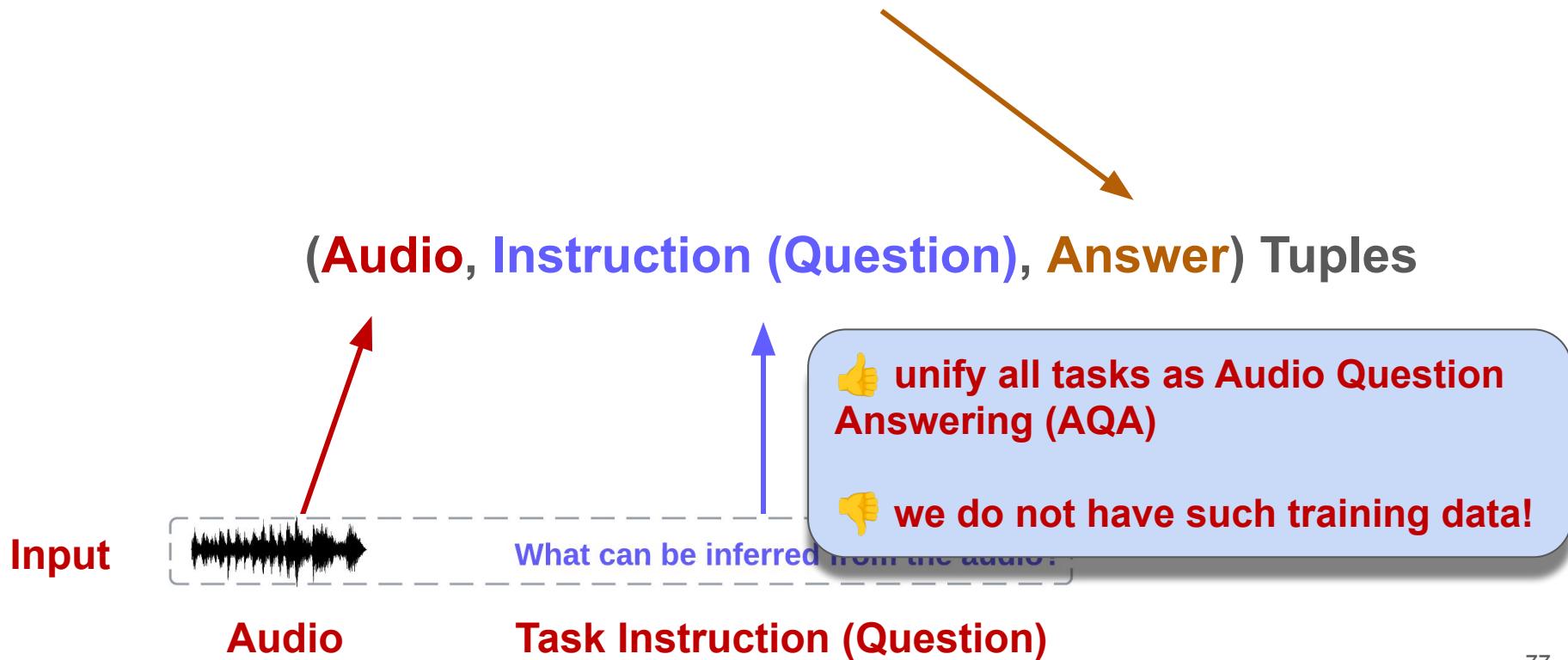
Task Instruction (Question)



What can be inferred from the audio?

LTU training data

Answer A woman is giving a cooking lesson.



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

**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.

Rule-Based Algorithm

Audio Event Classification
Question: Classify the sound events in the audio clip.
Answer: Ambulance (siren); Traffic noise, roadway noise; Accelerating, revving, vroom; Generic impact sounds.

Audio Captioning
Question: Write an audio caption describing the sound.
Answer: An ambulance siren echoes while traffic noise fades, and an engine revs.

Labels of Existing Datasets



**Closed-Ended
Audio Question Answer (AQA) Tuples**

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!

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?

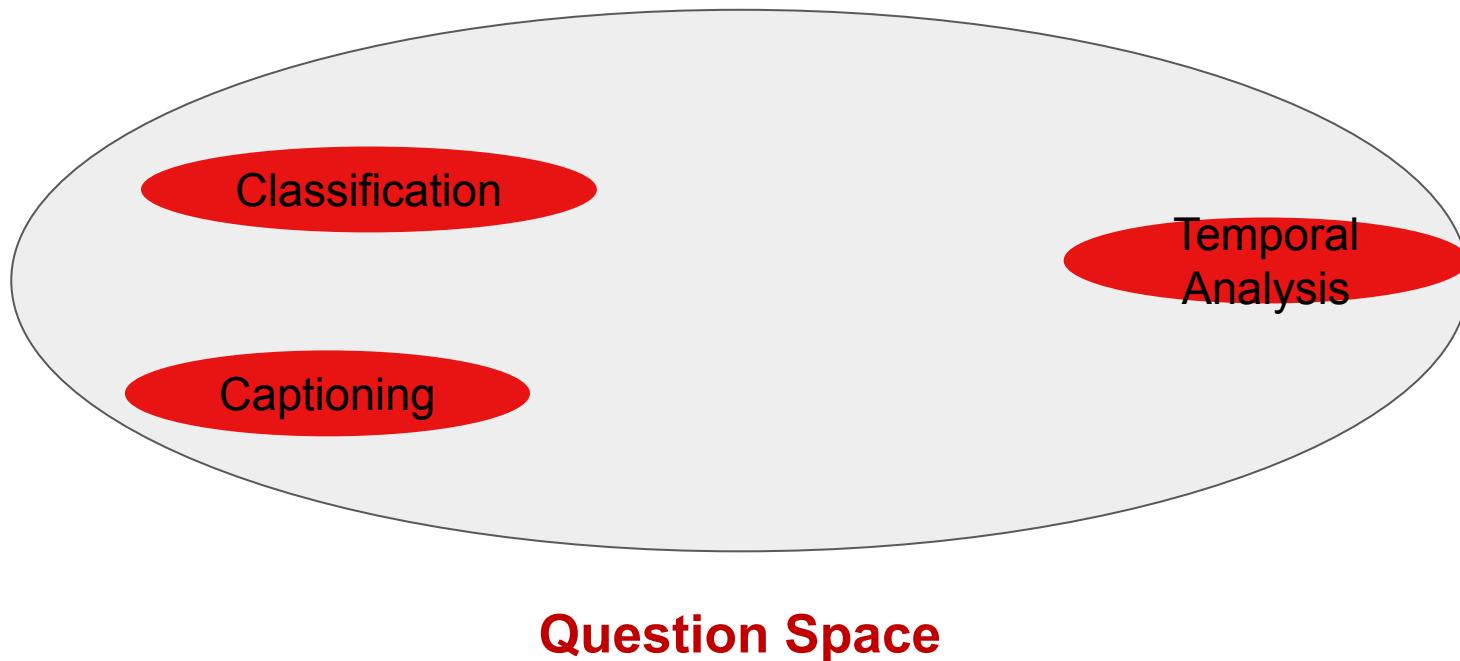
Answer: A woman speaks while frying food in a pan with background noise from a kitchen or dining room.



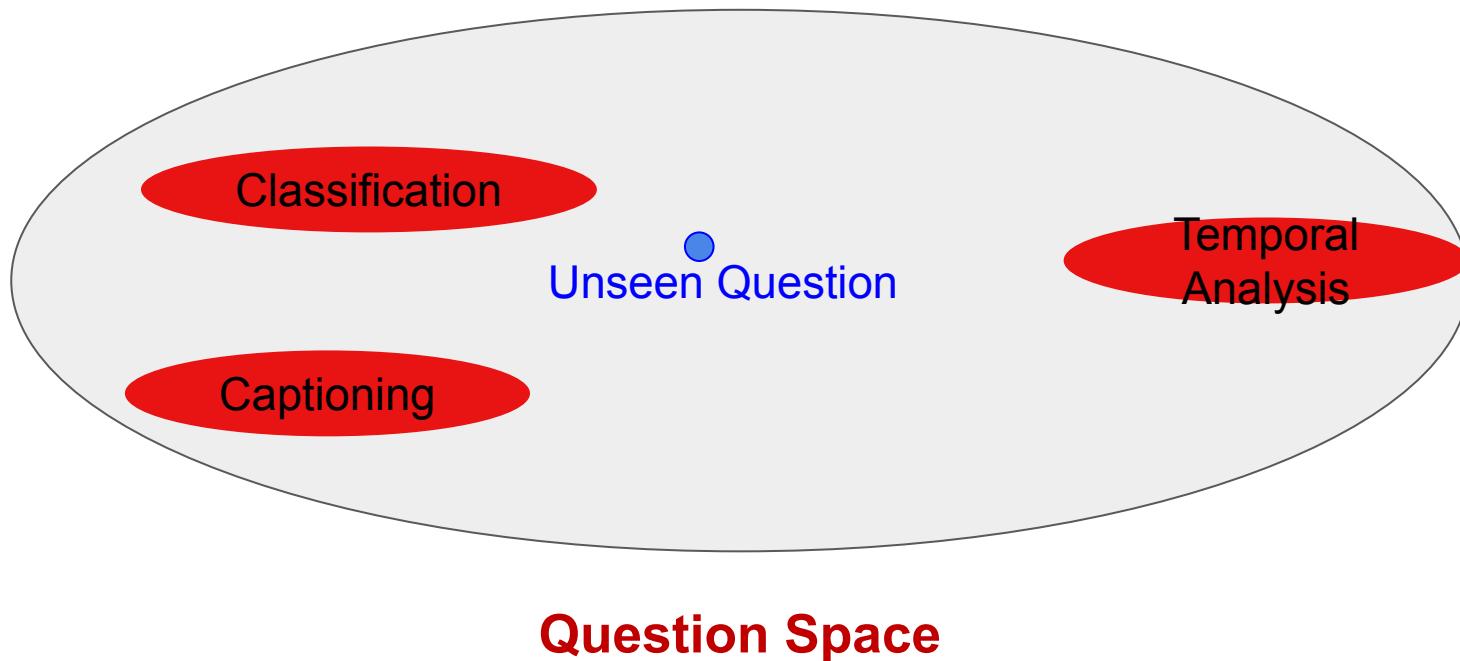
... The model cannot generalize!



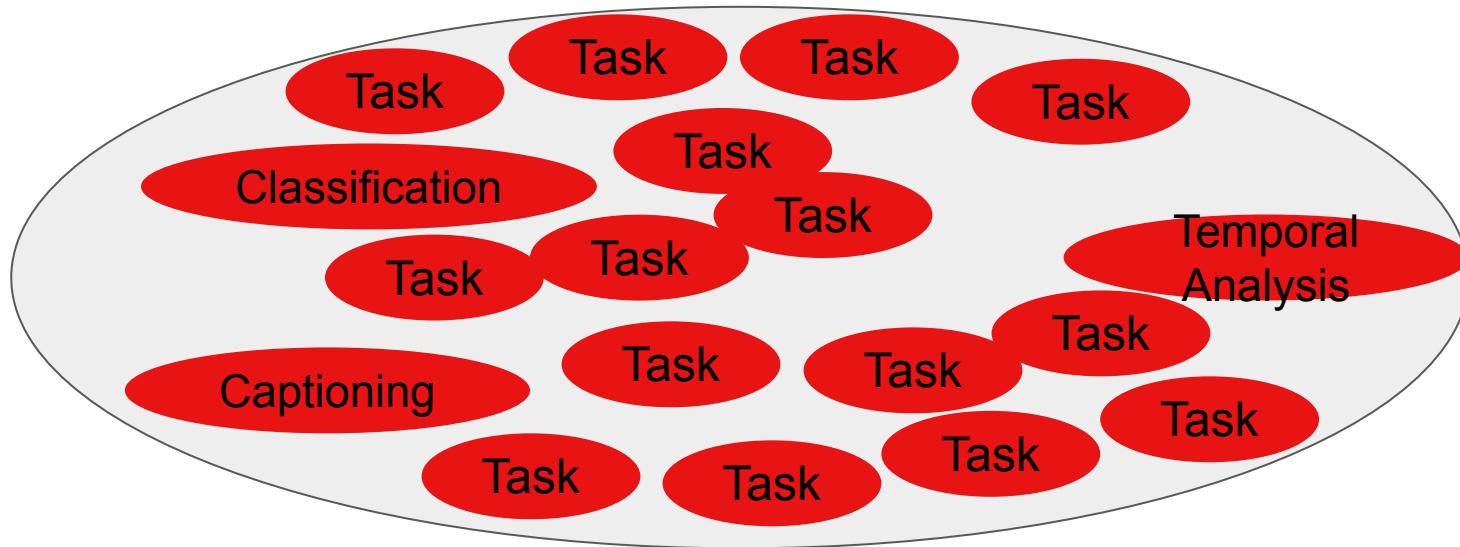
Closed-ended AQAs cannot train a generalized model



Closed-ended AQAs cannot train a generalized model

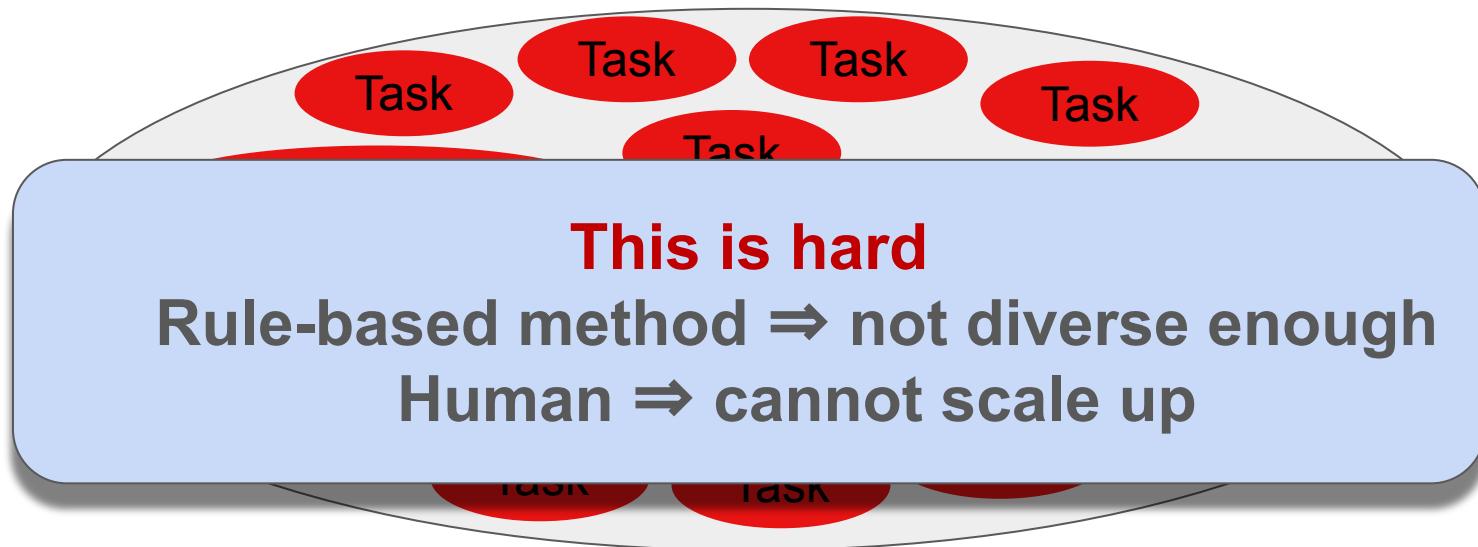


Ideally...



Question Space

Ideally...

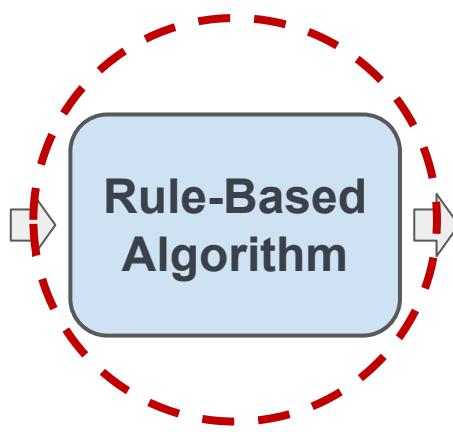


Question Space

Rule-based algorithm cannot generate open-ended AQAs

**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.



Audio Event Classification
Question: Classify the sound events in the audio clip.
Answer: Ambulance (siren); Traffic noise, roadway noise; Accelerating, revving, vroom; Generic impact sounds.

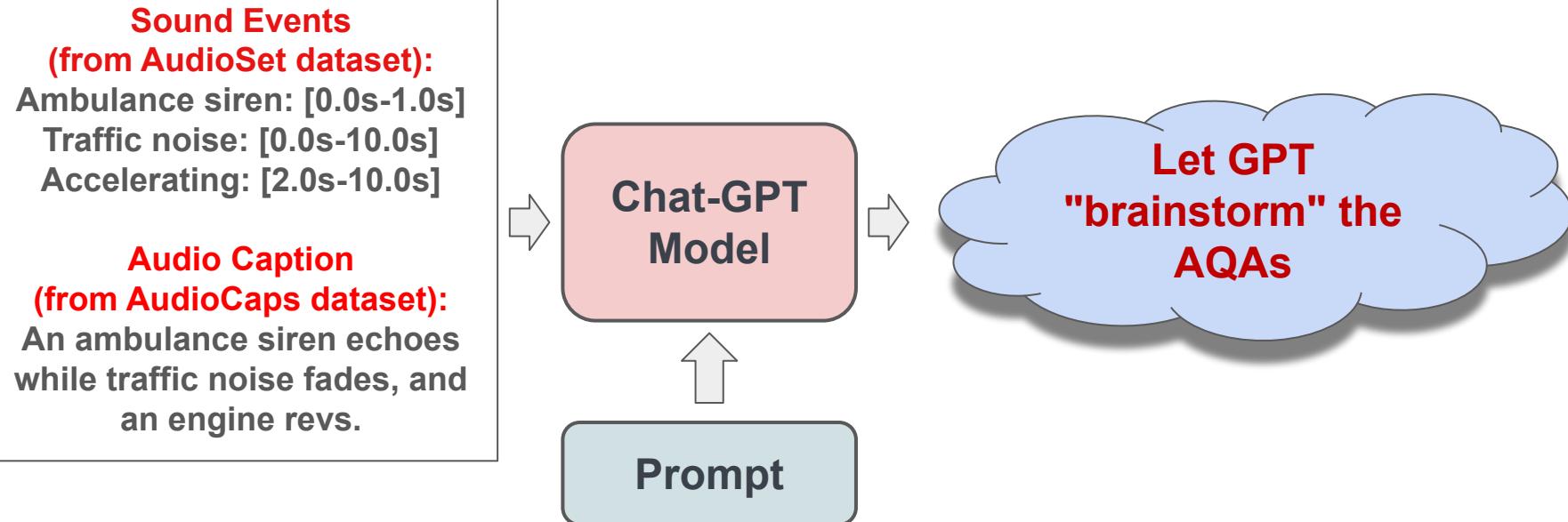
Audio Captioning
Question: Write an audio caption describing the sound.
Answer: An ambulance siren echoes while traffic noise fades, and an engine revs.

Labels of Existing Datasets



**Closed-Ended
Audio Question Answer (AQA) Tuples**

Audio instruction generation with GPT assistant



Labels of Existing Datasets

A closer look at the GPT prompt

GPT Prompt

Based on the audio clip, generate 10 different types of complex open-ended questions that require step-by-step thinking, and corresponding answers.

Task

The following information is provided: the sound events appear in the audio clip, together with its acoustic features, and corresponding onset and offset time stamps. A description of the content of the audio clip is also provided.

Provided Information

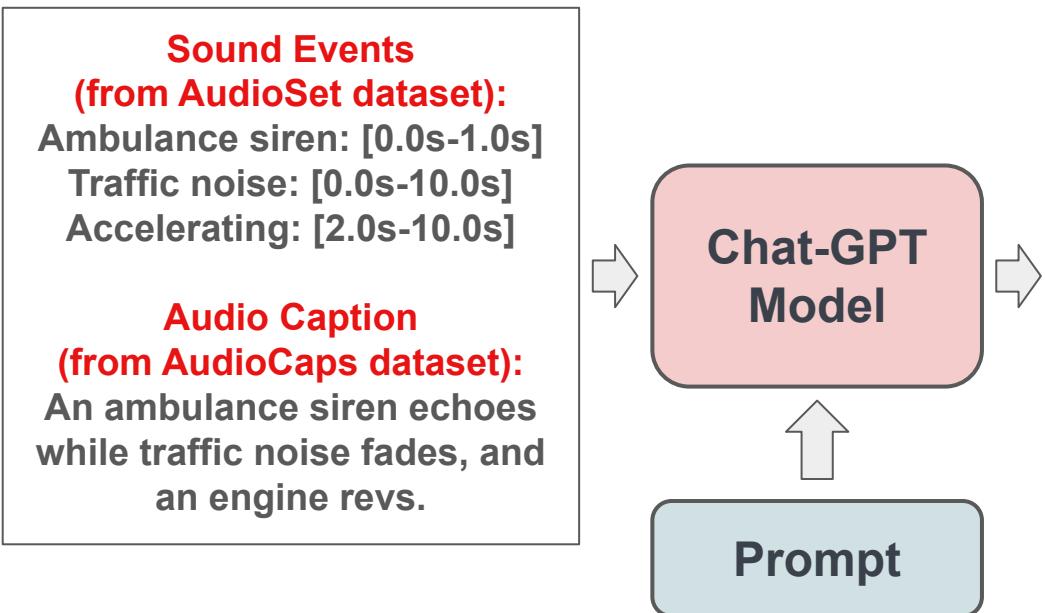
Questions should be about the audio, e.g., which sound event is recognized and why (e.g., based on its acoustic feature), what can be inferred based on the combination of sound events; the temporal relationship between the sound events and what can be inferred from that; the potential scenario that such an audio clip could happen, if the audio clip is special (e.g., urgent, funny, interesting, abnormal, unique, etc) and why, what mood or atmosphere this audio clip conveys, etc.

Sample Questions

The more complex and diverse the question, the better.

Requirements

Audio instruction generation

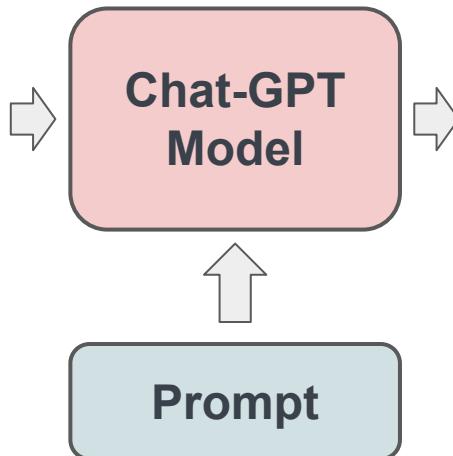


Labels of Existing Datasets

Audio instruction generation

**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.



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.

Labels of Existing Datasets

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

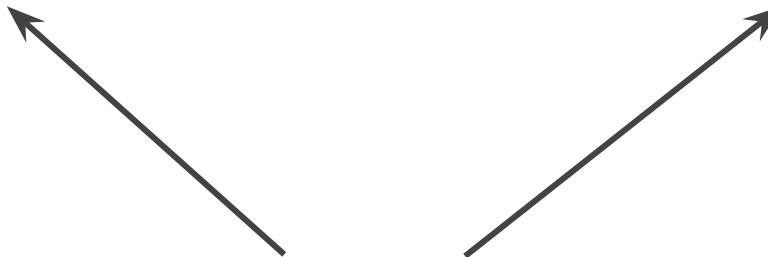


Question: What is the acoustic feature that distinguishes the sound of the ambulance siren from the generic impact sounds?

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



Not in original label

Generated based on ChatGPT's knowledge

abstract



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.

↑
Not in original label
Generated based on ChatGPT's common sense

Sound

low-level

Understand

abstract

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

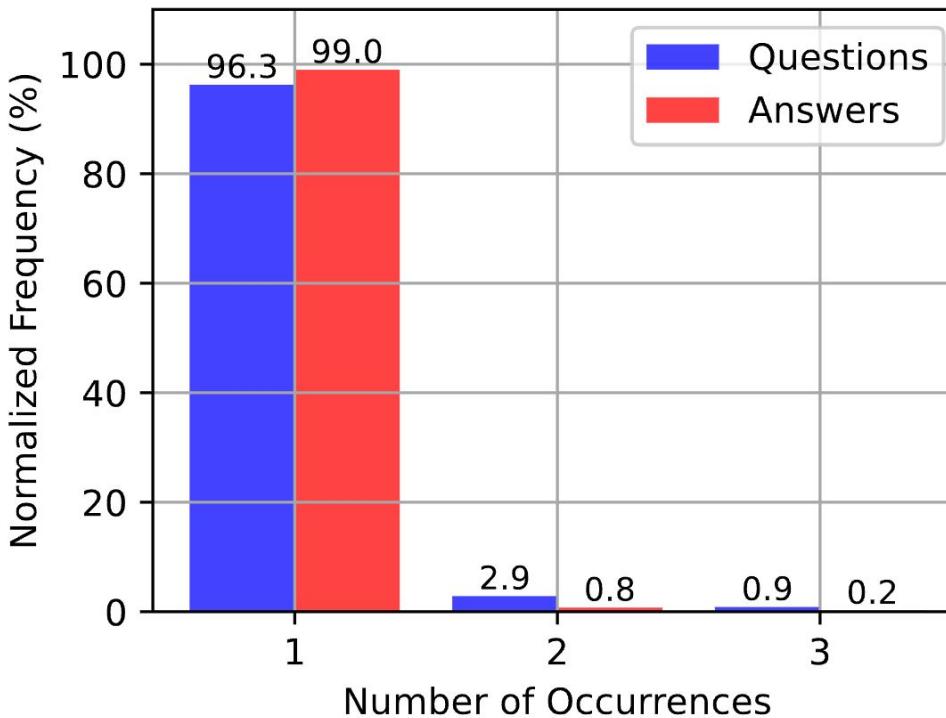
Understand

Mood

abstract

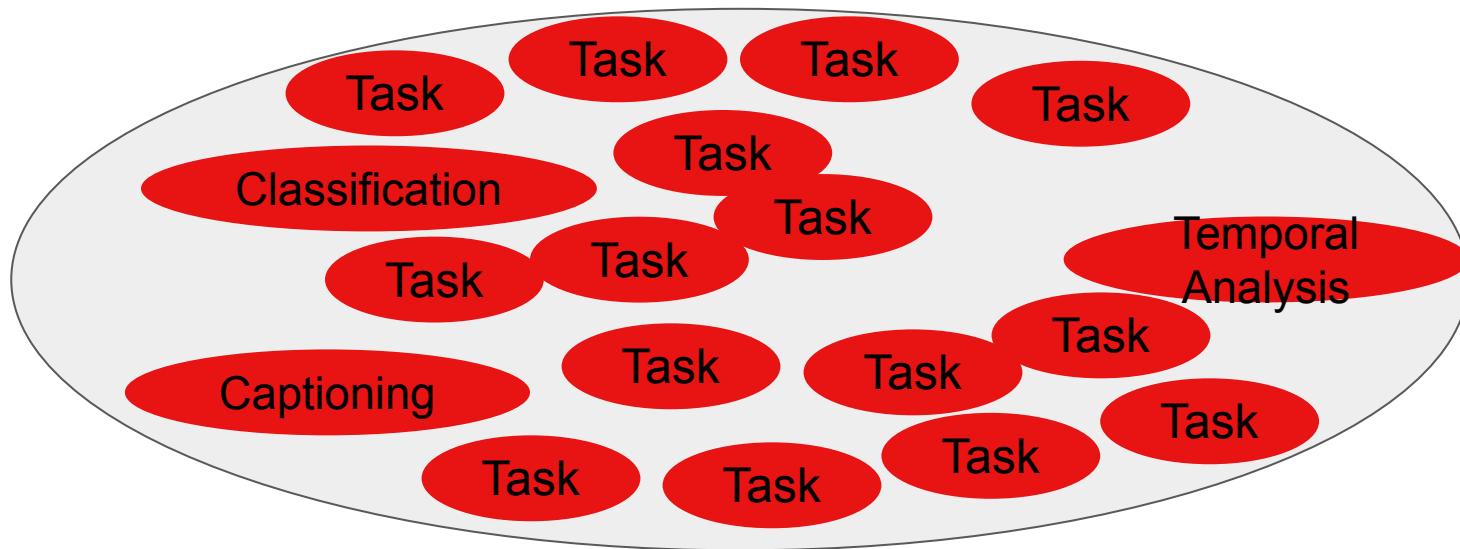


Generated open-ended questions are diverse



96.3% questions and 99.0% answers appears only once

Ideally...



Question Space

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) [5]</i>												
AS-Strong [30]	x	x	x	x	-	-	-	-	102k	683k	901k	
AudioSet [31]	x	-	x	x	-	-	-	x	500k	538k	184k	
VGGSound [102]									12.1k	2.57k	907k	
FSD50K									403k		403k	
AudioCaps [4]									478k		478k	
FreeSOUND [2]									791k		791k	
Clotho [3]									89k		89k	
SoundBNT [5]									10k		10k	
Sum									3,763k			
<i>Speech Datasets</i>												
IEMOCAP									83k			
LibriTTS									418k			
VoxCeleb									926k			
MOSEI [1]									355k			
Sum									1,784k			
<i>Joint AQA</i>												
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

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.

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.

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.



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

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.

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.

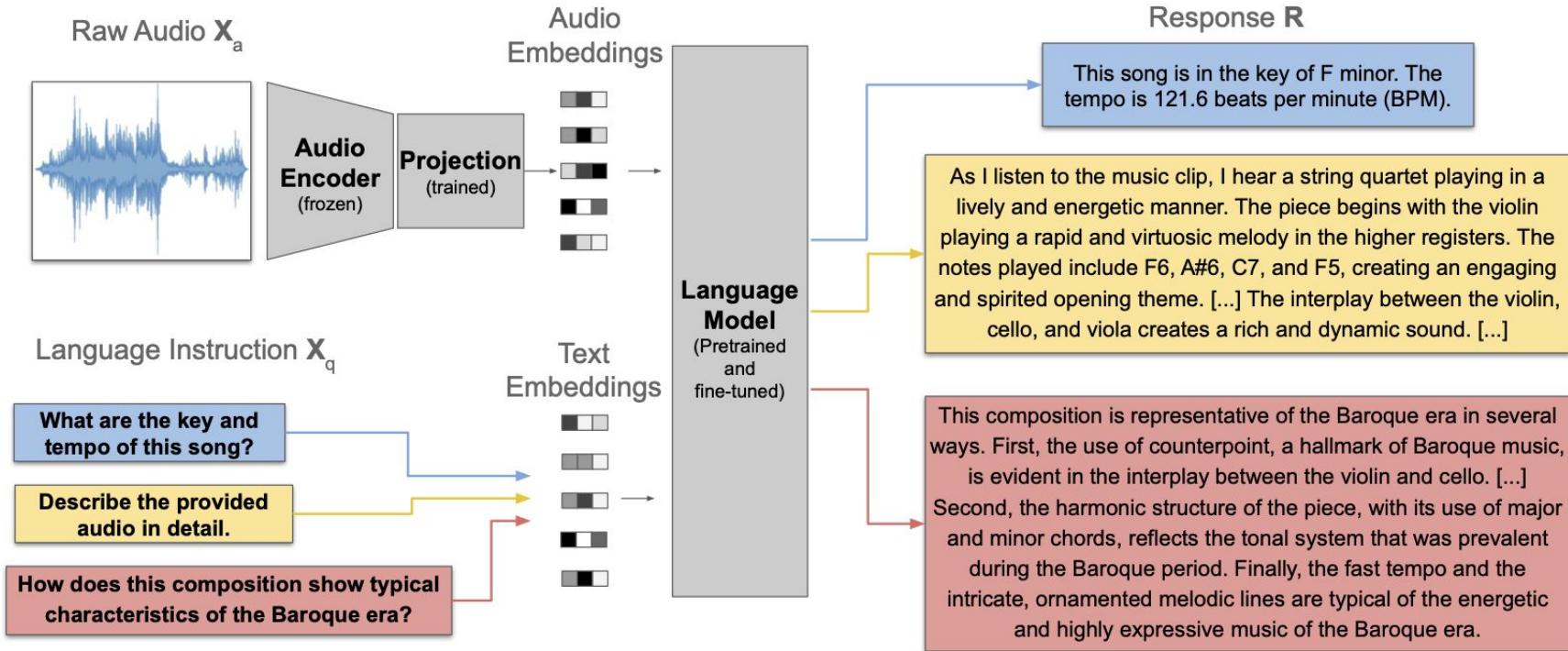


Question: Is this music appropriate to play in a primary school?

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.



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.

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 '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。





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. [\[Paper\]](#) [\[Github\]](#)

LTU is authored by Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass (MIT & MIT-IBM Watson AI Lab).

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: [\[sample audios from AudioSet evaluation set\]](#)

audio_path

Drop Audio Here
- or -
Click to Upload

question

Why this audio is fun?

Audio Meta Information

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

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.



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

0:00 / 0:10

Audio 1 of 5, Audio ID: MnEZwnRuQR8.flac

Question: What atmosphere or mood do the sound events in the audio clip convey?

Answer: The sound events in the audio clip convey a tense and potentially dangerous atmosphere, with the explosion and gunfire suggesting potential danger.

Did this answer directly address the question? (Regardless of whether the answer is correct or not.)

Yes, the answer directly addresses the question (regardless of whether the answer is correct or not) No, the answer is not related to the question

If select "no", please explain.

Is this answer factually correct?

Correct Partially Correct Wrong

If select "partially correct" or "wrong", please explain which part is wrong.

Please compare the following two answers:

Question: What atmosphere or mood do the sound events in the audio clip convey?

Answer 1: The sound events in the audio clip convey a tense and potentially dangerous atmosphere, with the explosion and gunfire suggesting potential danger.

Answer 2: The audio clip conveys a tense and dramatic mood due to the combination of narration, male speech, explosive sounds, and artillery fire.

Which answer is better?

Answer 1 is better Answer 2 is better

Please answer the question based on the audio by yourself (at least 5 words):

Ask a new question based on the audio and provide the corresponding answer:

Question (at least 5 words):

Answer (at least 5 words):

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.



Q: "What type of sound is it?"

LTU Answer: "The audio clip contains the sounds of a train, clickety-clack, and steam whistle blowing at a high pitch."

Human Answer: "It was train moving sound."

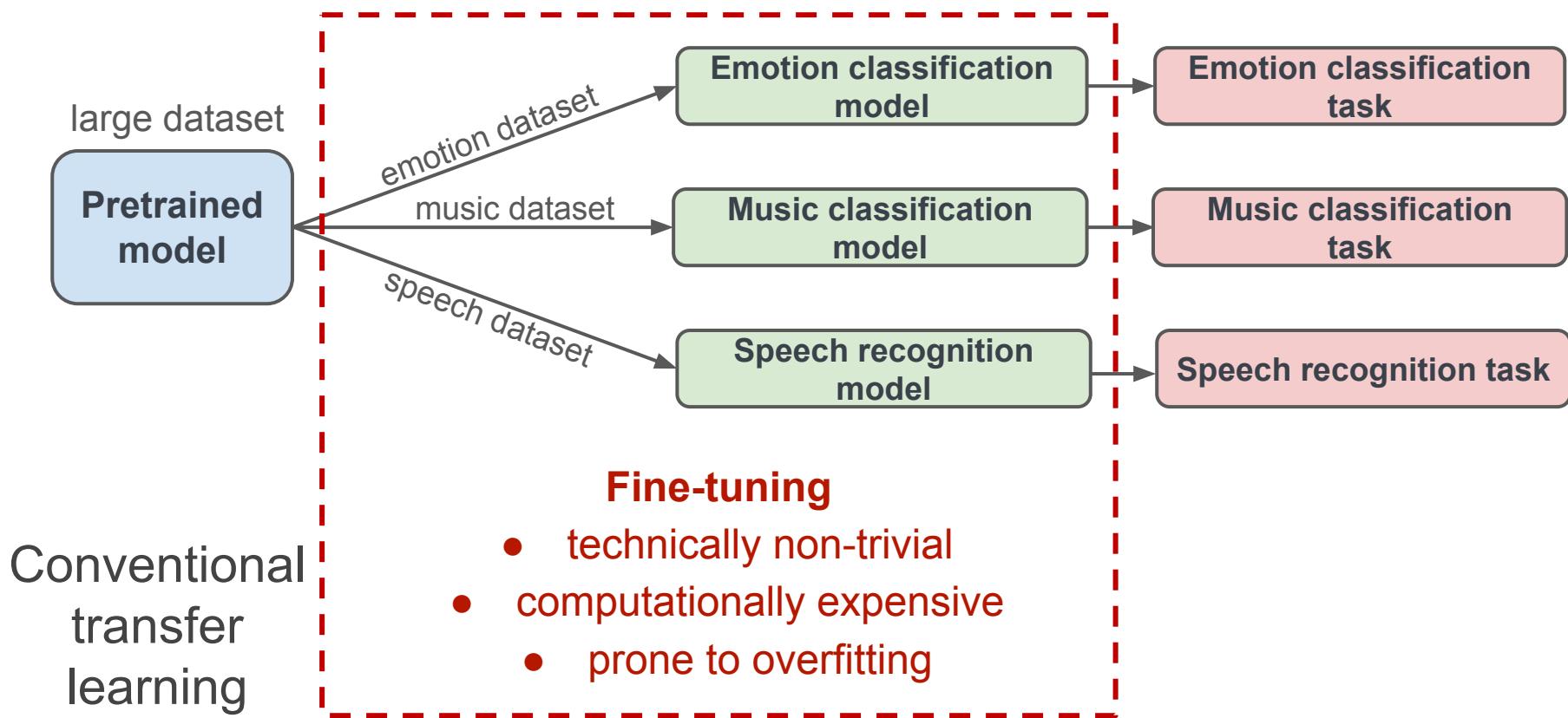
Advantage of LTU - 1/ understanding

LTU jointly understands
spoken text

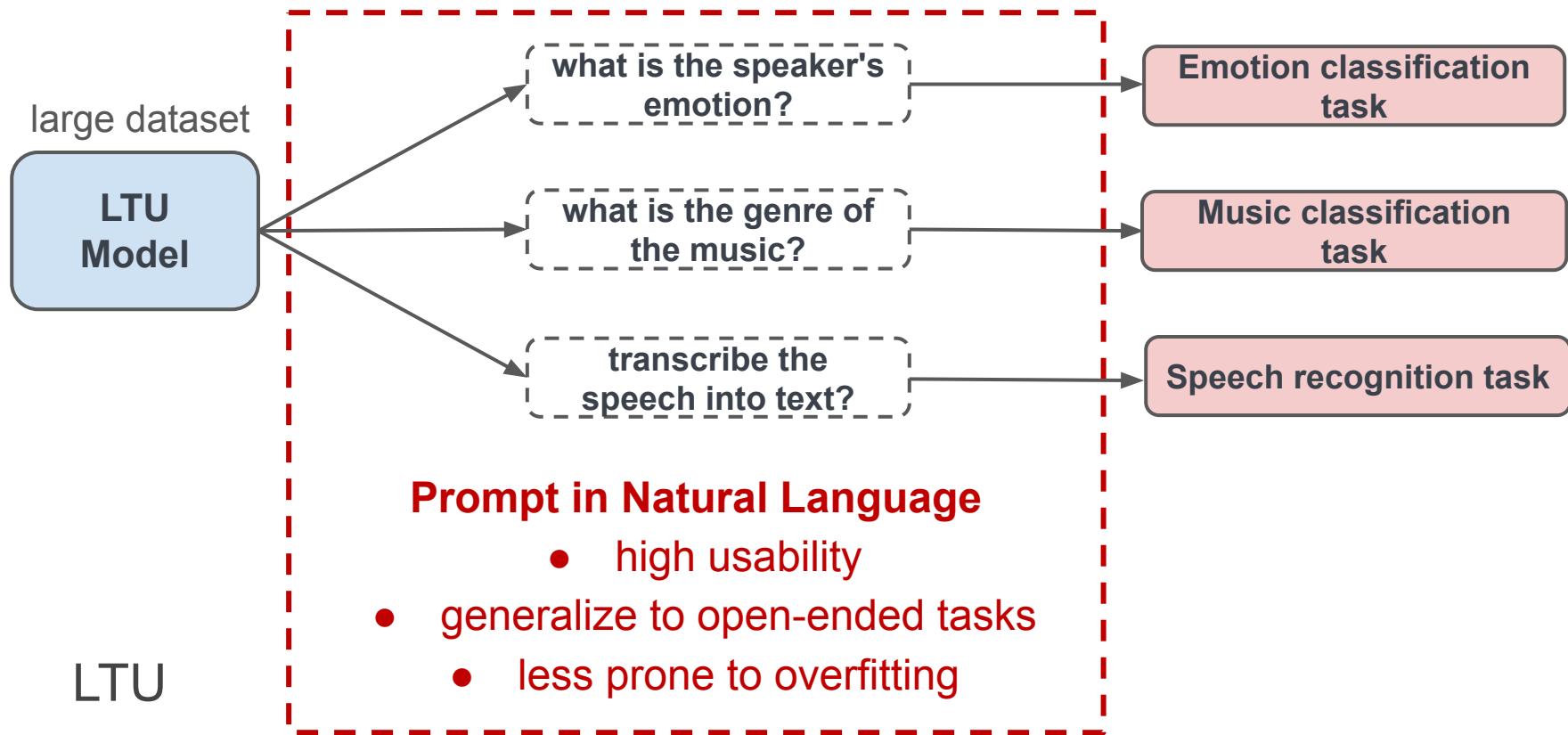
speech paralinguistics (e.g., emotion)
non-speech audio events

almost ***everything*** in audio

Advantage of LTU - 2/ broader capacity



Advantage of LTU - 2/ broader capacity

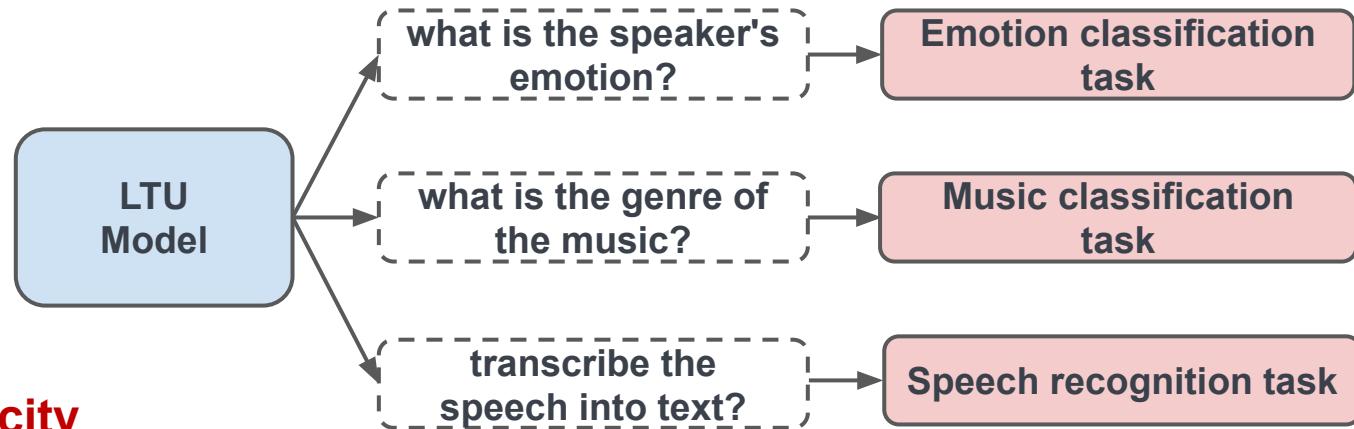


LTU is a step closer to artificial general intelligence (AGI)

1/ Understanding

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

2/ Broader capacity

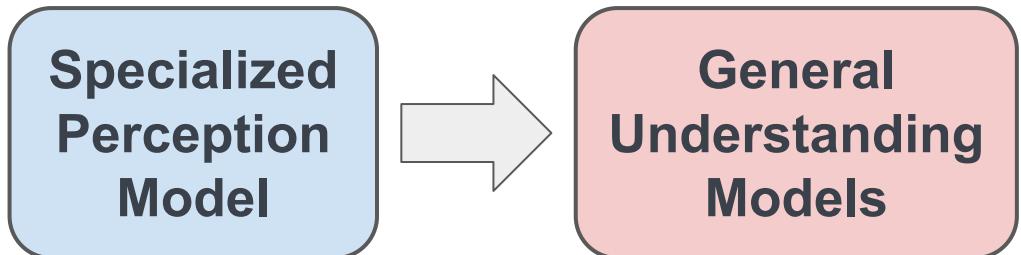


From a broader perspective...

There are many signals:

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

For many natural signals, we desire:



and many others ...

Outline

Part 1

Listen

General audio
recognition

Part 2

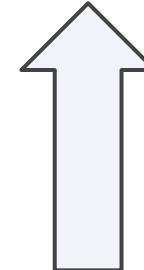
Understand

Audio large language model

Part 3

Future Research

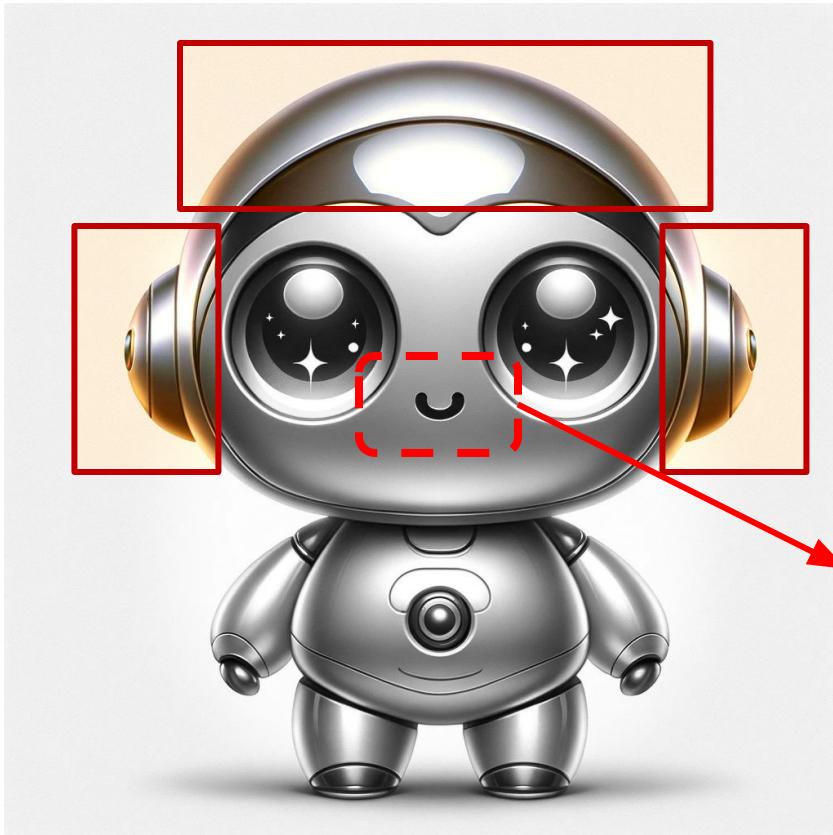
- Advancing audio AGI
- Audio AGI for social good
- Responsible audio AGI



Future Research

1 Advancing Audio AGI

Further unify audio generation



We modeled *listening and understanding*

We still left **sound generation** ability

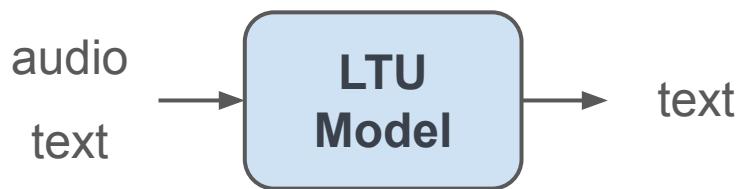
1/ Supports all audio tasks



Supports

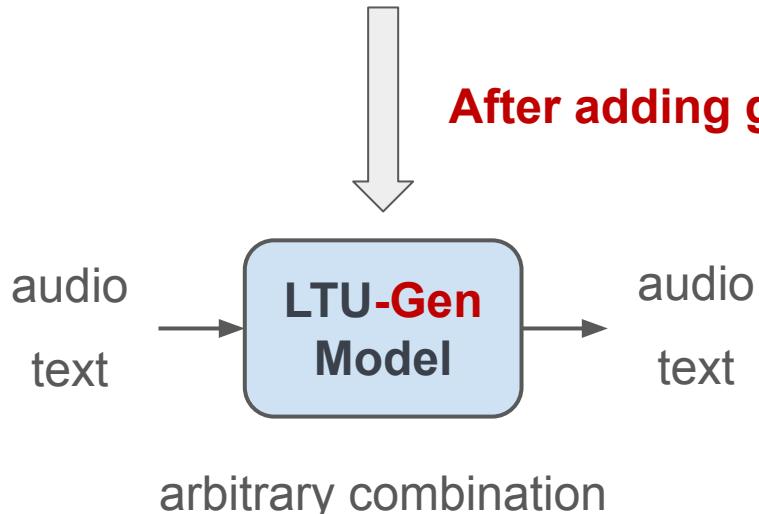
- classification
- understanding
- audio question answering

1/ Supports all audio tasks



Supports

- classification
- understanding
- audio question answering



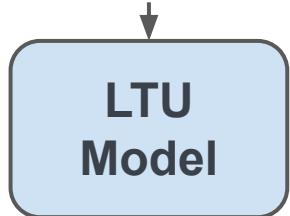
Further supports

- audio \Rightarrow audio
 - speech enhancement/separation
- text \Rightarrow audio
 - text-to-speech synthesis
 - text-based sound generation

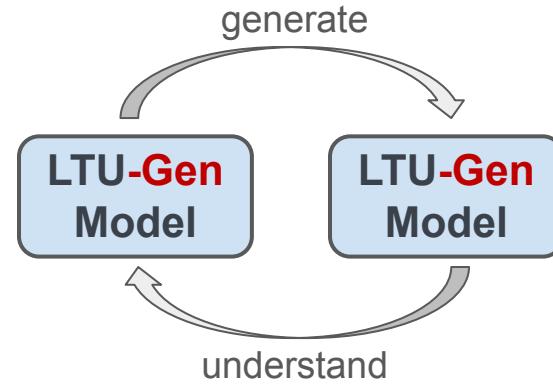
Almost all audio tasks!

2/ Reforming audio AI training

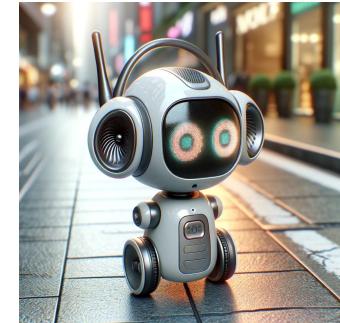
millions of (audio, label) pairs



Learn from samples



Self-play training



Learn by interacting with virtual/physical world

Future Research

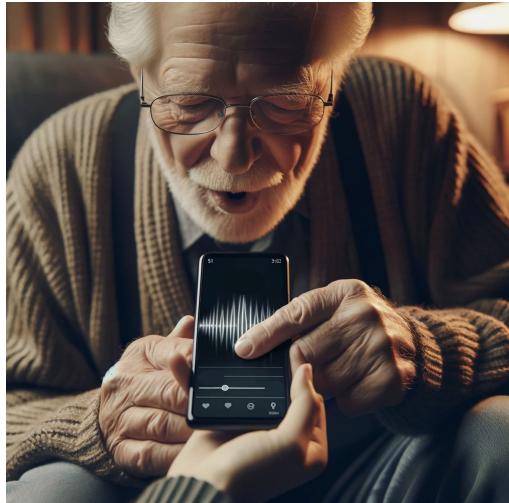
2 Audio AGI for Health

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

Audio AI for Health



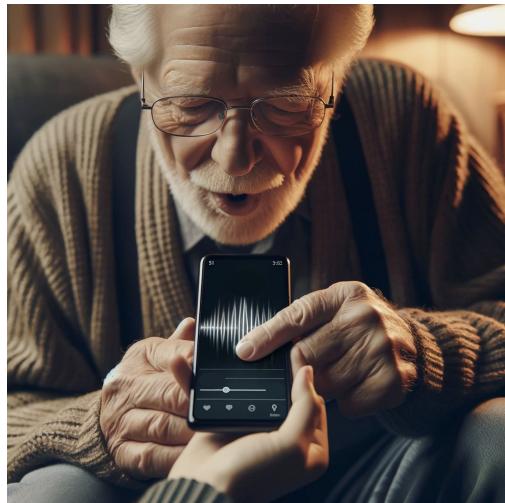
**Which is abnormal?
Why?**



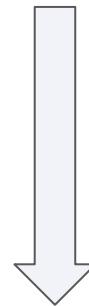
The second recording is abnormal

- Slurred speech
- Irregular rhythm

Amyotrophic Lateral Sclerosis (ALS)



2018 November

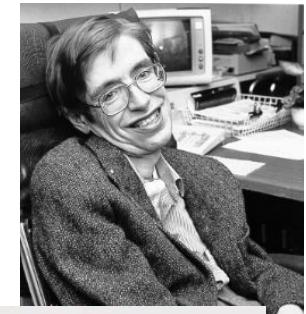


2019 June

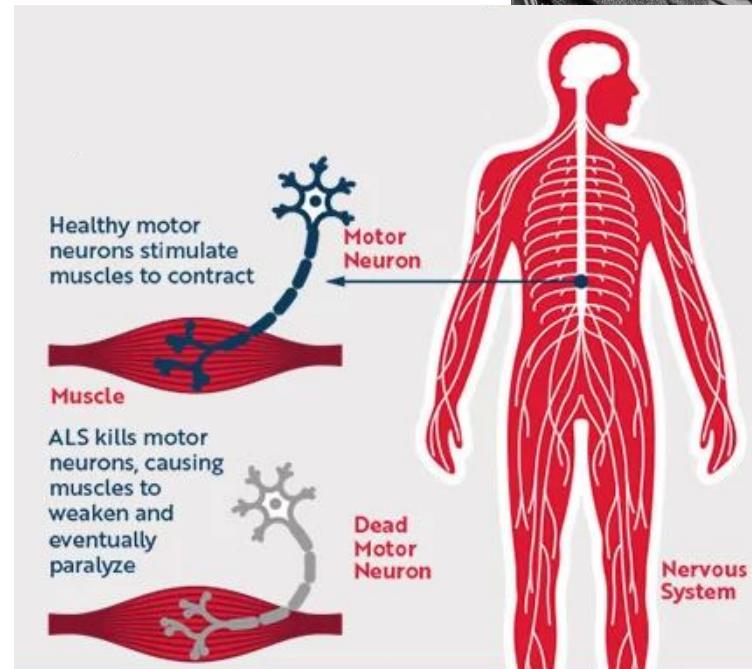
7 months

}

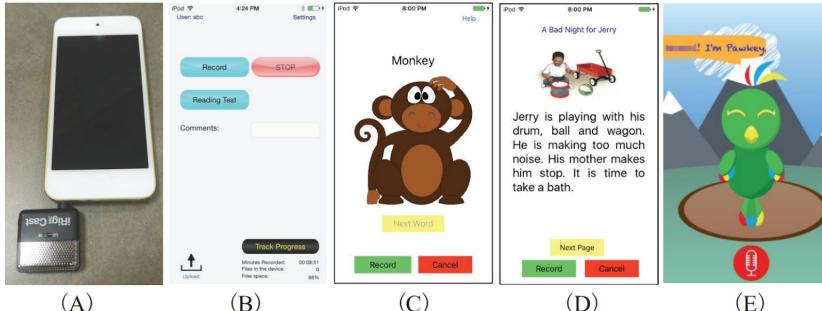
Amyotrophic Lateral Sclerosis (ALS)



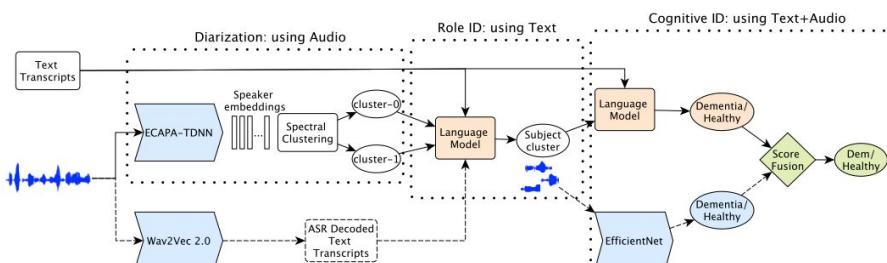
motor neurons gradually degenerate
⇒
loss of voluntary muscle movement
⇒
difficult to articulate words
⇒
speech becomes slurred and slow



Previous work



Child Autism (ASD) Detection



Dementia Detection



Emotional Disorder Detection

Yuan Gong et al., "Topic Modeling Based Multi-Modal Depression 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-BCB 2018.

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

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

Conventional AI

Prediction

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

“ALS speech is slurred”

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”

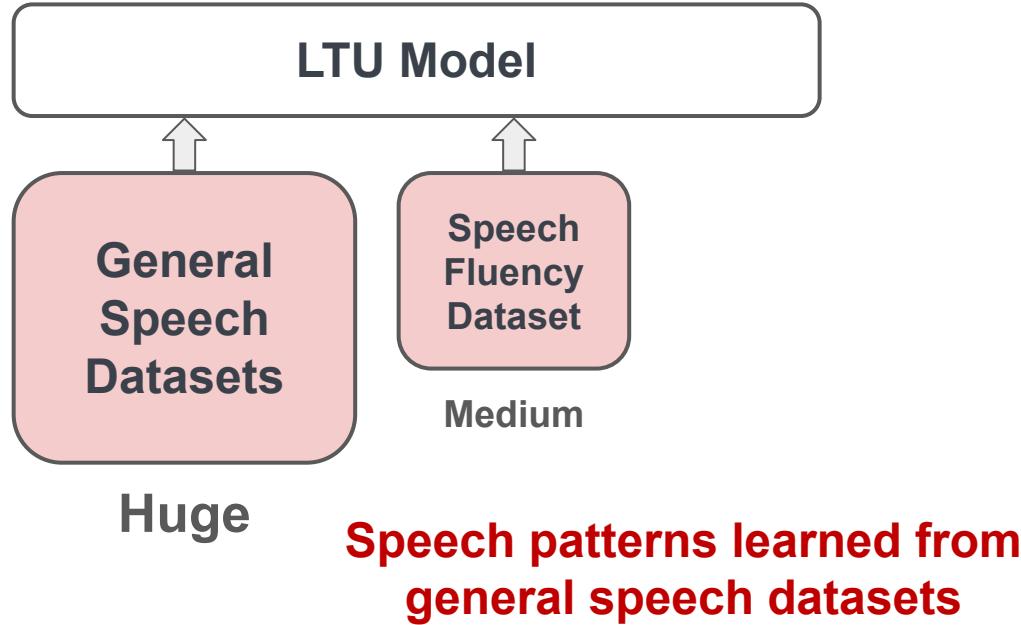
Do we have to fully rely on speech health datasets?

Recall that:

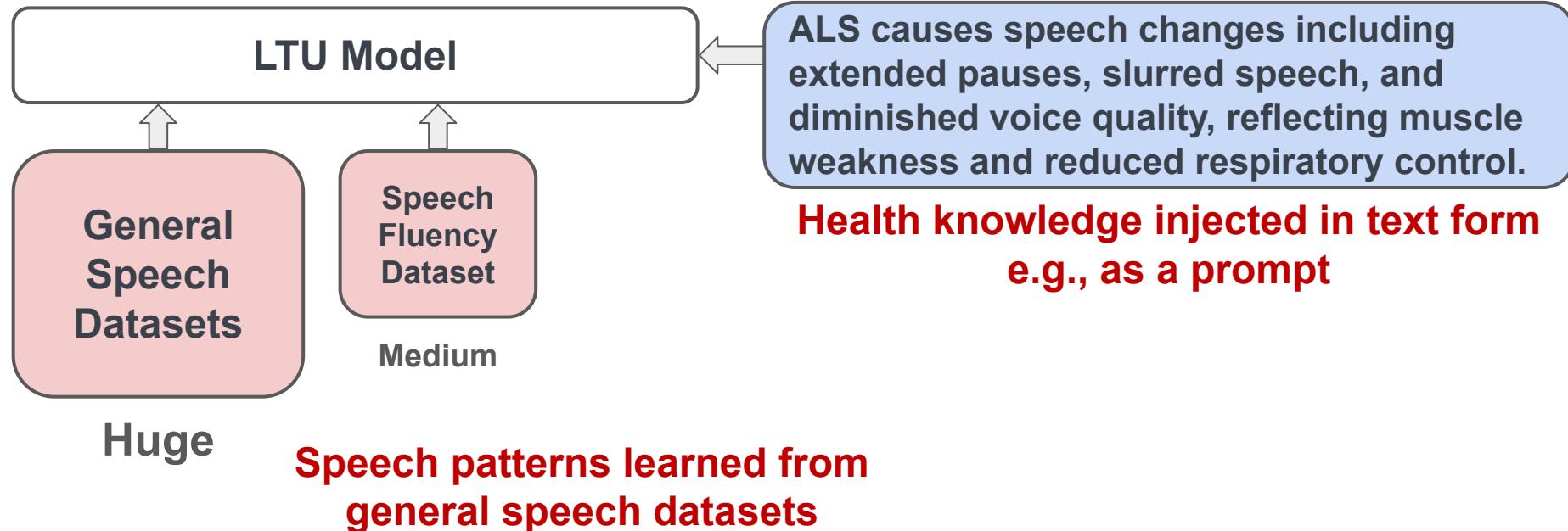
- We can identify which recording is abnormal **without** knowledge of ALS
- After learning about ALS, we can **associate** speech patterns with ALS

Humans not need learn from a large speech health dataset

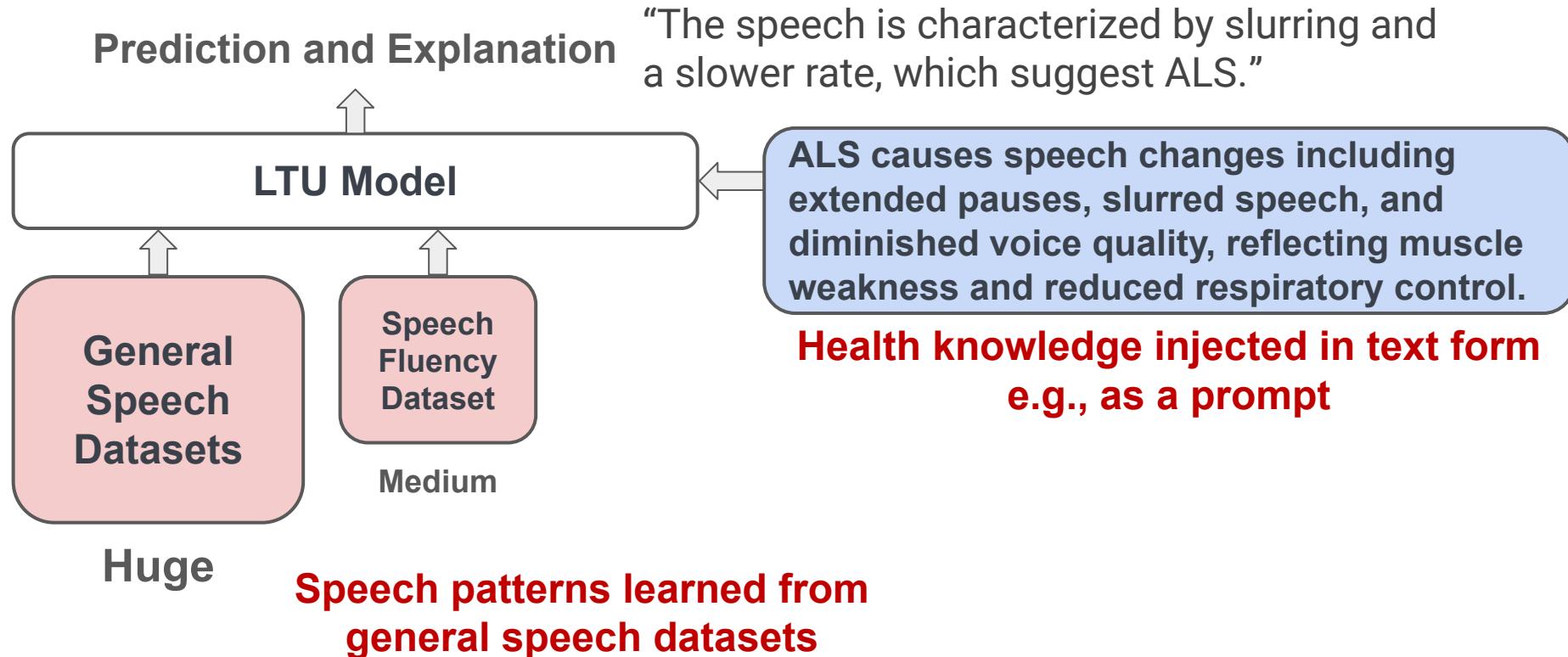
Audio AGI for health



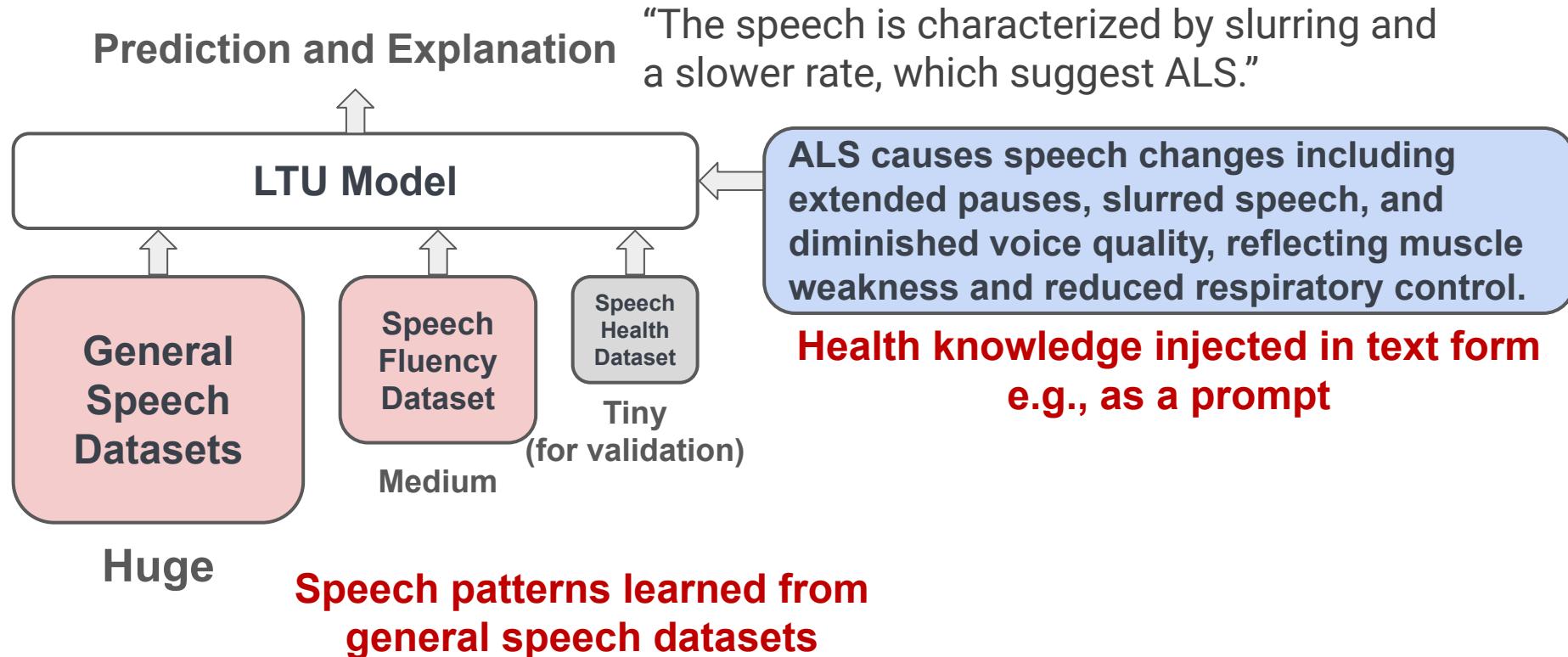
Audio AGI for health



Audio AGI for health



Audio AGI for health



Future Research

3 Responsible Audio AGI

Responsible audio AGI - security



AP News

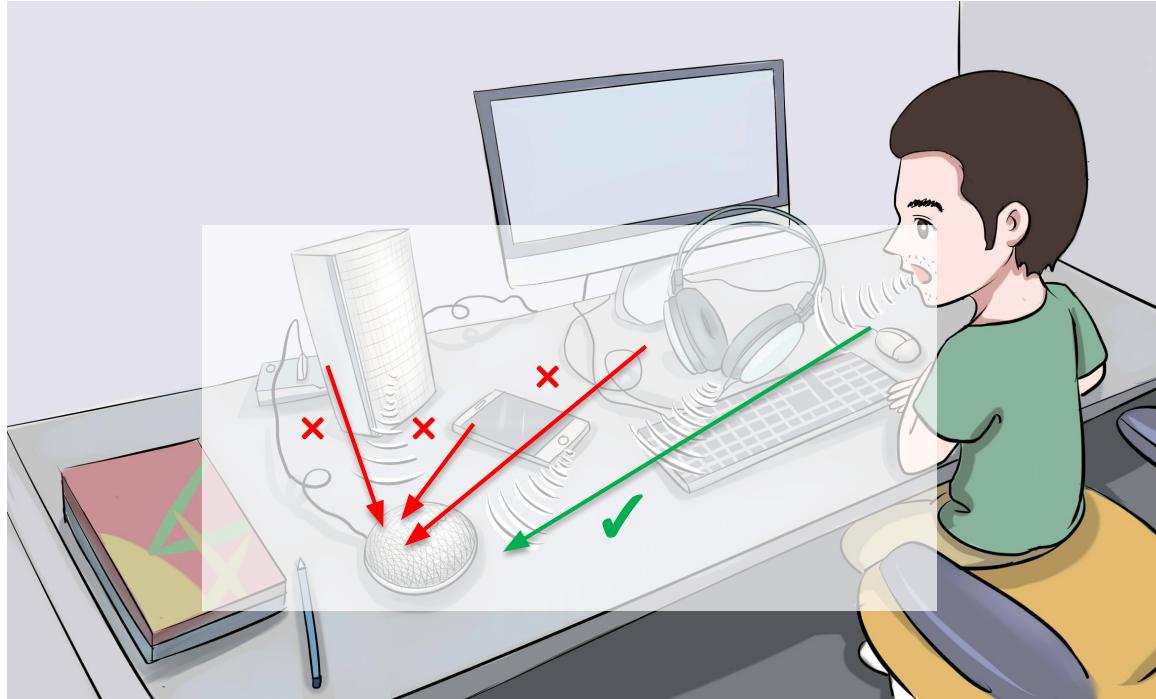
<https://apnews.com/article/new-hampshire-primary-bi...> ::

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 ...



Responsible audio AGI - security



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

Responsible audio AI - security



AP News

<https://apnews.com/article/new-hampshire-primary-bi...> ::

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

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

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。

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

前往苗栗市方向八斗子。

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。

Responsible audio AGI - fairness and trustworthiness

Fairness

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

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

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 noise-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.

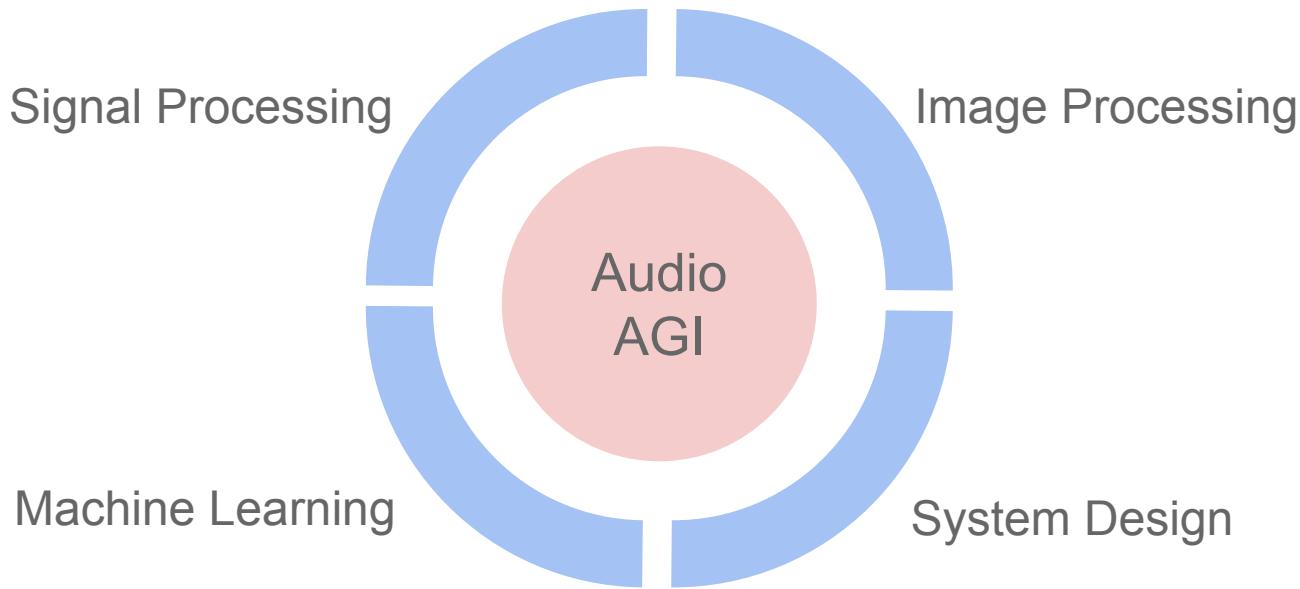
In this special session, we specialize in responsible aspects of speech foundation models, which are not adequately covered by regular sessions. We aim to facilitate knowledge sharing from diverse speech areas and pioneer discussions on both tech and non-tech issues. Furthermore, in line with the IS 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

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**

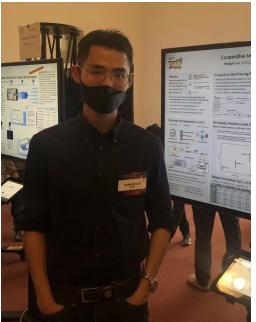
Audio AGI and other ECE areas



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



MIT-IBM
Watson
AI Lab

signify

Takeda