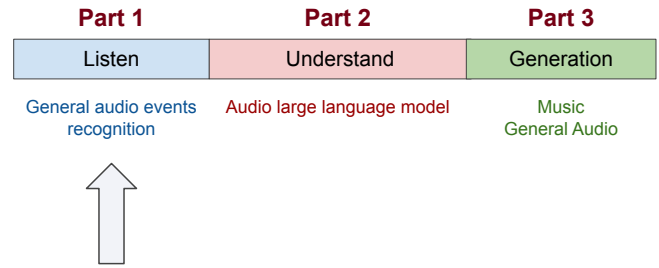


General Audio Processing Perception, Understanding, and Generation

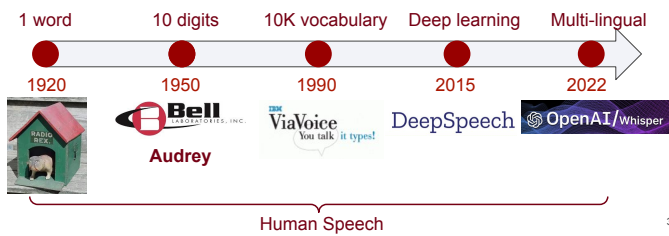
Yuan Gong / MIT 6.8620/HST.728 / Spring 2024

Outline



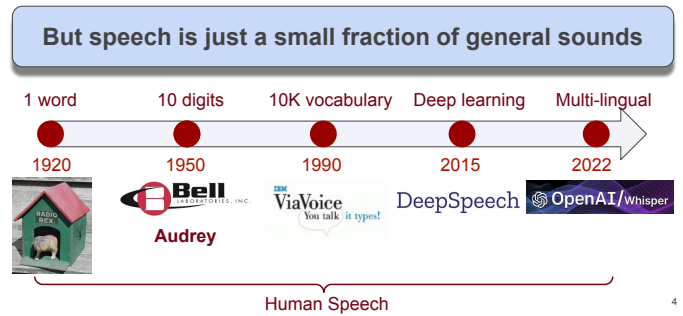
2

History of machine listening



3

History of machine listening



4

Sample



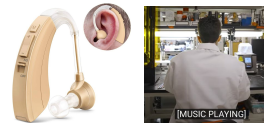
What did you hear?

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



Music Recognition



Embodied AI with Audio Capability



Sound-Based Surveillance



History of machine listening

First workshop on general audio recognition **DCASE**

First large general audio dataset **AudioSet**

What is the difference between speech and general audio signals?



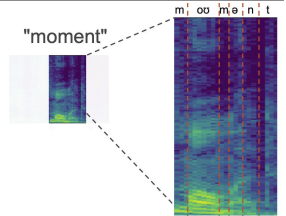
7

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]



8

There are unlimited number of general sounds

Speech

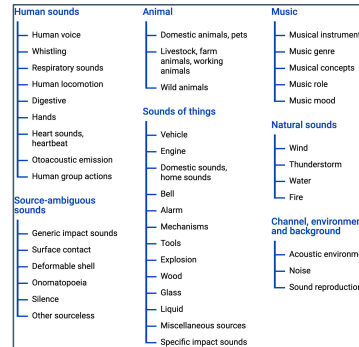
- composed of elementary units: phonemes
 - ~40 phonemes for English

General audio

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

9

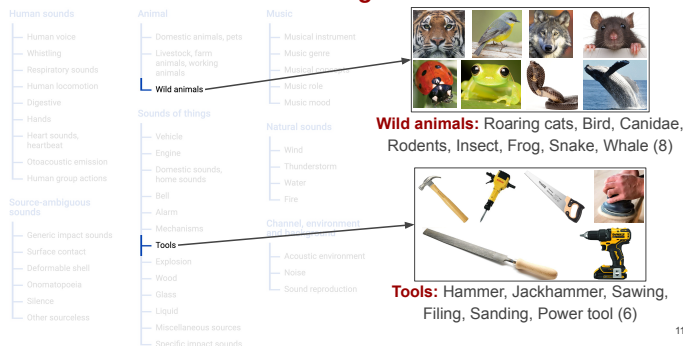
There are unlimited number of general sounds



AudioSet ontology (top-level)
527 classes in total

10

There are unlimited number of general sounds



11

AudioSet is already the largest dataset

Dataset	# Recordings	Avg Duration	Total Duration	# Sound Events	Label	Annotation
ESC-50	2,000	5s	2.8h	50	Single	Manual
UrbanSound	1,302	75s	27.0h	10	Single	Manual
DCASE 2017	52,762	10s	146.6h	17	Single	Manual
AudioSet (Strong Label)	120K	10s	~330h	456	Multiple	Automatic Suggestion and Manual Verification
AudioSet (Weak Label)	2.1 M	10s	5,800h ~ 8 months	527	Multiple	Manual Verification

12

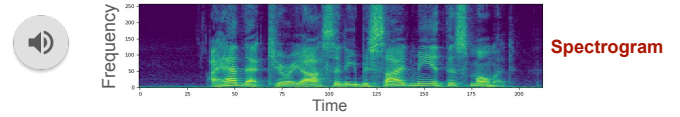
	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet

General audio classification has a larger search space

13

Speech has structured spectrogram patterns

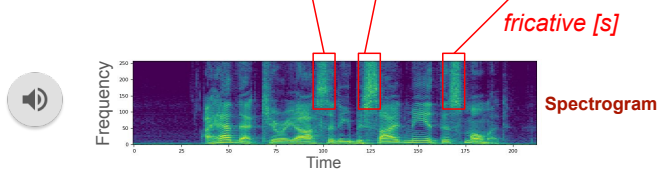
Transcript: I had that curiosity beside me at this moment



14

Speech has structured spectrogram patterns

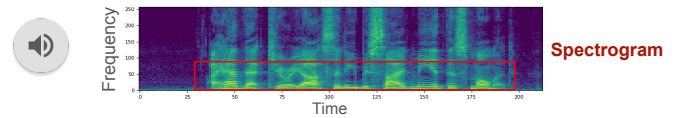
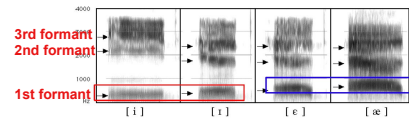
Transcript: I had that curiosi**s**ity beside me at thi**s** moment



Fricatives exhibit high-frequency turbulence

15

Speech has structured spectrogram patterns

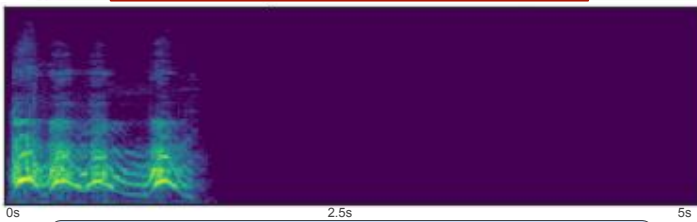


- Each vowel has relatively fixed formant patterns
- The duration of each phone is typically < 300ms

16

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

17

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

18

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 → Capture long-range context
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification → Attend to multiple sounds

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 → Attend to multiple sounds

20

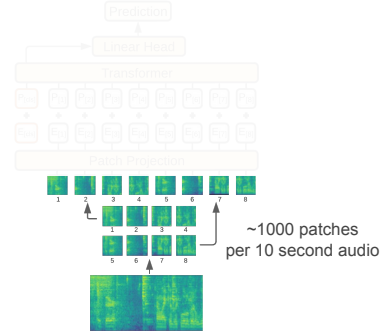
Audio spectrogram transformer (AST)



Yuan Gong et al., Audio Spectrogram Transformers, 2021

21

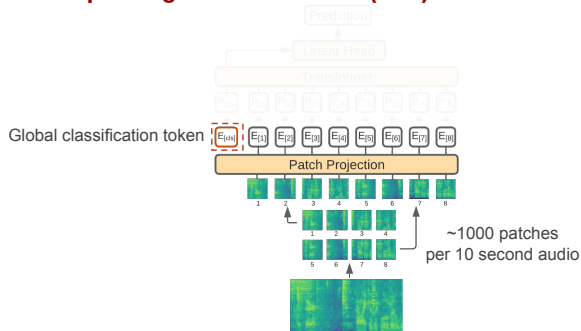
Audio spectrogram transformer (AST)



Yuan Gong et al., Audio Spectrogram Transformers, 2021

22

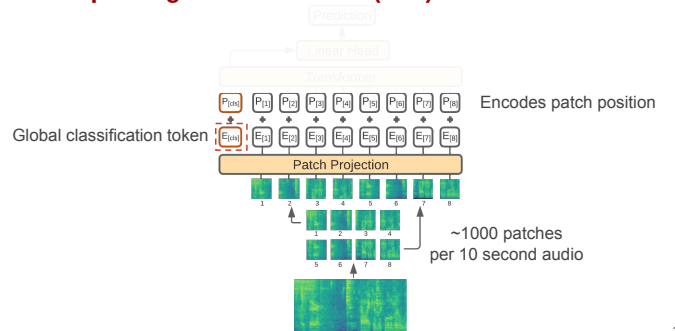
Audio spectrogram transformer (AST)



Yuan Gong et al., Audio Spectrogram Transformers, 2021

23

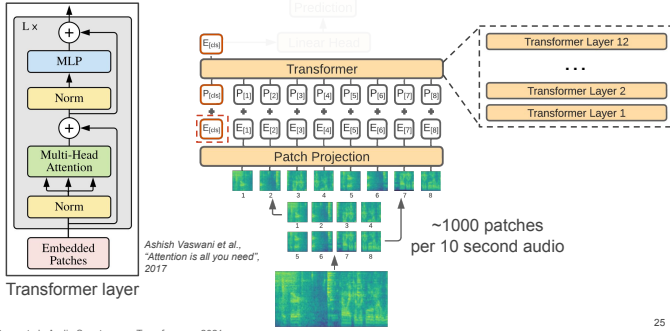
Audio spectrogram transformer (AST)



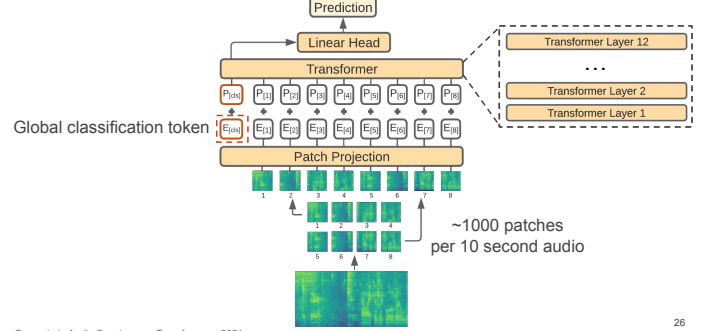
Yuan Gong et al., Audio Spectrogram Transformers, 2021

24

Audio spectrogram transformer (AST)



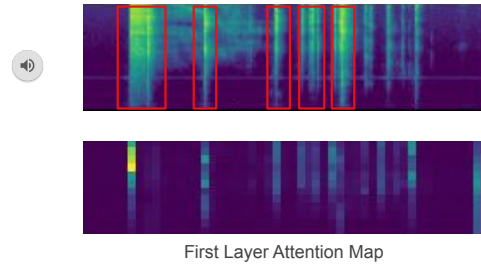
Audio spectrogram transformer (AST)



	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

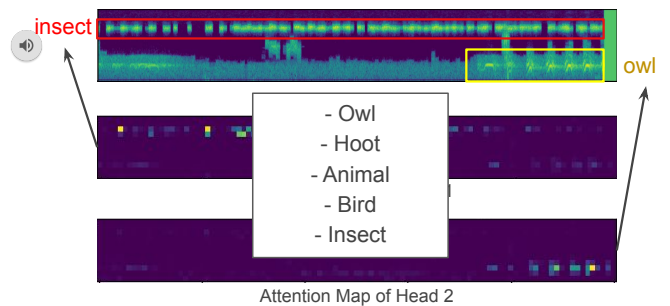
Advantage of AST: long-range dependencies



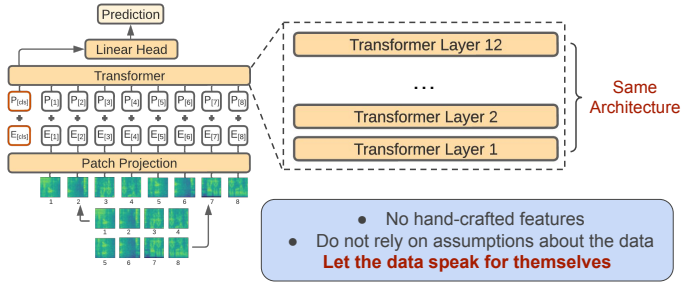
	Speech	General Audio
Elementary units	Phonemes	N/A
# Classes	~40 for English	Unlimited 527 for AudioSet
Duration	Typically < 300ms	Up to a few seconds
Pattern	Structured	Diverse & Complex
Task	Single-label Classification	Multi-label Classification

Attend to multiple sounds

Advantage of AST: dynamic attention

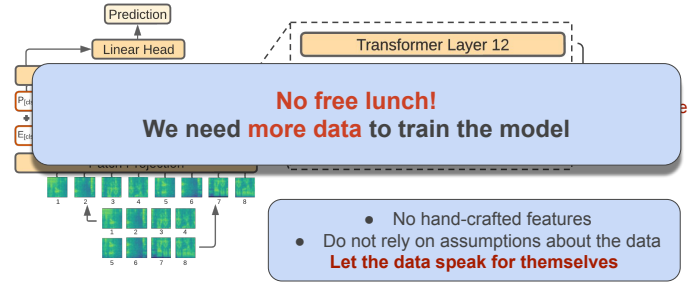


AST is a data-driven model



31

AST is a data-driven model



32

The data issue

Conventional supervised learning



- labeled data (audio, sound class) are expensive!
- Can we learn from unlabeled data?

33

Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data



Random masking, no human annotation required

34

Audio self-supervised learning

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

35

Audio self-supervised learning



The model learns the **spectrogram pattern**
 The model does **not** associate the pattern with sound class
 This sound features repetitive, high-frequency tones



36

Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data



At this point, the model cannot do classification

37

Audio self-supervised learning

Stage 1: Self-supervised pretraining

Large-scale unlabeled data



Weight Inheritance

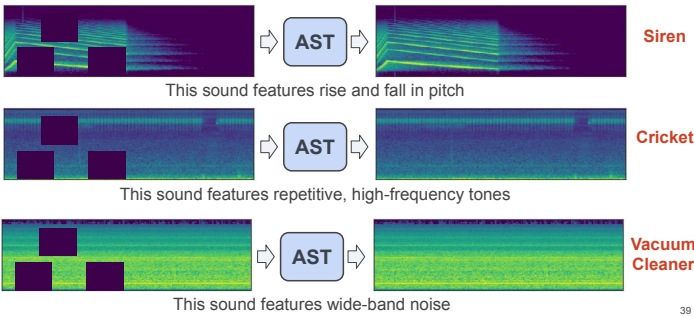
Stage 2: Supervised finetuning

Small-scale labeled data



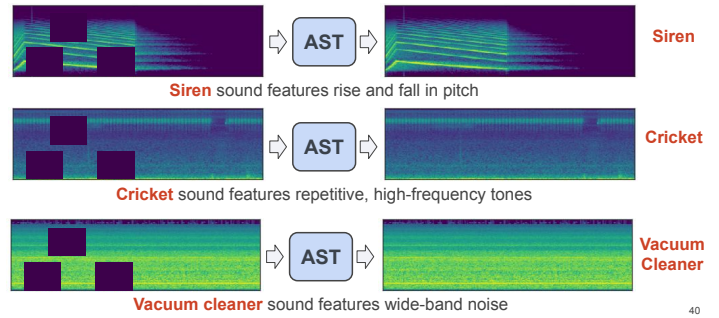
38

Audio self-supervised learning



39

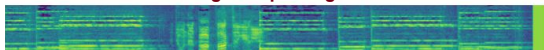
Audio self-supervised learning



40

Surprisingly strong reconstruction ability

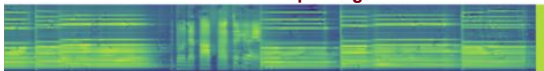
Original Spectrogram



Masked Spectrogram (75% Masked)



Reconstructed Spectrogram

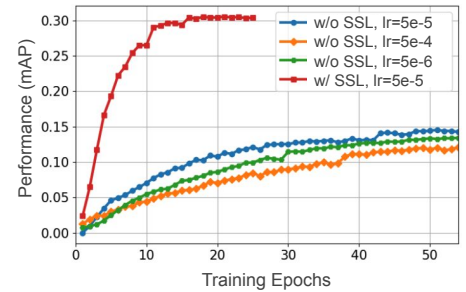


Reconstruction ability serves merely as a sanity check
Ultimately, we care about the performance in sound classification

41

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, 2022

42

Unified speech and non-speech recognition



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

43

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

44

MIT / ast-finetuned-audioset-10-10-0.4593 🔗 ❤️ like

Audio Classification 🔗 Transformers 🔗 PyTorch 🔗 Safetensors

audio-spectrogram-transformer 🔗 Inference Endpoints 🔗 arxiv:2104.01778 🔗 📄 📄 📄

Train 🔽 Deploy 🔽 Use in Transformers

Model card 🔗 Files 🔗 Community 10

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

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

45

Impact of AST

Wildlife Conservation

Bioelectrical Signal Classification

Multi-modal Systems

Meta

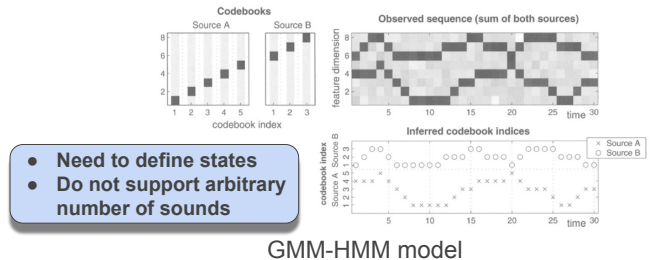
46

Limitations of AST

- Cannot predict onset and offset timestamps
- Not "true" self-supervised learning
 - Data is the unlabeled version of labeled data
- Computationally expensive
 - $O(n^2)$ w.r.t the audio length

47

Old methods for general audio recognition



Daniel Ellis, Model-based scene analysis, 2006

48

Old methods for general audio recognition

VGGish [1]	CNN6	CNN10	CNN14
Log-mel spectrogram 90 frames × 64 mel bins		Log-mel spectrogram 1000 frames × 64 mel bins	
3 × 3 @ 64 ReLU	5 × 5 @ 64 BN, ReLU	(2 × 3 @ 64) BN, ReLU × 2	(3 × 3 @ 64) BN, ReLU × 2
MP 2 × 2		Pooling 2 × 2	
3 × 3 @ 128 ReLU	5 × 5 @ 128 BN, ReLU	(3 × 3 @ 128) BN, ReLU × 2	(3 × 3 @ 128) BN, ReLU × 2
MP 2 × 2		Pooling 2 × 2	
(3 × 3 @ 256) ReLU × 2	5 × 5 @ 256 BN, ReLU	(3 × 3 @ 256) BN, ReLU × 2	(3 × 3 @ 256) BN, ReLU × 2
MP 2 × 2		Pooling 2 × 2	
(3 × 3 @ 512) ReLU × 2	5 × 5 @ 512 BN, ReLU	(3 × 3 @ 512) BN, ReLU × 2	(3 × 3 @ 512) BN, ReLU × 2
MP 2 × 2		Pooling 2 × 2	
Flatten	Global pooling	Global pooling	Global pooling
FC 4096 ReLU × 2	FC 512, ReLU	(3 × 3 @ 1024) BN, ReLU × 2	(3 × 3 @ 1024) BN, ReLU × 2
FC 527, Sigmoid	FC 527, Sigmoid	Pooling 2 × 2	Pooling 2 × 2
		(3 × 3 @ 2048) BN, ReLU × 2	(3 × 3 @ 2048) BN, ReLU × 2
		Global pooling	Global pooling
		FC 2048, ReLU	FC 2048, ReLU
		FC 527, Sigmoid	FC 527, Sigmoid

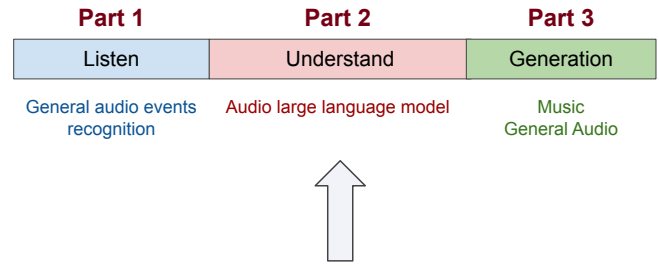
- More complex design

Convolutional neural networks

Qiuqiang Kong et al., PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition, 2020

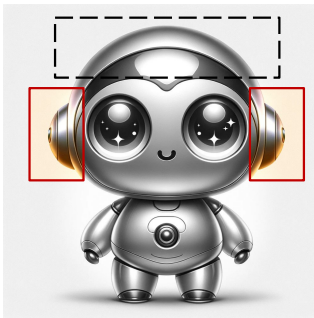
49

Outline



50

There is one thing missing...



No *thinking* and *understanding* ability!

We only modeled *listening* ability

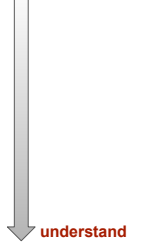
51

Human cognitive ability

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



perceive



understand

52

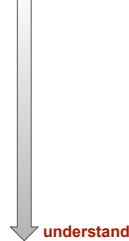
Human cognitive ability

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

[Think] Could be some burning or frying or arcing sound.



perceive



understand

53

Human cognitive ability

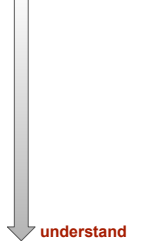
[Listen] High-frequency crackling noises, intermittent low-frequency pops.

[Think] Could be some burning or frying or arcing sound.

[Think] Combined with the speech, it is a cooking sizzling sound.



perceive



understand

54

Human cognitive ability

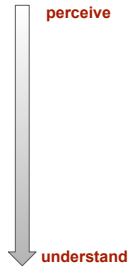


[Listen] High-frequency crackling noises, intermittent low-frequency pops.

[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.



55

Human cognitive ability



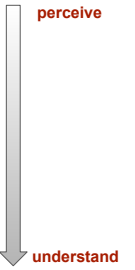
[Listen] High-frequency crackling noises, intermittent low-frequency pops.

[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?



56

Human cognitive ability

[Listen] High-frequency crackling noises, intermittent **perceive**

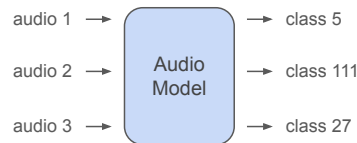
- We do more than just listen; we also comprehend sounds.
- Our comprehension integrates speech and non-speech sounds.

Can we make AI having a similar understanding ability?

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

57

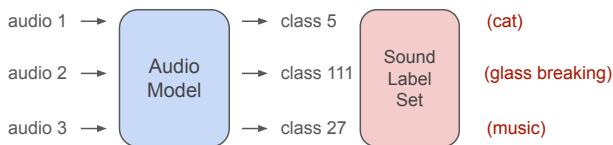
Conventional training won't work



"Listening" model only knows class indices, not sounds

58

Conventional training won't work

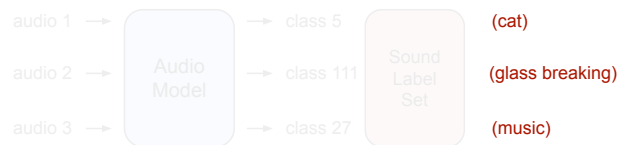


"Listening" model only knows class indices, not sounds

59

Conventional training won't work

"cat's meow is soft" **knowledge about sound**



"Listening" model only knows class indices, not sounds

what we need to model additionally

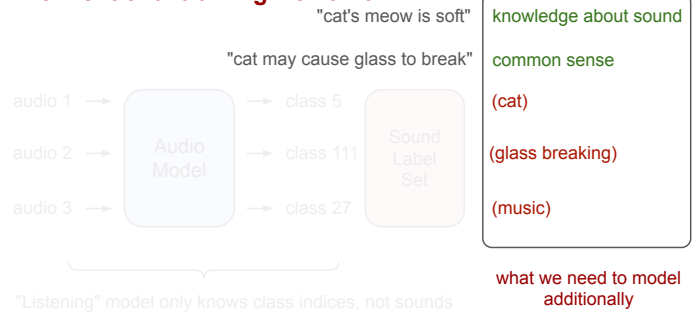
60

Conventional training won't work



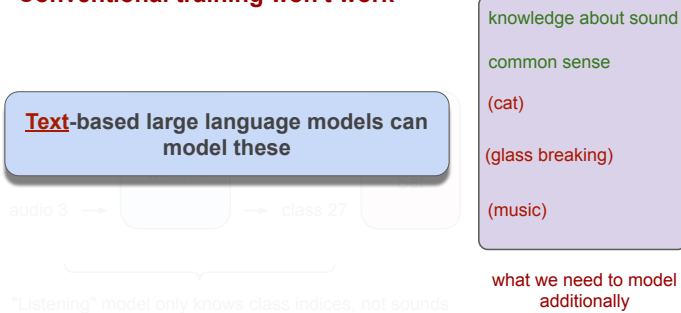
61

Conventional training won't work



62

Conventional training won't work



63

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:

- 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.
- Duration:** The duration of individual barks can vary, but a typical single bark lasts between 0.1 to 1 second.
- 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.

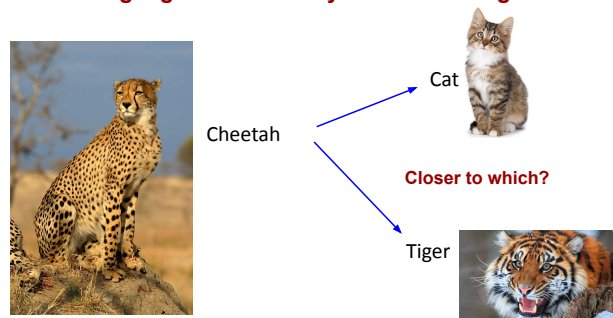
64

Pure language LLMs already have knowledge about sounds

- 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.
- 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.
- 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.
- 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.
- 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.

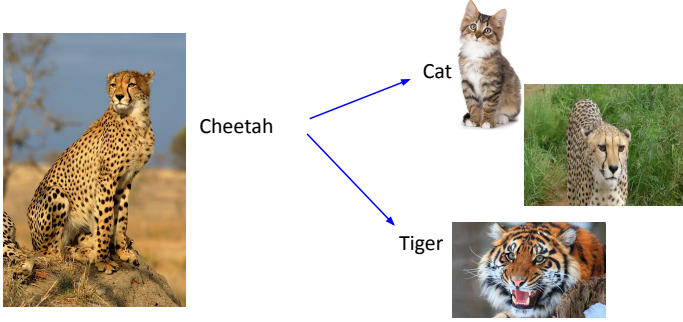
65

Pure language LLMs already have knowledge about sounds



66

Pure language LLMs already have knowledge about sounds



“Small” language model (LaBSE) gives wrong answer

Sentence Similarity

Source Sentence

Sound of cheetah

Sound of cat

Sound of tiger

Sound of lion

0.771

least similar



0.821

most similar



0.773

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?

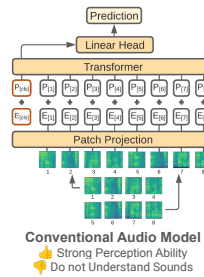


When comparing the sounds of cheetahs to those of domestic cats and tigers, the vocalizations of cheetahs are more akin to domestic cats. Here's a breakdown:

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

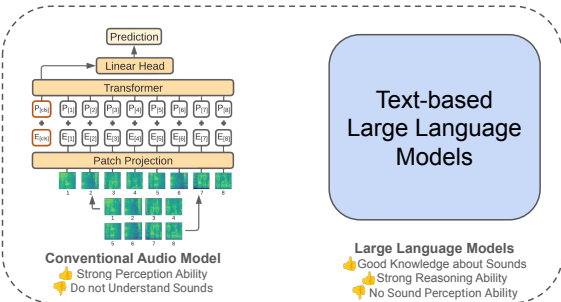


Conventional Audio Model
 🟡 Strong Perception Ability
 🟡 Do not Understand Sounds

Text-based Large Language Models

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

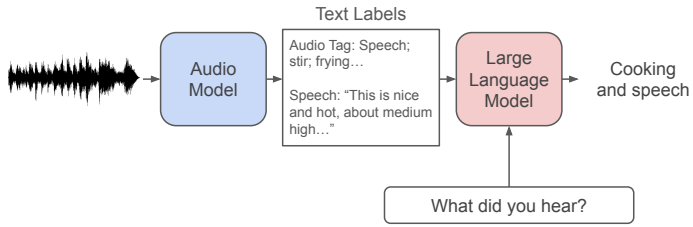
Text-based Large Language Models

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

Bridging sound perception and understanding

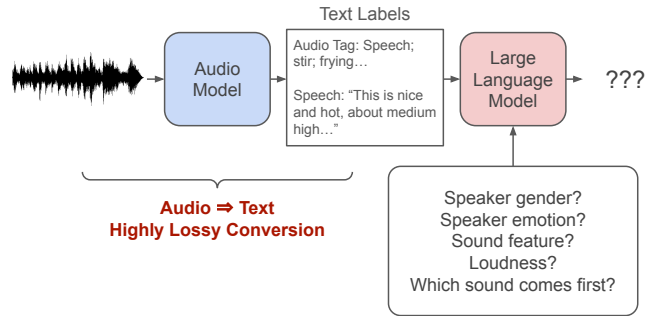
Listen, Think, and Understand (LTU)

Can we just concatenate an audio model and an LLM?



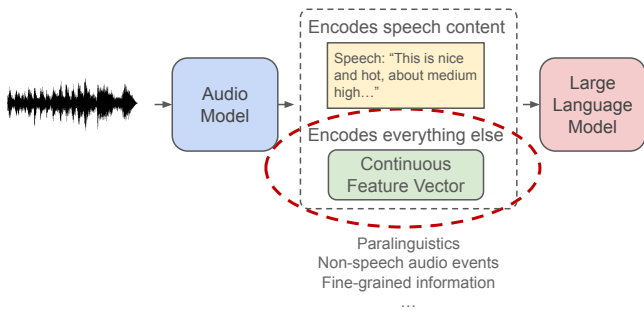
73

Can we just concatenate an audio model and an LLM?



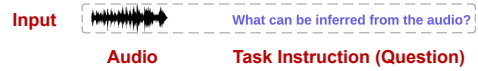
74

Continuous acoustic features are needed



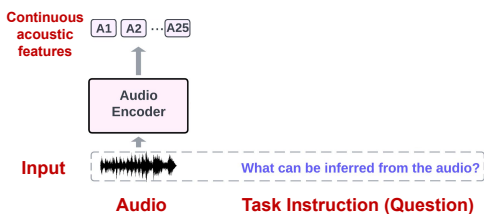
75

LTU model architecture



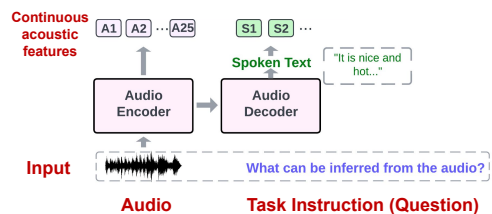
76

LTU model architecture



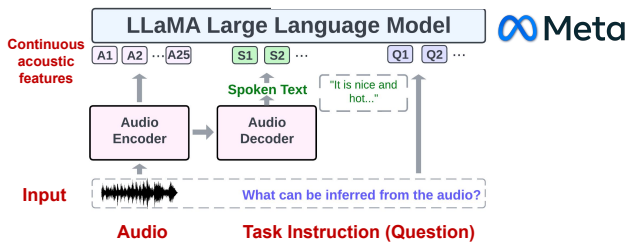
77

LTU model architecture



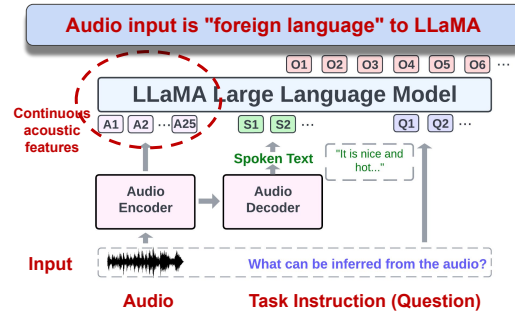
78

LTU model architecture



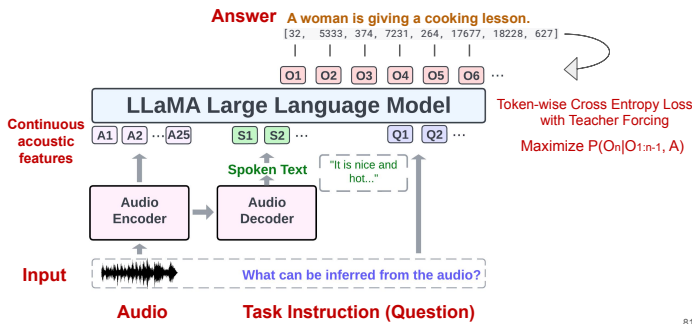
79

LTU model architecture



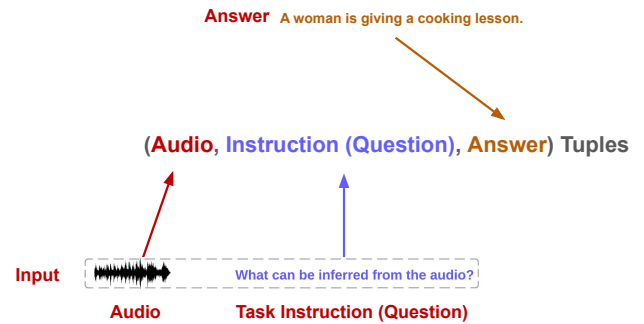
80

LTU training objective



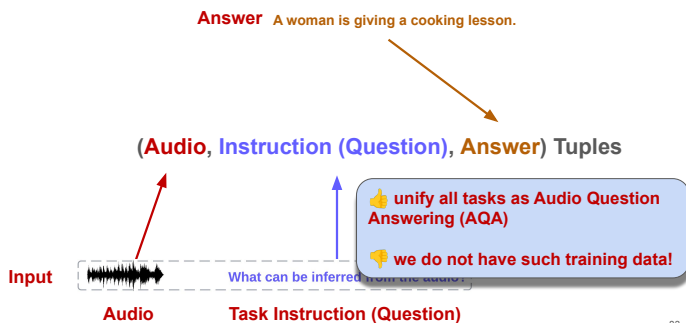
81

LTU training data



82

LTU training data



83

Closed-ended AQAs are easy to generate

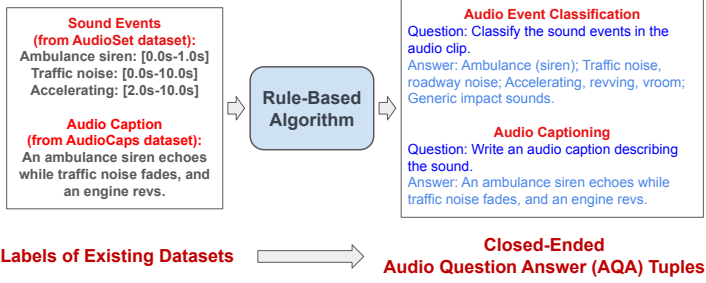
Sound Events
(from AudioSet dataset):
Ambulance siren: [0.0s-1.0s]
Traffic noise: [0.0s-10.0s]
Accelerating: [2.0s-10.0s]

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

Labels of Existing Datasets

84

Closed-ended AQAs are easy to generate



85

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!

86

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?

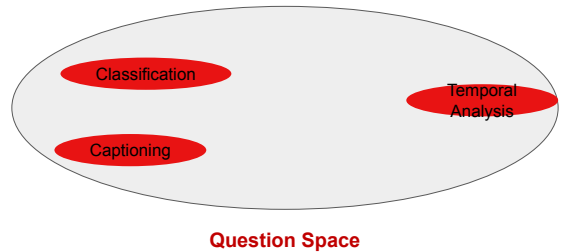
Answer: A woman speaks while frying food in a pan with background noise from a kitchen or dining room.



... The model cannot generalize!

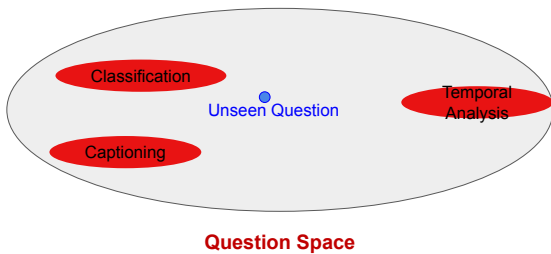
87

Closed-ended AQAs cannot train a generalized model



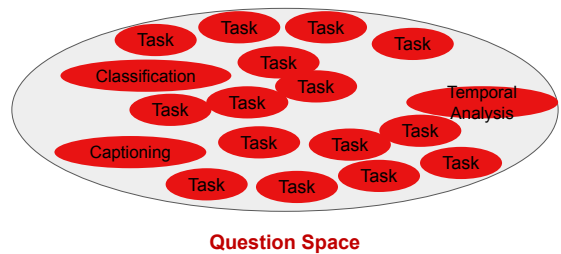
88

Closed-ended AQAs cannot train a generalized model



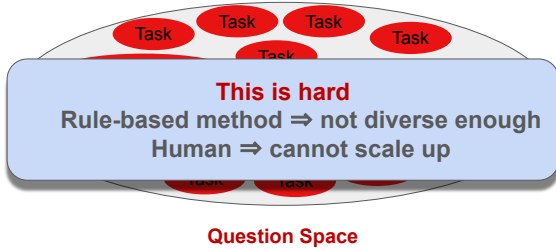
89

Ideally...



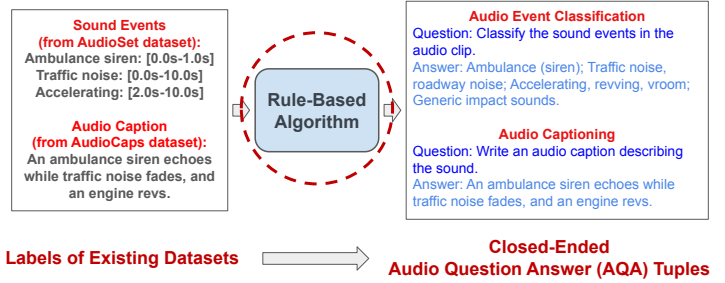
90

Ideally...



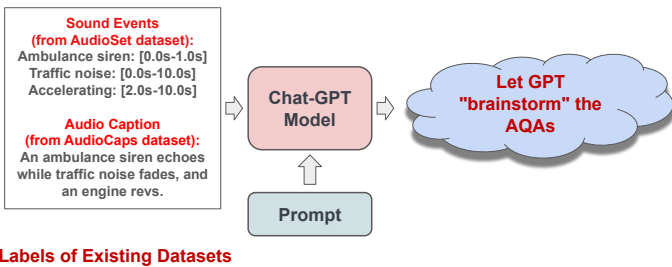
91

Rule-based algorithm cannot generate open-ended AQAs



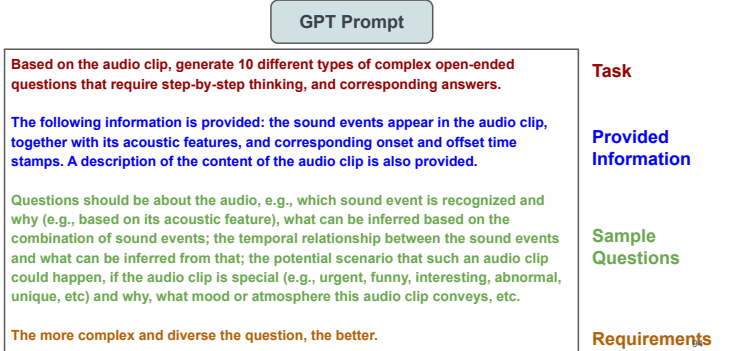
92

Audio instruction generation with GPT assistant

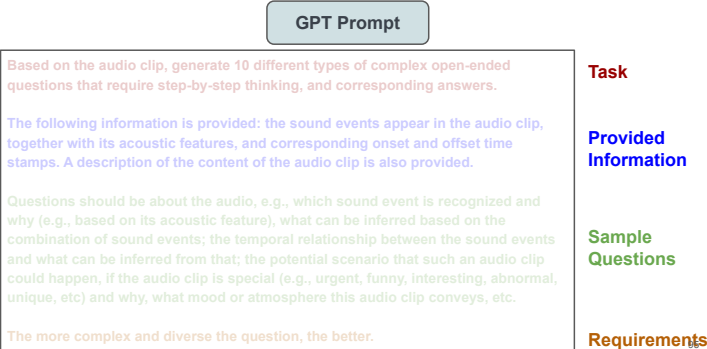


93

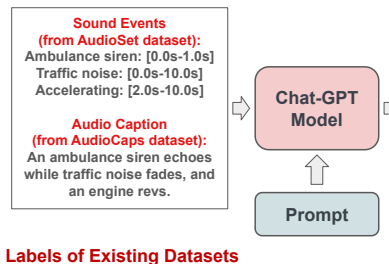
A closer look at the GPT prompt



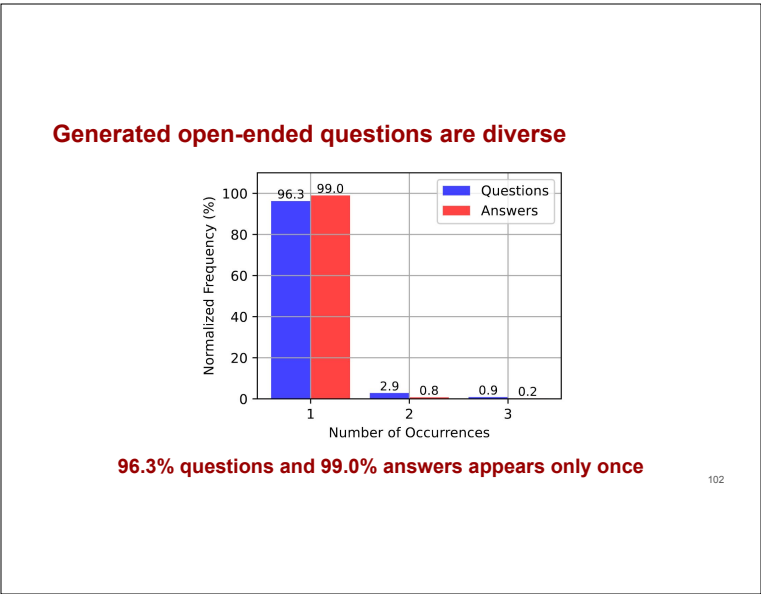
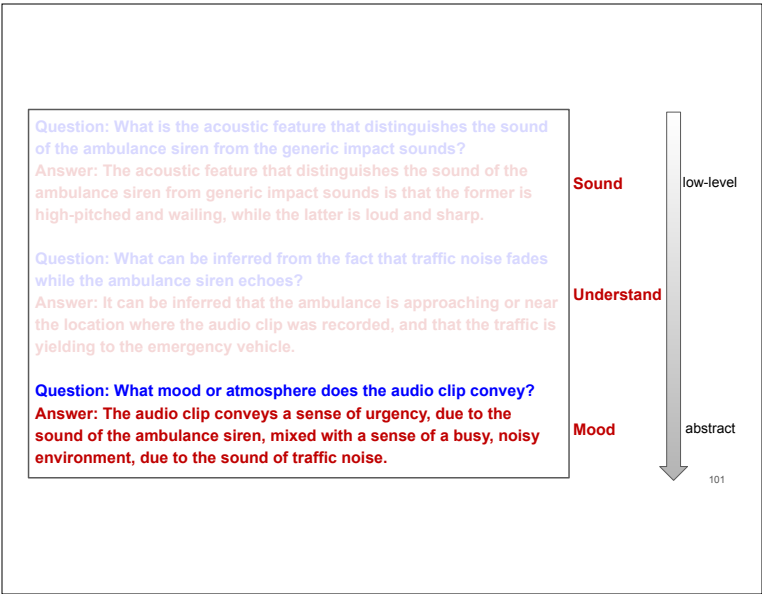
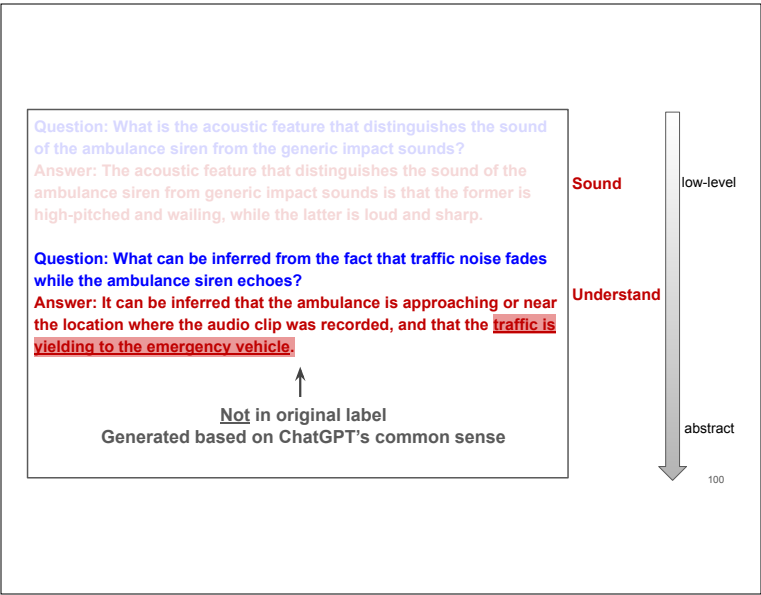
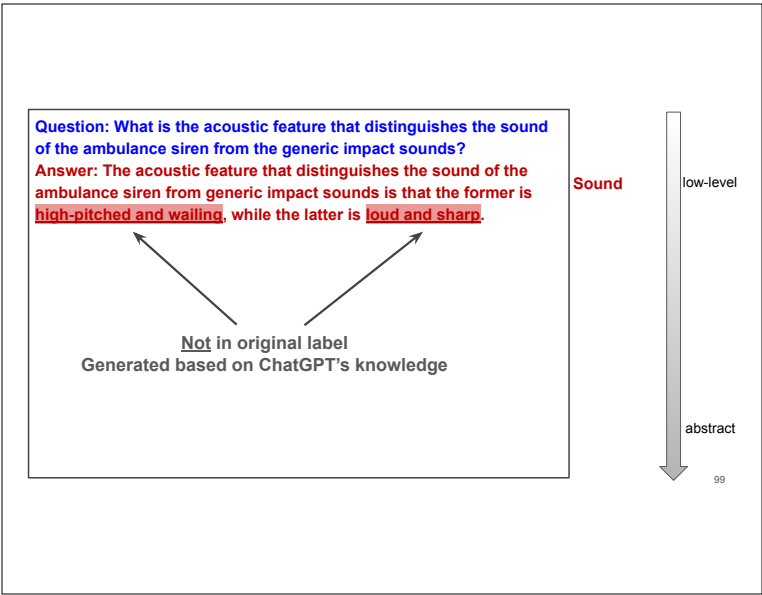
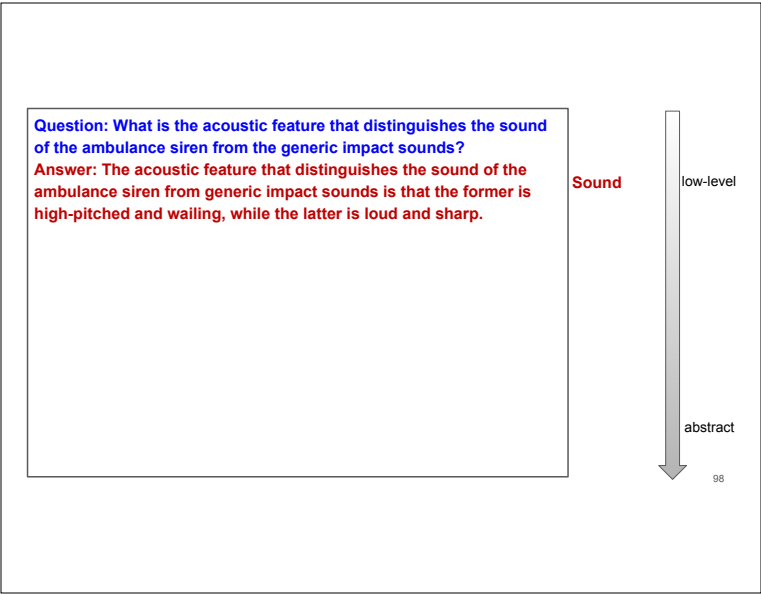
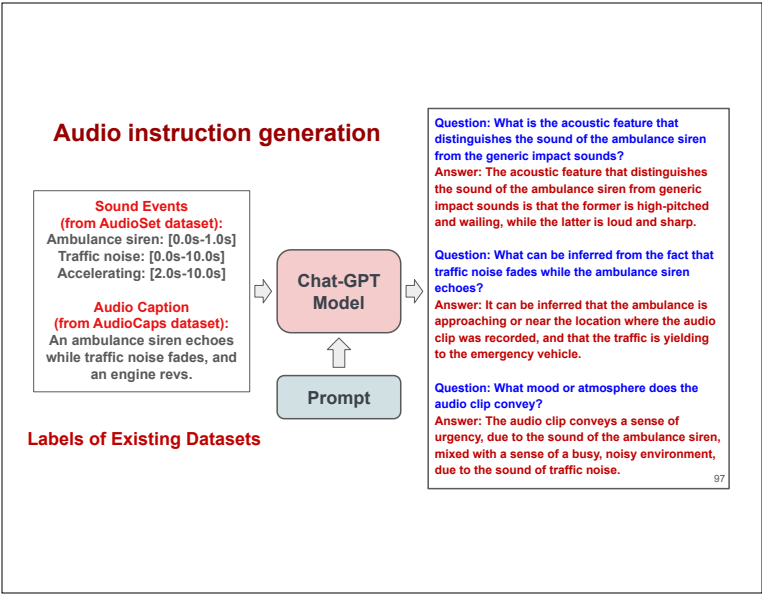
A closer look at the GPT prompt



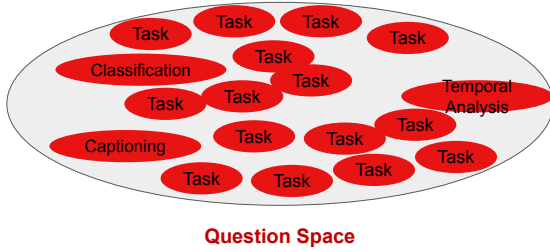
Audio instruction generation



96



Ideally...



103

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 (Opened) (4) (5)</i>											
AS-Strong (30)	x	x	x	x	-	-	-	-	102k	683k	901k
AudioSet (11)	x	-	x	x	-	-	-	x	500k	538k	184k
VGGSound	-	-	x	-	-	-	-	-	-	-	907k
FSD50k	-	-	x	-	-	-	-	-	-	-	403k
AudioCaps	-	-	x	-	-	-	-	-	-	-	478k
FreeSound	-	-	x	-	-	-	-	-	-	-	791k
Clothe (1)	-	-	x	-	-	-	-	-	-	-	89k
Sound B	-	-	x	-	-	-	-	-	-	-	10k
Sum	-	-	-	-	-	-	-	-	-	-	3,763k
<i>Speech</i>											
IEMOCAP	-	-	x	-	-	-	-	-	-	-	83k
LibriTTS	-	-	x	-	-	-	-	-	-	-	418k
VoxCele	-	-	x	-	-	-	-	-	-	-	926k
MUSDB18	-	-	x	-	-	-	-	-	-	-	358k
Sum	-	-	-	-	-	-	-	-	-	-	1,784k
<i>AudioSet (1)</i>											
AudioSet (1)	-	-	x	-	-	-	-	x	62k	93k	747k
FMA (29)	-	-	x	-	-	-	-	x	93k	93k	396k
Sum	-	-	-	-	-	-	-	-	175k	93k	1,143k
Total	-	-	-	-	-	-	-	-	1,089k	2,951k	6,690k
(9,641k Question Answer Pairs)											

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

104

Revisit the cooking demo



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

105

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.

106

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.

107

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.

108

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?

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?

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.

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?

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.

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?

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?

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.

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.

121

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.

122

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.

123

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.

124



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..."

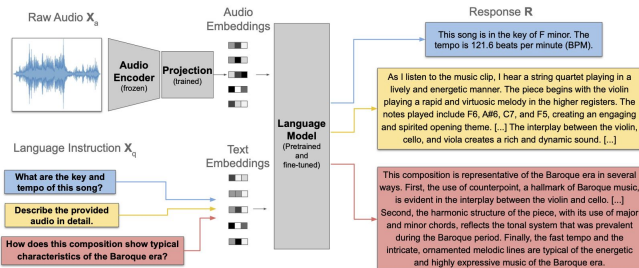
Question: Which line does this station transfer to?

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.

125



125

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。请在购票门口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 Karlinisky, 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.

Amazon Mechanical Turk

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

Audio: 1: A Audio ID: 10.030000000000000

Questions: What does the audio contain?
 Answer: The audio contains a train and possibly someone announcing, with the explosion and gunfire suggesting potential danger.

On 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 which part is wrong

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:
 Questions: What does the audio contain?
 Answer 1: The sound events in the audio clip convey a sense of tension and potential danger, with the explosion and gunfire suggesting potential danger.
 Answer 2: The audio clip conveys a sense of tension and potential danger due to the combination of explosion, train speech, and gunfire sounds, and further the danger.

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."

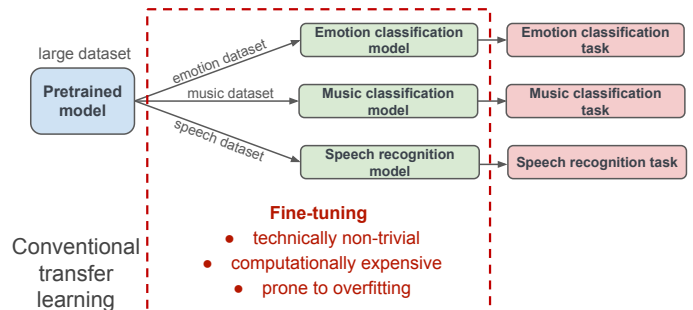
Human Answer: "It was train moving sound."

Advantage of LTU - 1/ understanding

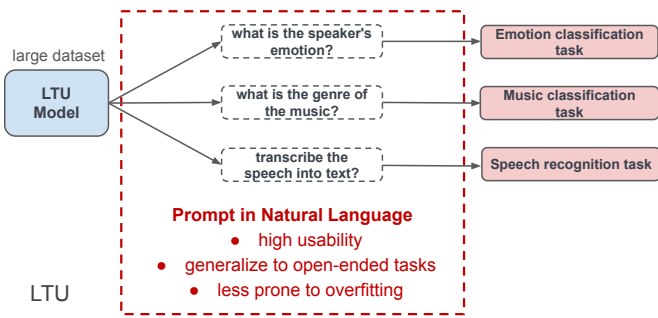
LTU jointly understands
spoken text
speech paralinguistics (e.g., emotion)
non-speech audio events

almost **everything** in audio

Advantage of LTU - 2/ broader capacity



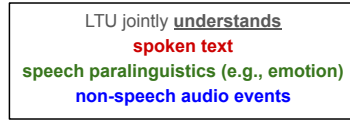
Advantage of LTU - 2/ broader capacity



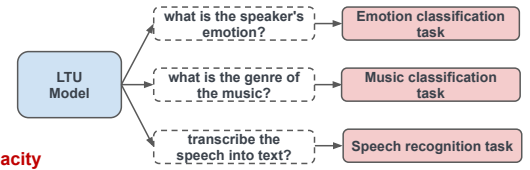
133

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

1/ Understanding

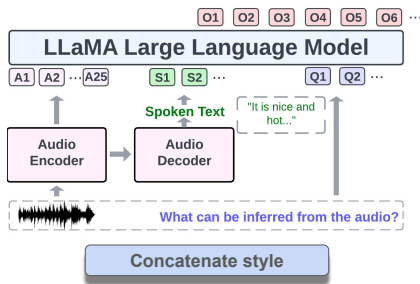


2/ Broader capacity



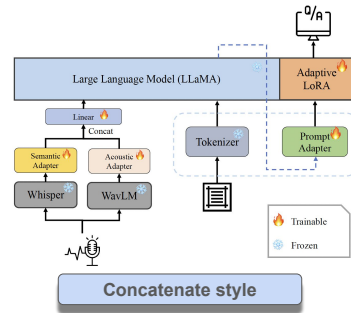
134

Other types of audio large language models



135

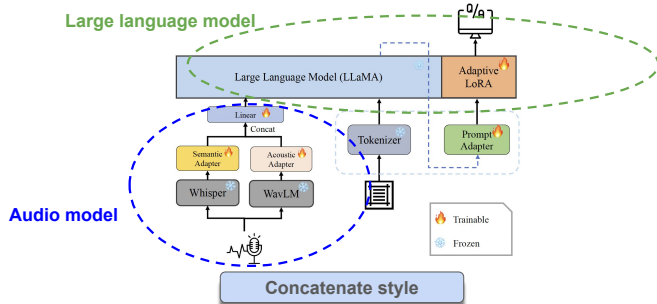
Other types of audio large language models



Shujie Hu et al., WavLM: Towards Robust and Adaptive Speech Large Language Model, 2024

136

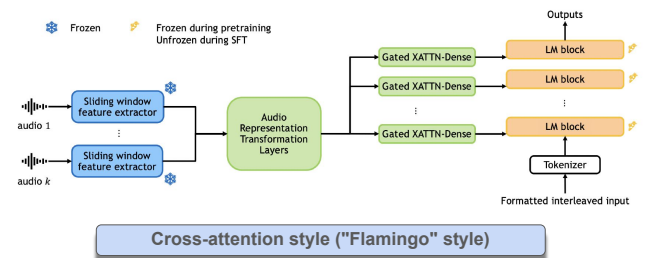
Other types of audio large language models



137

Shujie Hu et al., WavLM: Towards Robust and Adaptive Speech Large Language Model, 2024

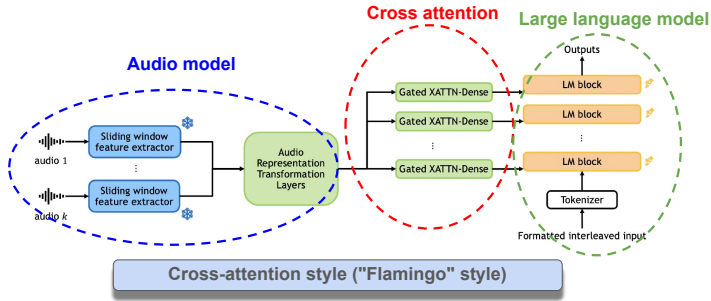
Other types of audio large language models



Zhifeng Kong et al., Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities, 2024

138

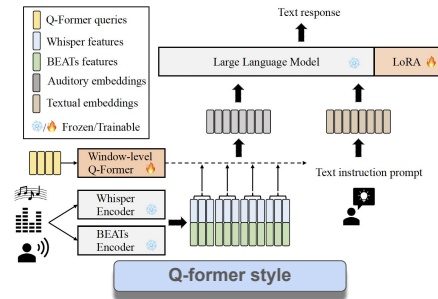
Other types of audio large language models



Zhifeng Kong et al., Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities, 2024

139

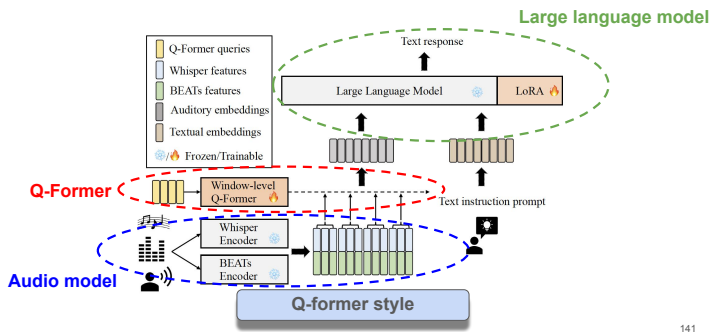
Other types of audio large language models



Changli Tang et al., SALMONN: Towards Generic Hearing Abilities For Large Language Models, 2023

140

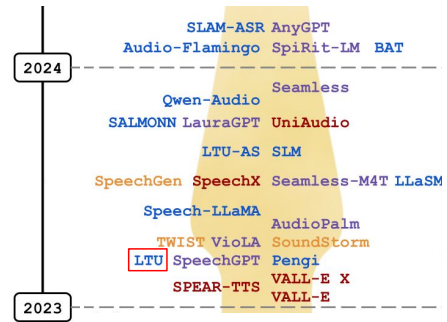
Other types of audio large language models



Changli Tang et al., SALMONN: Towards Generic Hearing Abilities For Large Language Models, 2023

141

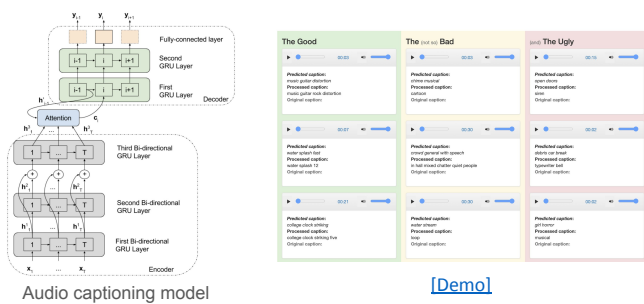
A list of audio large language models



<https://github.com/ga642381/speech-trident>

142

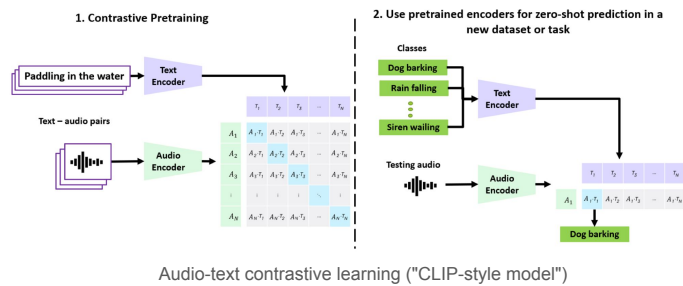
Other audio-text models



Konstantinos Drossos et al., Automated audio captioning with recurrent neural networks, 2017

143

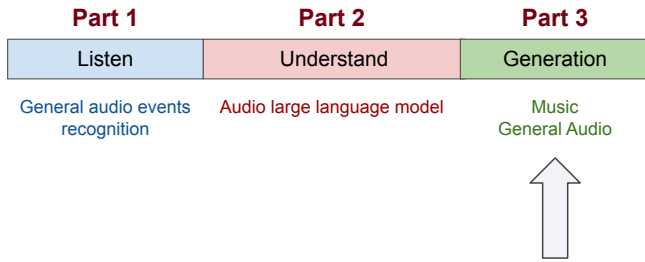
Other audio-text models



Benjamin Elizalde et al., CLAP: Learning Audio Concepts From Natural Language Supervision, 2022

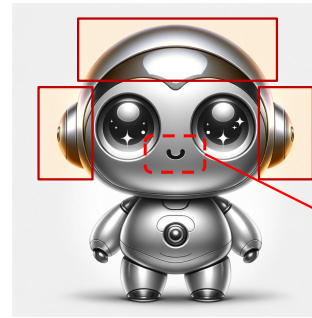
144

Outline



145

Audio generation



We modeled *listening and understanding*

We still left **sound generation** ability

146

Audio generation

Categorized by generation targets:

- Speech synthesis
 - [NaturalSpeech \(v1, v2, v3\)](#)
 - [Vall-E](#)
- Music generation
 - [Suno](#)
 - [MusicLM](#)
- Unified generation model
 - [AudioLDM \(v1, v2\)](#)
 - [Make-An-Audio](#)

147

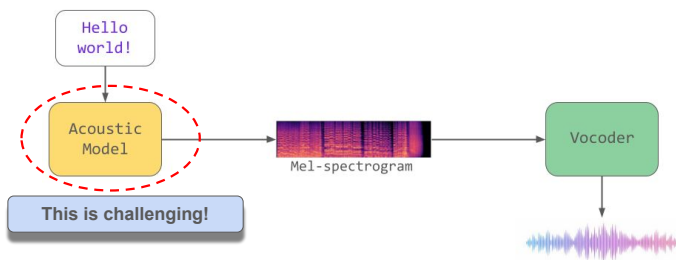
Audio generation

Categorized by generation methods:

- Diffusion-based models
- Audio codec-based models

148

Diffusion-based models

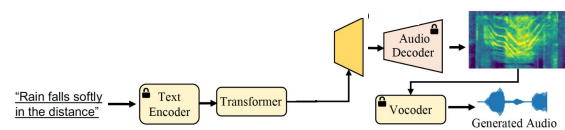


<https://www.assemblyai.com/blog/recent-developments-in-generative-ai-for-audio>

149

Diffusion-based models

Make-An-Audio Model

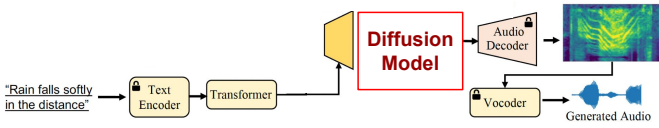


Rongjie Huang et al., Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models, 2024

150

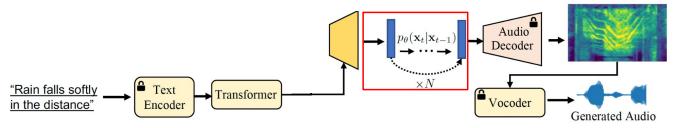
Diffusion-based models

Make-An-Audio Model

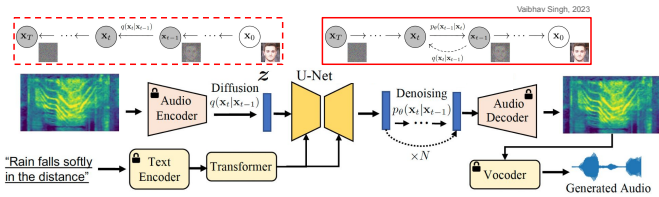


Diffusion-based models

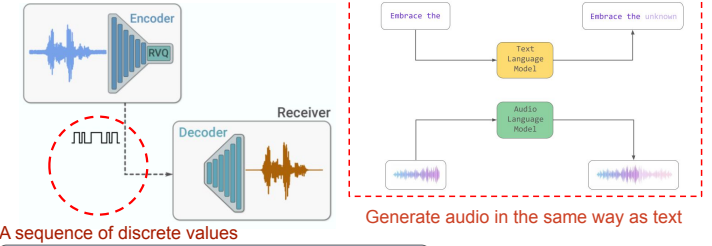
Make-An-Audio Model



Diffusion-based models



Audio codec based models

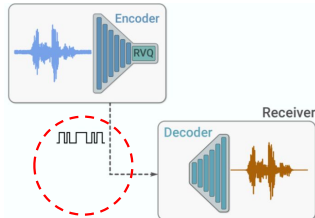


A sequence of discrete values

Generate audio in the same way as text

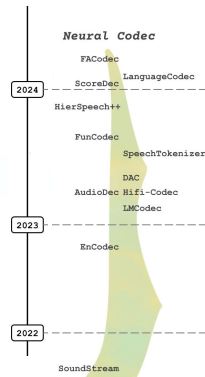
- Original designed for compression
- Later found ideal as generation units

Audio codec based models



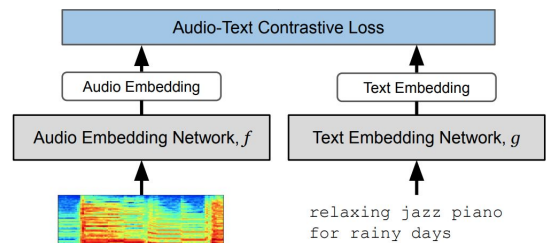
A sequence of discrete values

- Original designed for compression
- Later found ideal as generation units



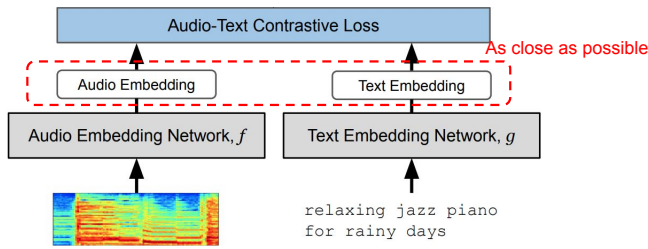
Audio codec based models

MuLan (Music-text embedding model)



Audio codec based models

MuLan (Music-text embedding model)

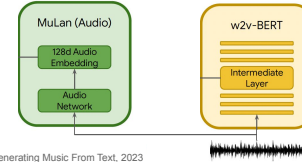


Audio codec based models

MusicLM

Step 1. Convert audio to representations

- An audio-text representation M^A
 - MuLan embedding, aligned with text
- An audio representation S to model coarse semantics (e.g., melody)
 - w2v-BERT embedding

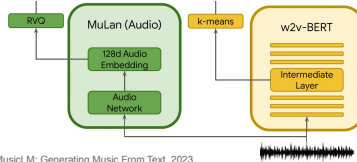


Audio codec based models

MusicLM

Step 2. Quantize representations to discrete values

- Residual vector quantization (RVQ) for MuLan embeddings M^A
- K-means for w2v-BERT embeddings S

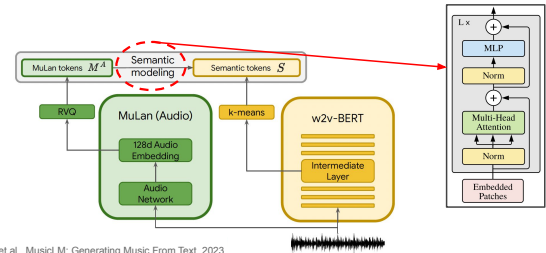


Audio codec based models

MusicLM

Step 3. Predict semantic tokens S from audio-text tokens M^A

- Modeled by a Transformer

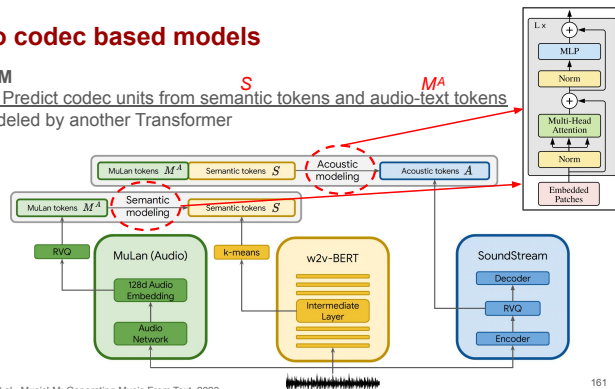


Audio codec based models

MusicLM

Step 4. Predict codec units from semantic tokens S and audio-text tokens M^A

- Modeled by another Transformer

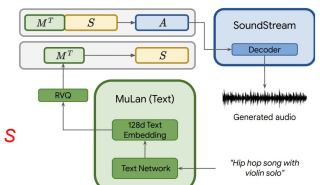


Audio codec based models

MusicLM

Inference

- Input text description
- Get MuLan embedding M^T ($\sim M^A$)
- Predict S conditioned on M^T
 - Use learned Transformer 1
- Predict codec A conditioned on M^T and S
 - Use learned Transformer 2
- Convert A to *waveform*
 - Use pretrained audio codec model SoundStream



Responsible audio AI

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
- Policies for responsible use of audio AI

163

Responsible audio AI

Fairness

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

Trustworthiness

- Hallucination
- Not good at tasks requiring high reliability

164

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。请在购票门口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。

165

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。

166

Responsible audio AI

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

167

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。请在购票门口前准备出票并进入车站。'

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

Responsible audio AI

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

Trustworthiness

- Hallucination
- Not good at tasks requiring high reliability

169

References

1. Mesaros et al., Sound event detection: A tutorial, 2021
2. Yuan Gong et al., Audio Spectrogram Transformers, 2021
3. Yuan Gong et al., Self-supervised Audio Spectrogram Transformers, 2022
4. Yuan Gong et al., Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong Audio Event Taggers, 2023.
5. Daniel Ellis, Model-based scene analysis, 2006
6. Qiuqiang Kong et al., PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition, 2020
7. Yuan Gong et al., Listen, Think, and Understand, 2024
8. Yuan Gong et al., Joint Audio and Speech Understanding, 2023
9. Gardner et al., Lark: A multimodal foundation model for music, 2023
10. Shujie Hu et al., WavLLM: Towards Robust and Adaptive Speech Large Language Model, 2024
11. Zhifeng Kong et al., Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities, 2024
12. Changli Tang et al., SALMONN: Towards Generic Hearing Abilities For Large Language Models, 2023
13. <https://github.com/ga642381/speech-trident>
14. Konstantinos Drossos et al., Automated audio captioning with recurrent neural networks, 2017
15. Benjamin Elizalde et al., CLAP: Learning Audio Concepts From Natural Language Supervision, 2022
16. <https://www.assemblyai.com/blog/recent-developments-in-generative-ai-for-audio>
17. Rongjie Huang et al., Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models, 2024
18. Neil Zeghidour et al., SoundStream: An End-to-End Neural Audio Codec 2021
19. <https://github.com/ga642381/speech-trident>
20. Qingqing Huang et al., MuLan: A Joint Embedding of Music Audio and Natural Language, 2022
21. Andrea Agostinelli et al., MusicLM: Generating Music From Text, 2023
22. Cao et al., [Super Musician: Music Style Transfer Based on Autoencoder](#), 2020

170