



Beyond Perception: Making Audio AI Understand Sounds

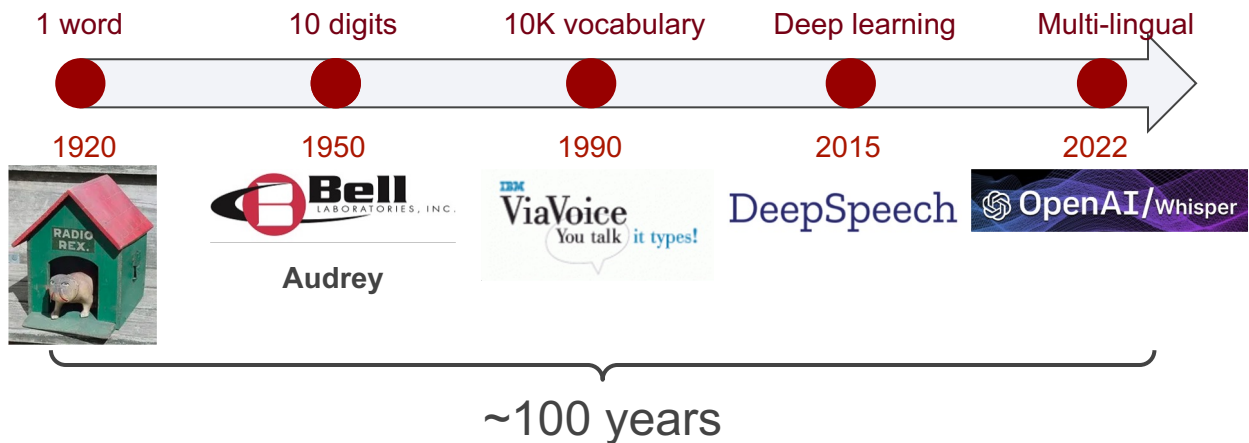
Yuan Gong

MIT Spoken Language Systems Group



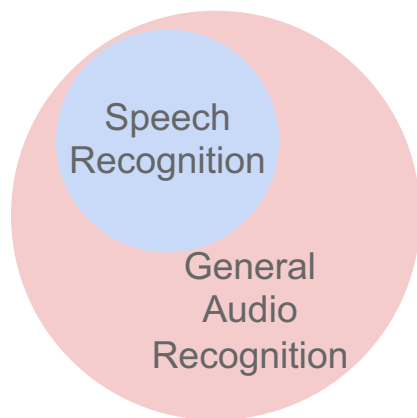
Computer Science &
Artificial Intelligence
Laboratory

The evolution of machine audio perception

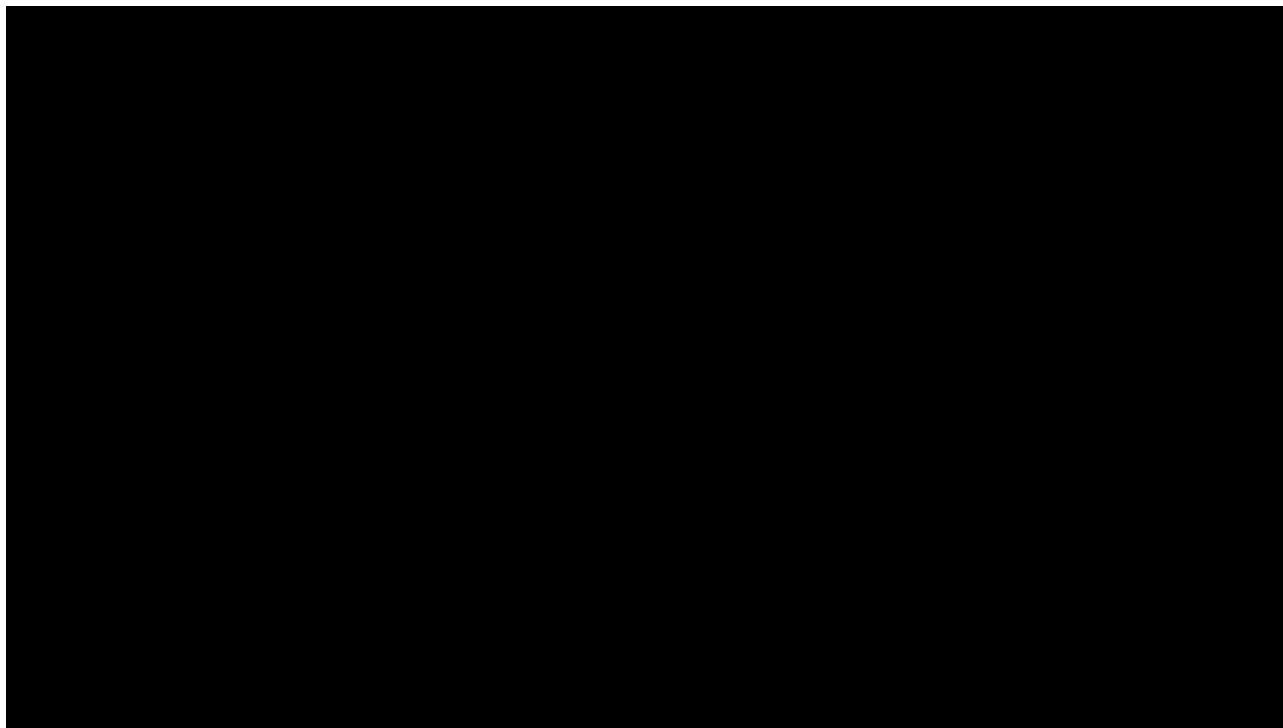
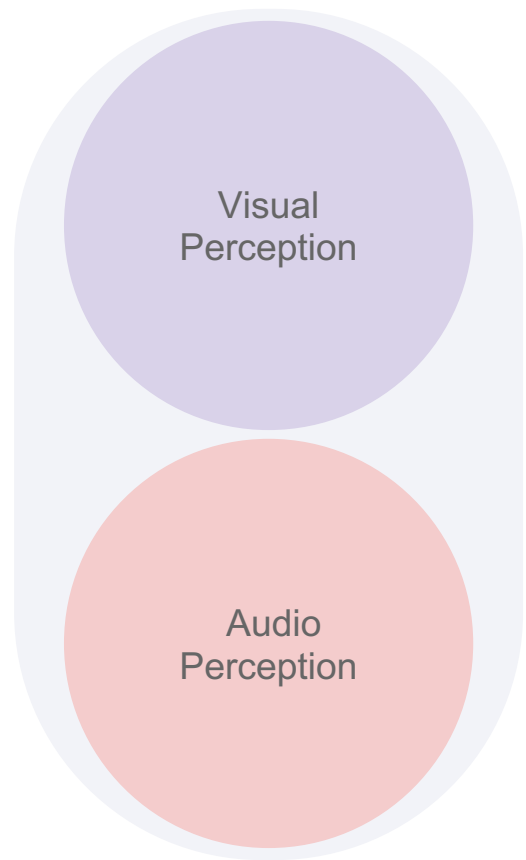


But speech is just a small fraction of general sounds

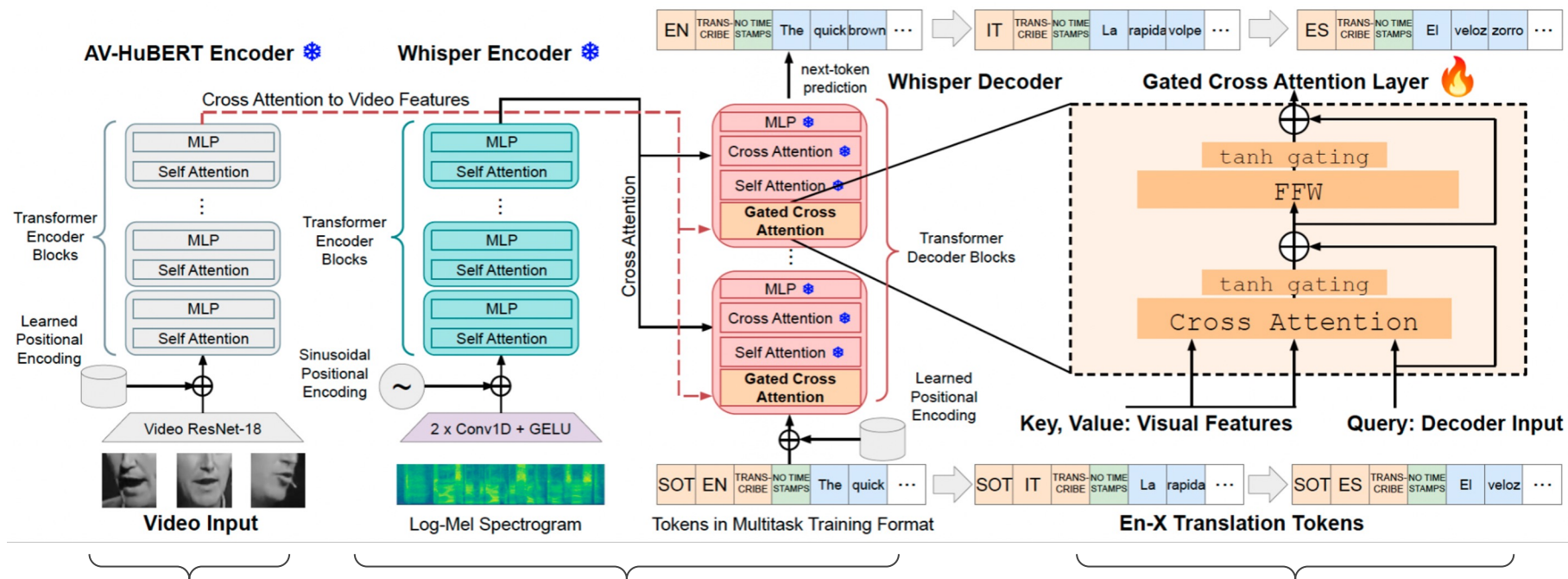
Unify speech and non-speech recognition



Connect audio and visual perception



Connect audio and visual perception



Video Model

Audio Model

Connector

Despite modern audio perception models coming close to, or even surpassing, human capabilities, they do not understand the (implicit) meaning of sounds

Human Cognitive Abilities



Sound of a cat \Rightarrow Cute animal, probably requesting food

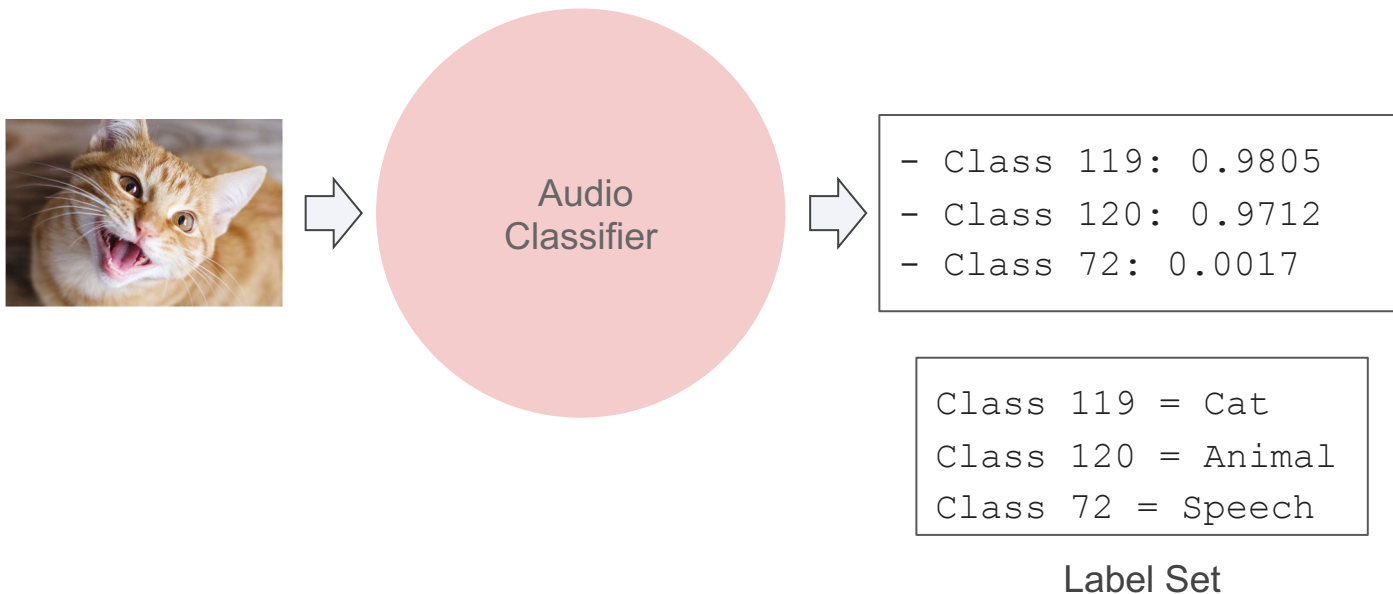


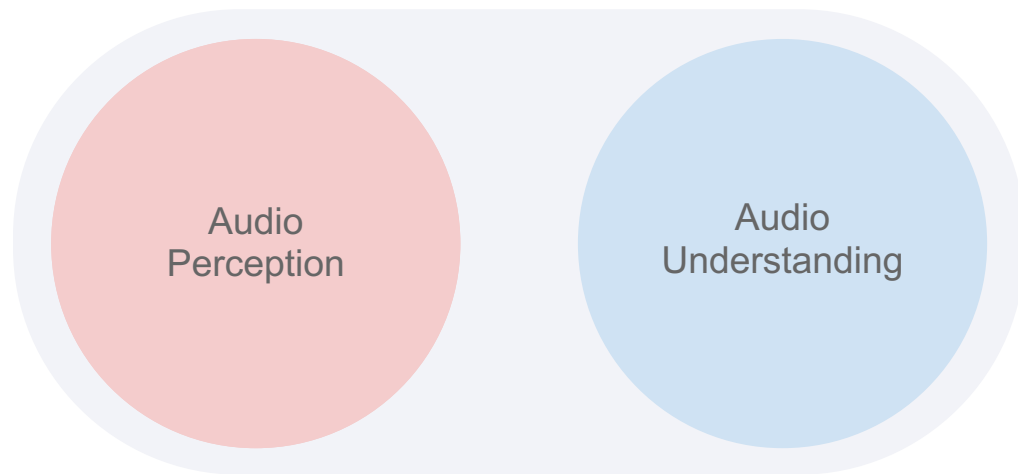
Sound of a tiger \Rightarrow Dangerous; run!

"What great weather today!" \Rightarrow Ideal day for travel with sunny skies

Despite modern audio perception models coming close to, or even surpassing, human capabilities, they do not understand the (implicit) meaning of sounds

Automatic Audio Perception Models





Text-based 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.

Text-based 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.

Bridging sound perception and understanding

Audio
Perception
Models



- 👍 Strong Perception Ability
- 👎 Do not Understand Sounds

Text-based
Large Language
Models



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

Bridging sound perception and understanding

Listen, Think, and Understand
(LTU)

Bridging sound perception and understanding

Answer A woman is giving a cooking lesson.

[32, 5333, 374, 7231, 264, 17677, 18228, 627]

O1 O2 O3 O4 O5 O6 ...

LLaMA Large Language Model

Token-wise Cross Entropy Loss with Teacher Forcing

Maximize $P(O_n | O_{1:n-1}, A)$

S1 S2 ... Q1 Q2 ...

Spoken Text

"It is nice and hot..."

Audio Encoder

Audio Decoder

Input



Audio

What can be inferred

Task Instruction

LTU can simultaneously recognize and jointly understand **spoken text**, **speech paralinguistics** (e.g., emotion), and **non-speech audio events**.

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											907k
FSD50K											403k
AudioCaps											478k
FreeSound											791k
Clotho [3]											89k
Sound Bites											10k
Sum											3,763k
<i>Speech Datasets</i>											
IEMOCAP											83k
LibriTTS											418k
VoxCeleb											926k
MOSEI [10]											355k
Sum											1,784k
<i>Joint Audio-Speech Datasets</i>											
AudioSet [31]	x	-	x	x	-	-	-	x	82k	-	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

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.



Audio Tags: Speech; Dishes, pots, and pans; Stir; Frying
Speech: This is nice and hot, about medium high.

Demo of non-speech audio 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.



Bridging sound perception and understanding



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

Takeaways and future works

- Improve machine understanding capabilities
 - Can machine understand sarcasm?
 - Reliability
 - Over/under interpretation
 - How to measure understanding capability?
- How to apply these models?
 - Personal assistant
 - Customer service
 - Embodied AI systems
- Integrate with audio generation models
- Audio-visual understanding



Yuan



Sameer



Andrew



Alex



Hongyin

Leonid
(MIT-IBM)

Jim