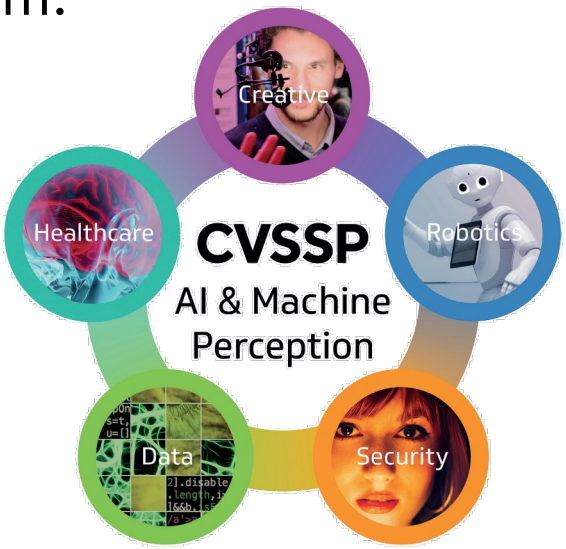
Introduction to Audio Artificial Intelligence (AI)

Guest Lecture of EEEM068 – Applied Machine Learning

Haohe Liu
https://haoheliu.github.io/
Final Year PhD Student
Centre for Vision, Speech and Signal Processing (CVSSP)

About CVSSP

- CVSSP: Creating machines that can see and **hear** to understand the world around them.



CVSSP
Centre for Vision, Speech and Signal Processing

CVSSP

One of the largest audio and vision research groups in the UK, started in 1986.

150+ researchers

About me

Haohe Liu
Final year PhD Student
University of Surrey
CVSSP

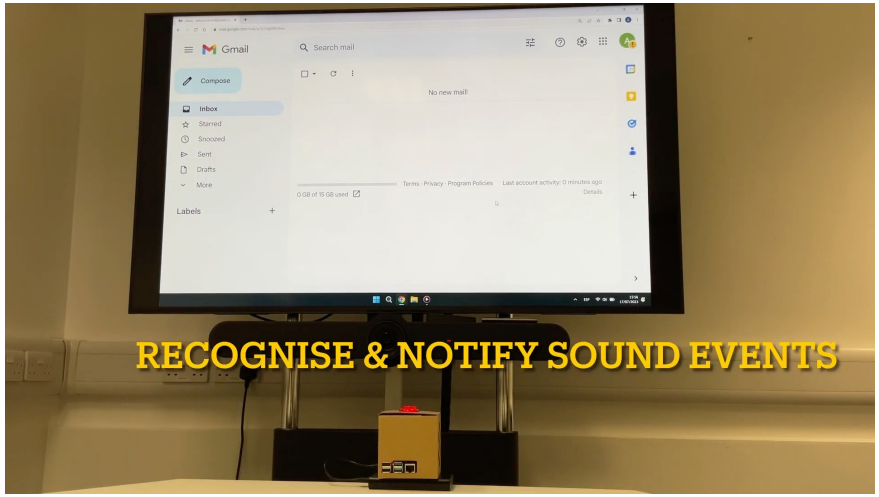
Supervisors:
Prof. Mark D. Plumbley
Prof. Wenwu Wang

AI for sound

CVSSP
Centre for Vision, Speech and Signal Processing

- Build audio technology that **inspires creativity** and **enhances communications**
- 🔬 Research
  - Audio and Music Generation; Text-to-Speech; Audio Recognition; Audio Quality Enhancement; Audio Source Separation, etc.
- 📊 Stats
  - 7000+ GitHub stars; 900+ citations; 100,000+ checkpoint downloads
  - ICML, AAAI, NeurIPS, TPAMI, TASLP, ICASSP, INTERSPEECH, etc.

Can AI understand sound?



Credit: Bibbo, G., Singh, A., & Plumbley, M. D. (2023). Audio Tagging on an Embedded Hardware Platform. arXiv preprint arXiv:2306.09106.

Can AI create sound?

Credit: Elevenlabs.io

Can AI create sound?

What you going to learn from this lecture

- Understand the background of audio technology.
- Intuitively understand how AI percept audio.
- Intuitively understand how auto-regressive models are used for audio generation.
- Gain general intuition of the latest research of audio AI.

Overview

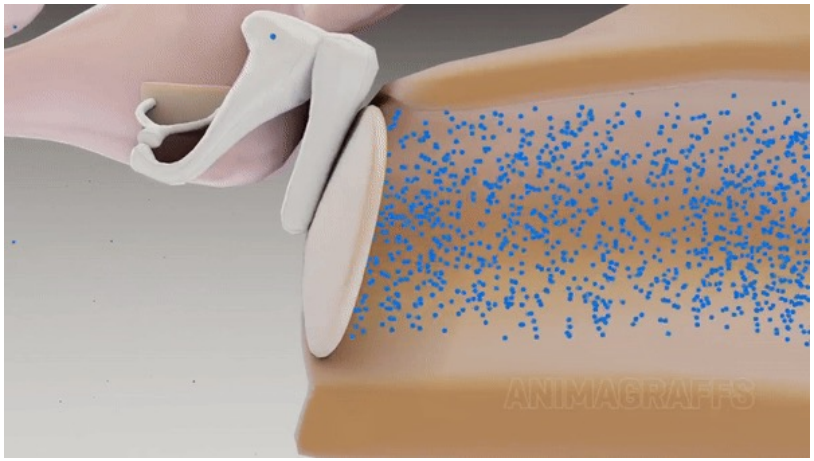
- (10 mins) Background: What is audio?
- *(15 mins) Can machines understand sound?* 👂
- *(15 mins) Can machines express themselves with sound?* 🗣️
- (5 mins) Interesting Recent Works
- Questions

Background

What is audio?

What is audio?

- Audio or Sound: The pressure wave through the air particle
- Human hearing is just movement detection



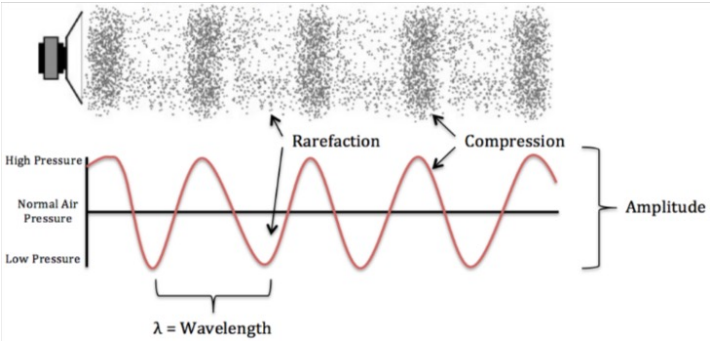
Credit: How Speakers Make Sound (youtube.com)

Credit: How Speakers Make Sound (youtube.com)

Before we work on audio, we need to capture it. But how?

The capture of audio wave

- Microphone
- The vibration level is electrically collected and saved.



We will refer to the captured "vibration level" as *samples* or *waveform*

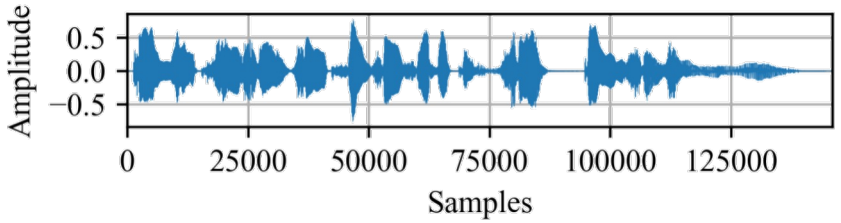
Credit: (2) How Microphone Works? (3D Animation) - YouTube

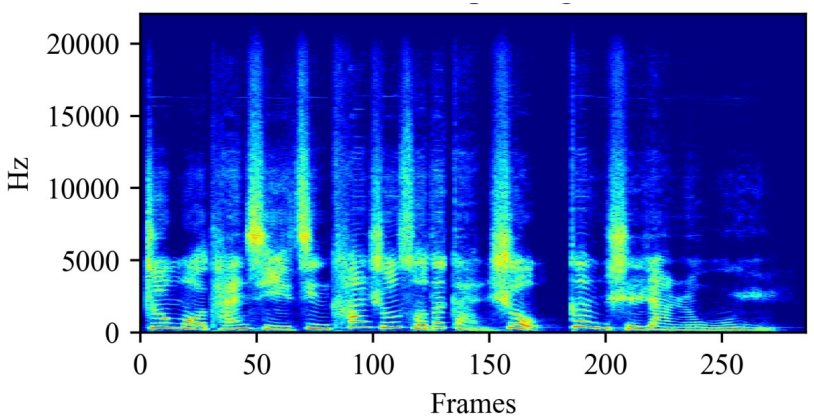
Let's try to capture the audio wave!!

Can we “see” the audio?

STFT

- **Waveform** (1-Dimensional)
  - The direct plot of “vibration”.
  - ✅ Original format
  - ❌ Hard to read
  - ❌ High-dimensional
- **Spectrogram** (2-Dimensional)
  - Short-time Fourier Transform (STFT)
  - ✅ Frequency information unfolded
  - ✅ Easy to understand

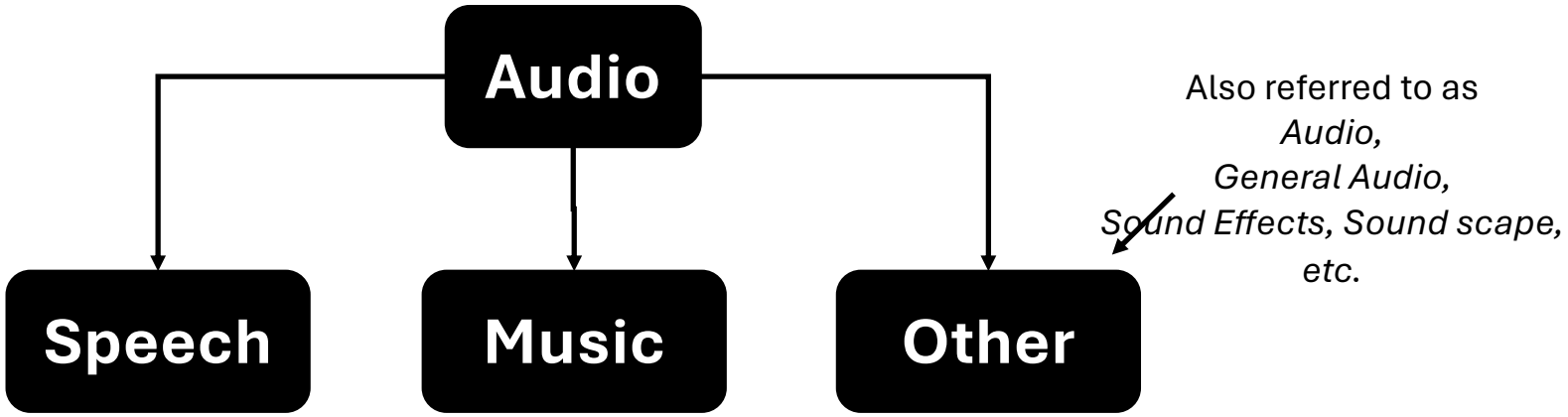




Let’s get back to the recording we made before.

In a research perspective

- Audio research usually consists of the following domain

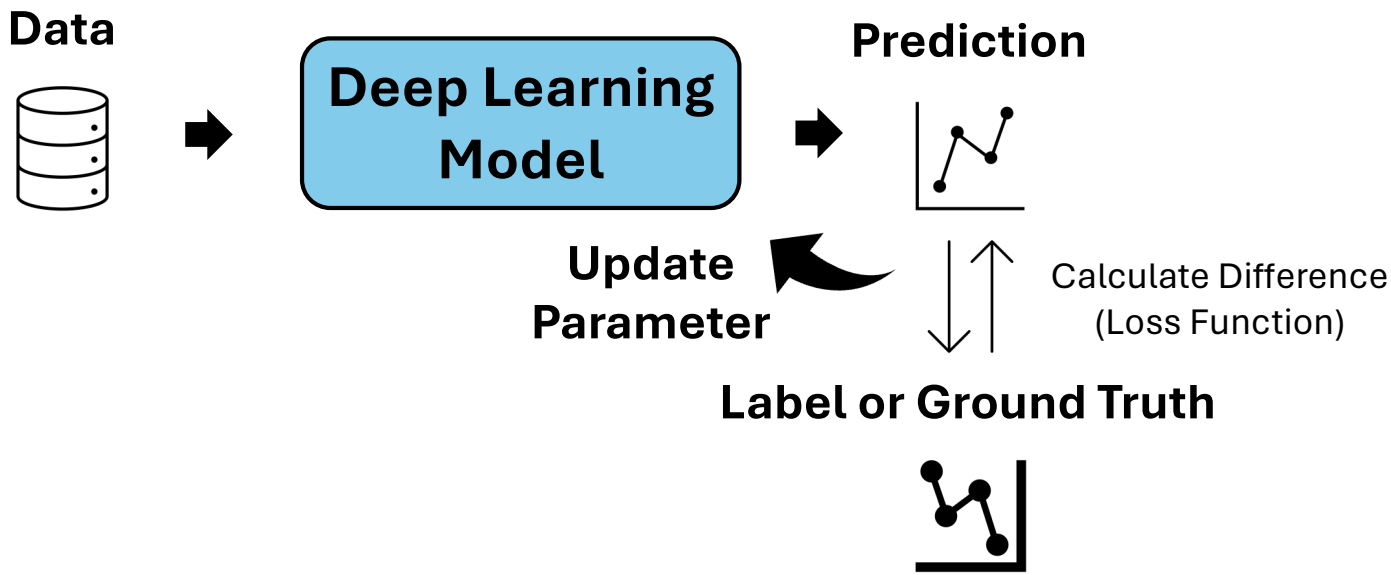


Can machines understand sound? 👂

1. (main) Supervised Learning
2. Self-supervised Learning
3. Contrastive Learning

Recap: Supervised Learning

- **Learning from labelled data**



The Dataset Matters

- **Audio & Labels**
  - e.g., AudioSet
    - 527 sound classes
    - 2 Million audio files from YouTube
- Audio & Text
- Audio & Labels + Timestamps
- …

audio-dataset

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

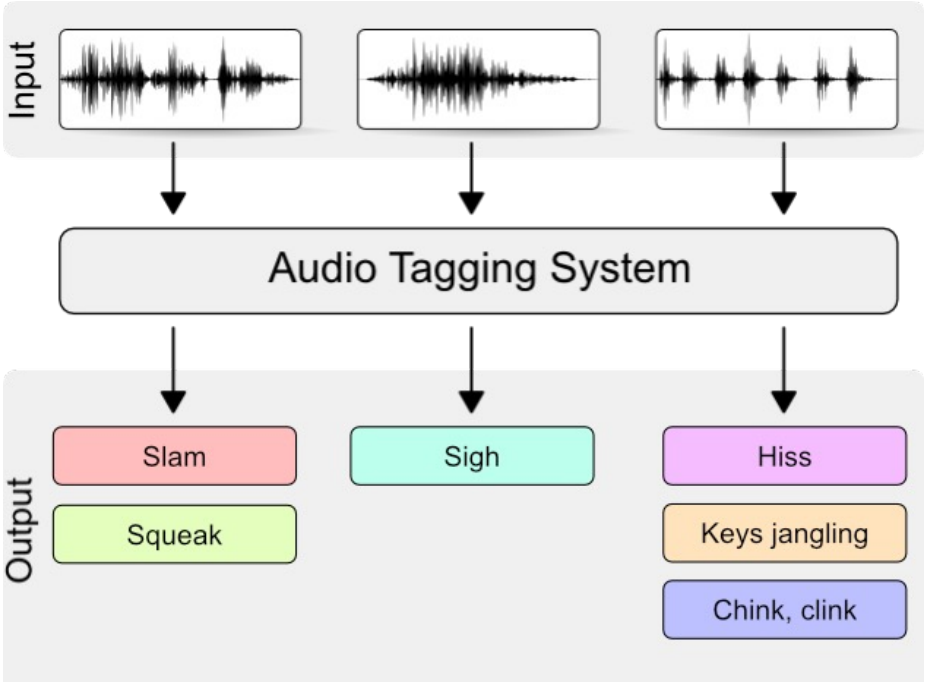
Channel, environment and background
- Acoustic environment
- Noise
- Sound reproduction

https://research.google.com/audioset/ontology/index.html

🔊 Example: [Domestic animals, pets, Squeak, Dog, Animal]

Supervised Learning

- **Example: Audio Tagging**
- Input: Audio
- Output: One or Multiple Labels
- Applications:
  - Organizing sound archive
  - Recommendation system
  - Surveillance monitoring
  - …



Supervised Learning

Let's say we have 527 classes in total.

Class ID

Example: [Domestic animals (**129**), pets (**130**), Squeak (**161**), Dog (**153**), Animal (**128**)]

Prediction Target:

One-hot Vector 1*527

Model Prediction:

Class Probability Prediction 1*527

How to get this prediction?

Audio Tagging - Model

- **Audio can be processed similarly as images**

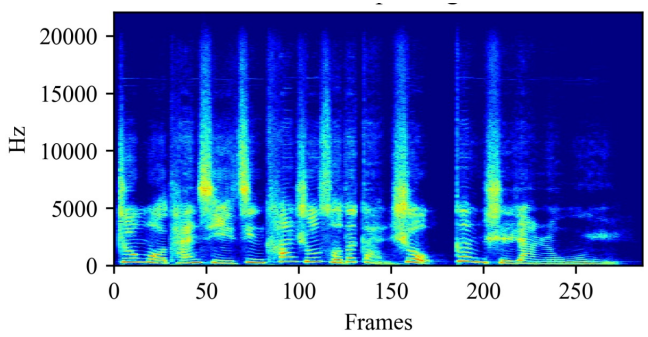


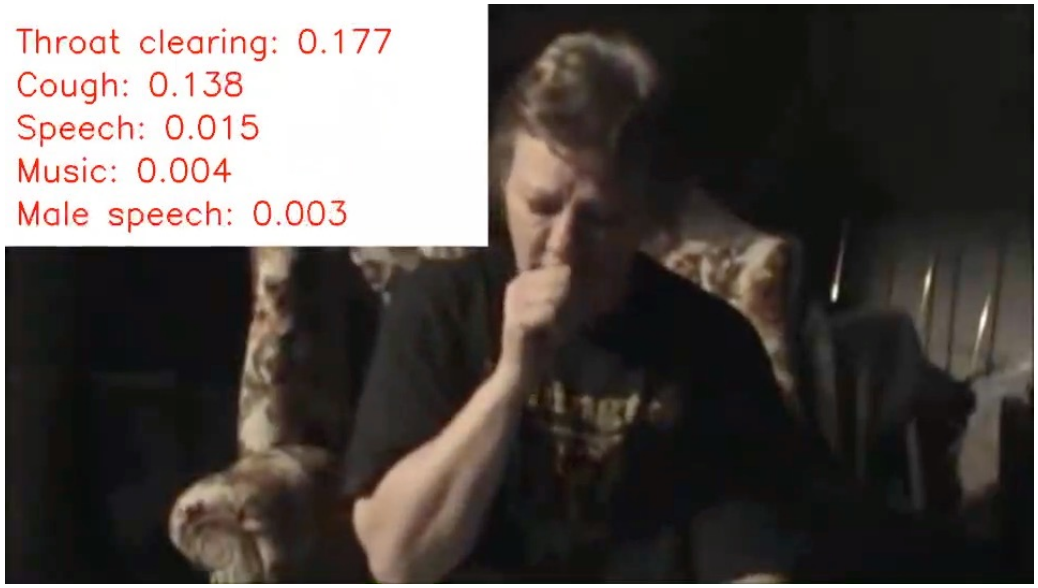
TABLE II
RESNETS FOR AUDIOSET TAGGING

ResNet22	ResNet38	ResNet54
Log mel spectrogram 1000 frames ×64 mel bins		
$(3 \times 3 \text{ @ } 512, \text{BN}, \text{ReLU}) \times 2$		
Pooling 2×2		
$(\text{BasicB @ } 64) \times 2$	$(\text{BasicB @ } 64) \times 3$	$(\text{BottleneckB @ } 64) \times 3$
Pooling 2×2		
$(\text{BasicB @ } 128) \times 2$	$(\text{BasicB @ } 128) \times 4$	$(\text{BottleneckB @ } 128) \times 4$
Pooling 2×2		
$(\text{BasicB @ } 256) \times 2$	$(\text{BasicB @ } 256) \times 6$	$(\text{BottleneckB @ } 256) \times 6$
Pooling 2×2		
$(\text{BasicB @ } 512) \times 2$	$(\text{BasicB @ } 512) \times 3$	$(\text{BottleneckB @ } 512) \times 3$
Pooling 2×2		
$(3 \times 3 \text{ @ } 2048, \text{BN}, \text{ReLU}) \times 2$		
Global pooling		
FC 2048, ReLU		
FC 527, Sigmoid		

Baseline ResNet system for PANNs

Kong, Qiuqiang, et al. "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 2880-2894.

Audio Tagging Example

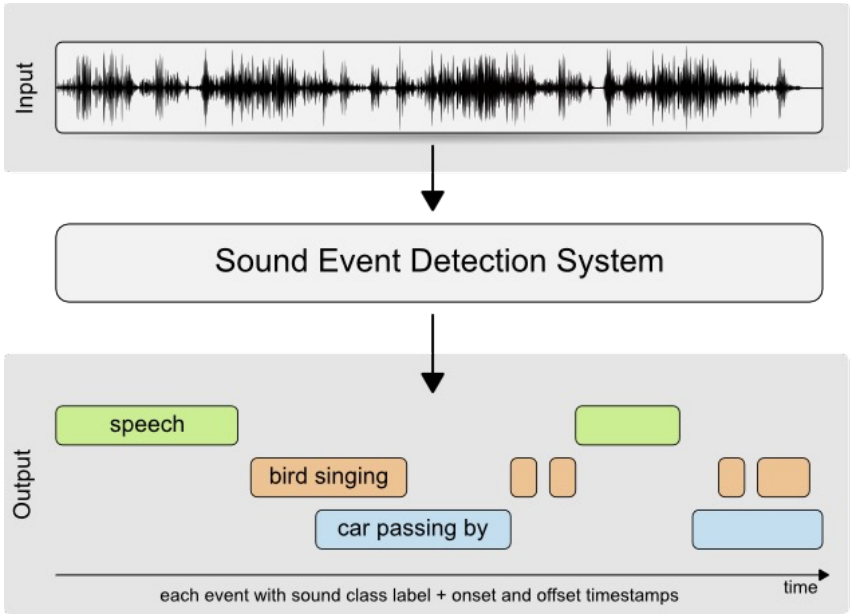


Code open-sourced: qiuqiangkong/audioset_tagging_cnn (github.com)

Kong, Qiuqiang, et al. "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 2880-2894.

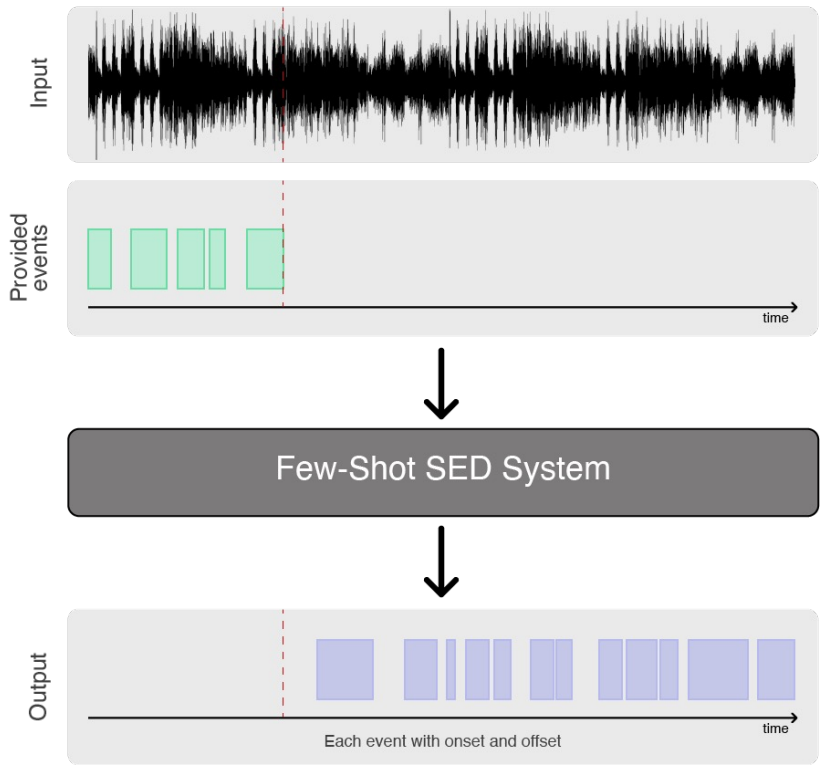
Moving Forward

- **Sound Event Detection**
  - Predict label + timestamp
- Recognize novel sound
  - One/Few-shot Sound Recognition
- Recognize piano key
  - Piano transcription
- …



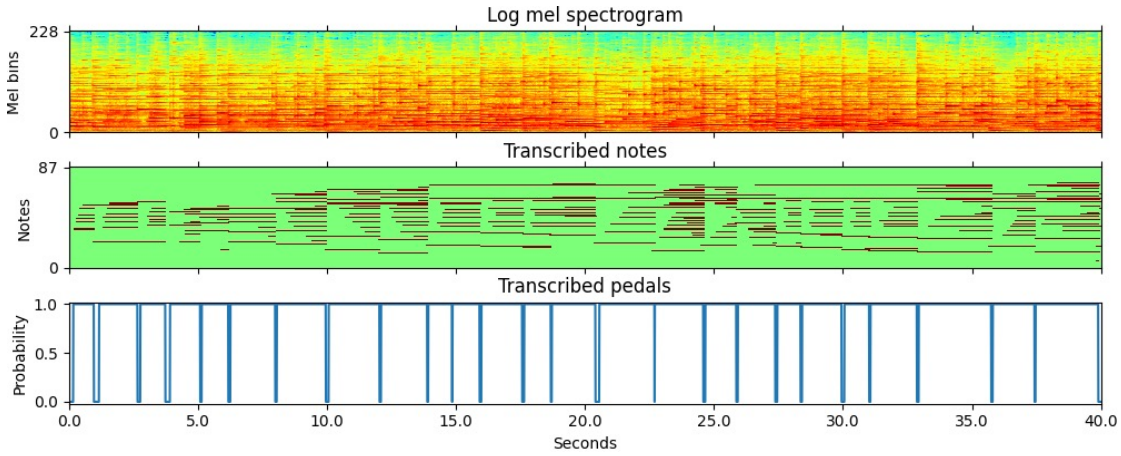
Moving Forward

- Sound Event Detection
  - Predict label + timestamp
- **Recognize novel sound**
  - One/Few-shot Sound Recognition
- Recognize piano key
  - Piano transcription
- …



Moving Forward

- Sound Event Detection
  - Predict label + timestamp
- Recognize novel sound
  - One/Few-shot Sound Recognition
- **Recognize piano key**
  - Piano transcription
- …



Suggest Reading

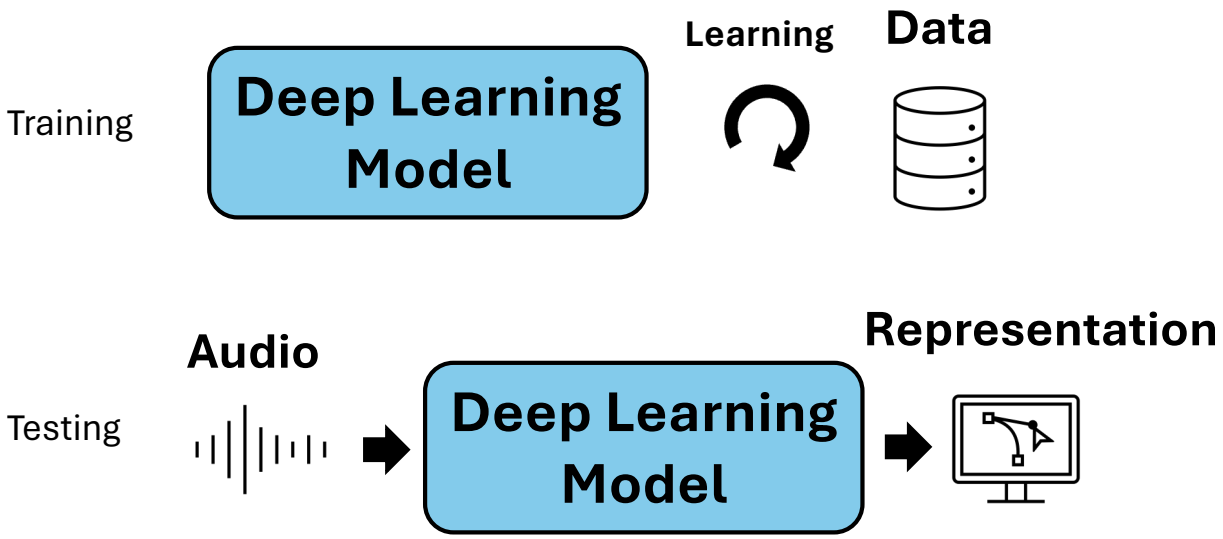
- Mesaros, Annamaria, et al. "Sound event detection: A tutorial." IEEE Signal Processing Magazine, 2021
- Kong, Qiuqiang, et al. ”**PANNs**: Large-scale pretrained audio neural networks for audio pattern recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020
- Gong, Yuan, Yu-An Chung, and James Glass. ”**PSLA**: Improving audio tagging with pretraining, sampling, labelling, and aggregation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021

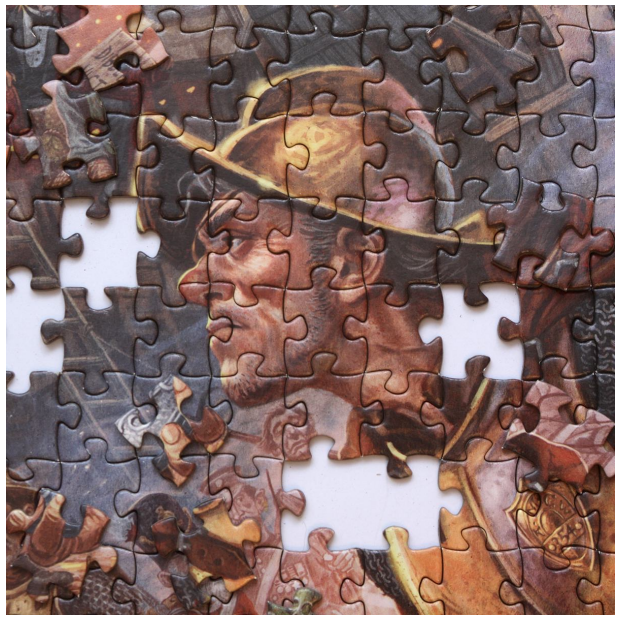
Audio Annotation is Painful

How to learn without the audio annotation?

Recap: Self-Supervised Learning (SSL)

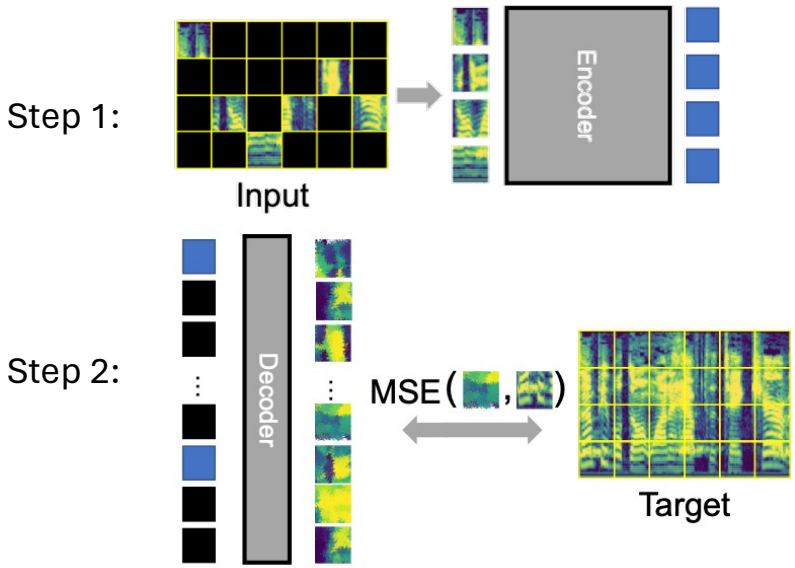
- **Learning from massive unlabelled data**

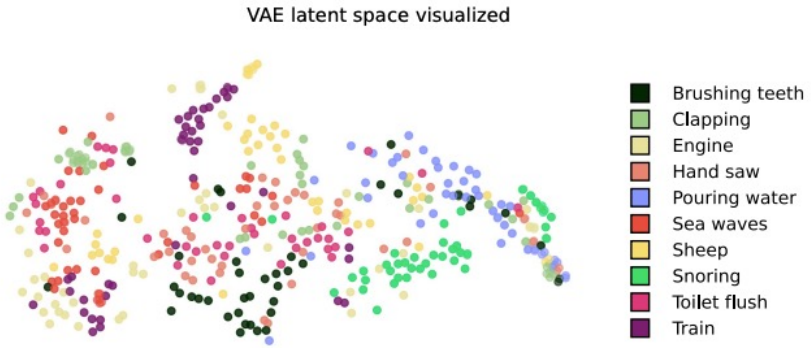


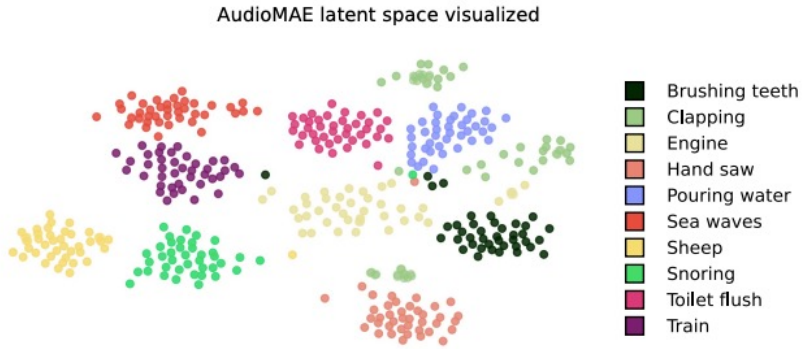


SSL example: AudioMAE

- **Learning to restore missing data**





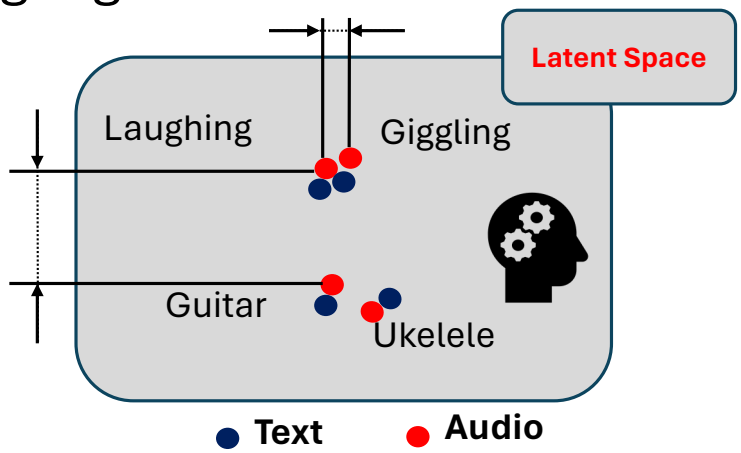


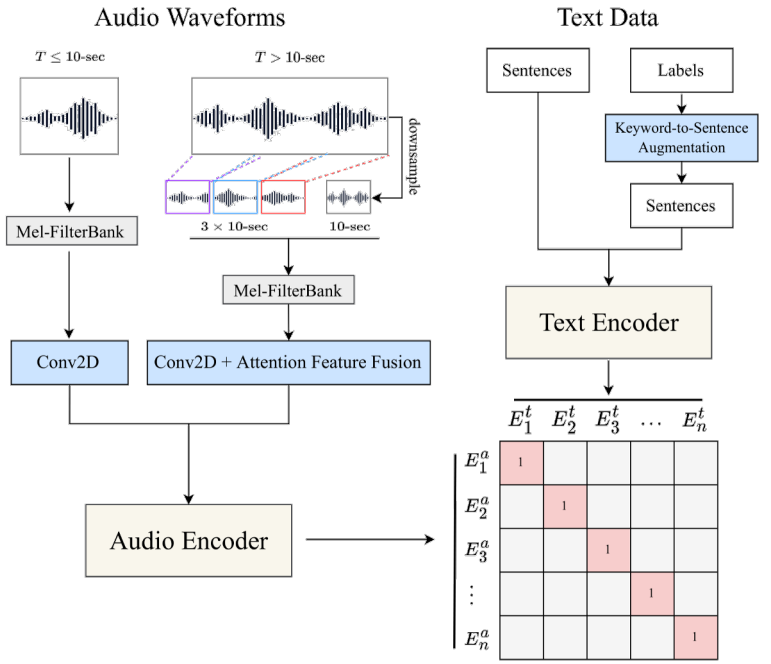
Suggest Reading

- Baevski, Alexei, et al. "**wav2vec 2.0**: A framework for self-supervised learning of speech representations." NeurIPS, 2020
- Hsu, Wei-Ning, et al. "**HuBERT**: Self-supervised speech representation learning by masked prediction of hidden units." IEEE TASLP 2021
- Chen, Sanyuan, et al. "**BEATs**: Audio Pre-Training with Acoustic Tokenizers." *ICML*, 2023.
- Chen, Sanyuan, et al. ”**WavLM**: Large-scale self-supervised pre-training for full stack speech processing." IEEE JSTSP 2022

Contrastive Learning

- Learning an aligned space between language and audio





🔗 Welcome to the AI Sound Searching System | surrey-audio (audioldm.github.io)

Wu, Yusong, et al. "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

Suggest Reading

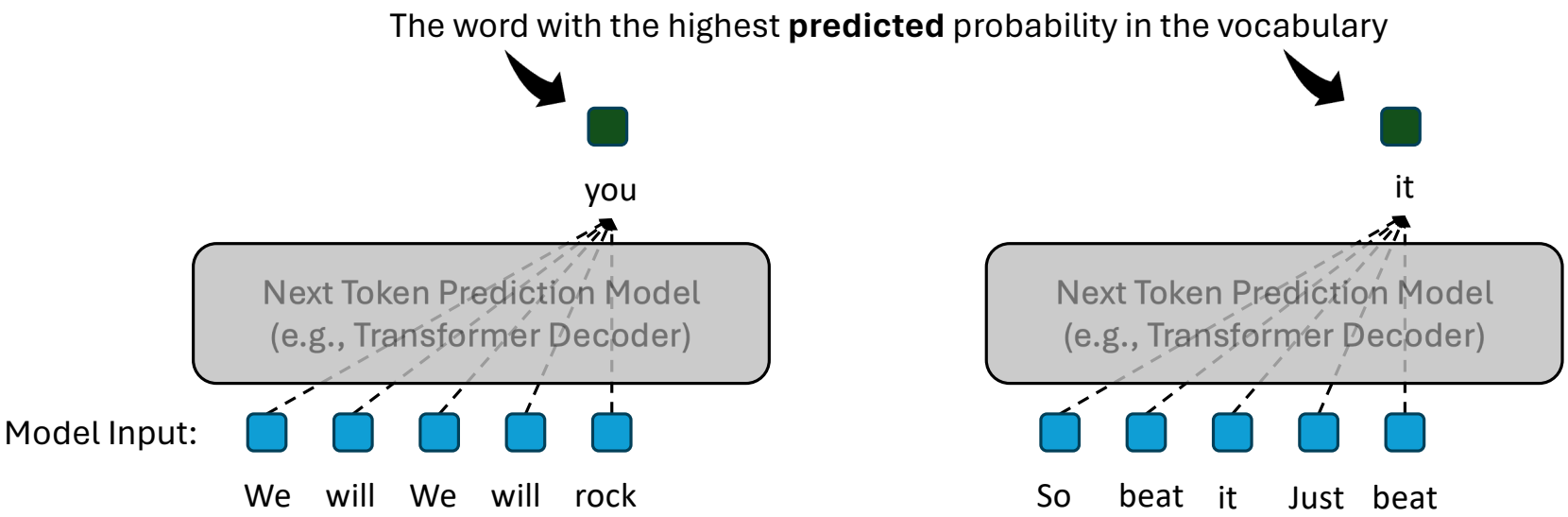
- **CLAP:** Wu, Yusong, et al. "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation." IEEE ICASSP 2023.
- **Imagebind:** Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
- **CLIP:** Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Can machines express themselves with sound? 🗣️

The Auto-regressive Approaches

Auto-regressive modeling

- **Language modeling:** Predict the probability of the **next token**

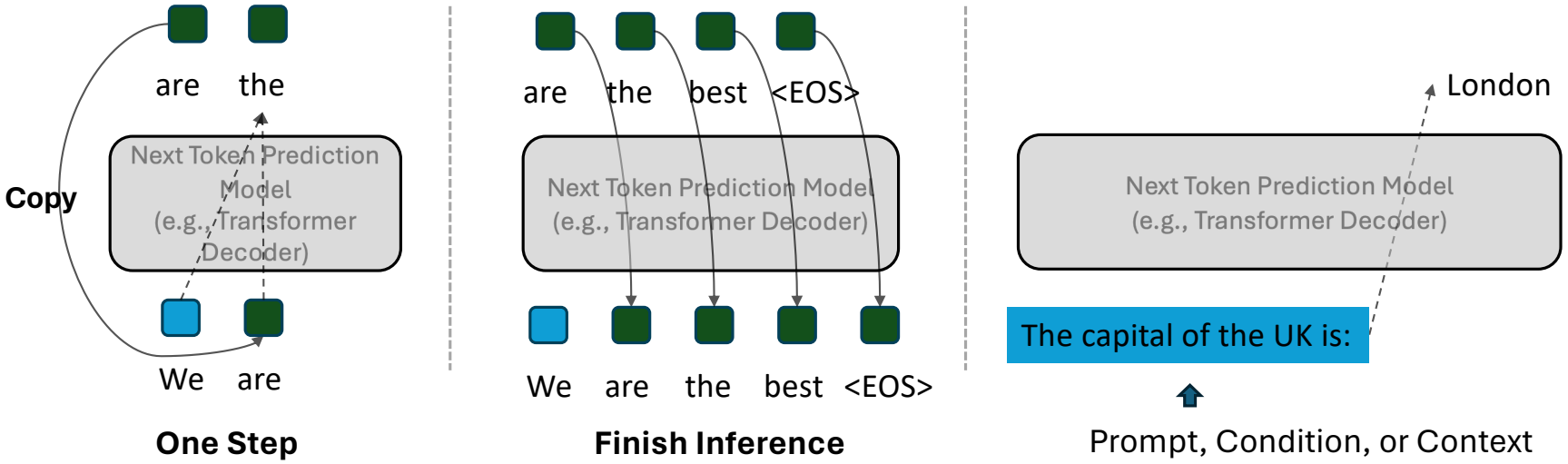


The **training process** of the language model

Auto-regressive Inference

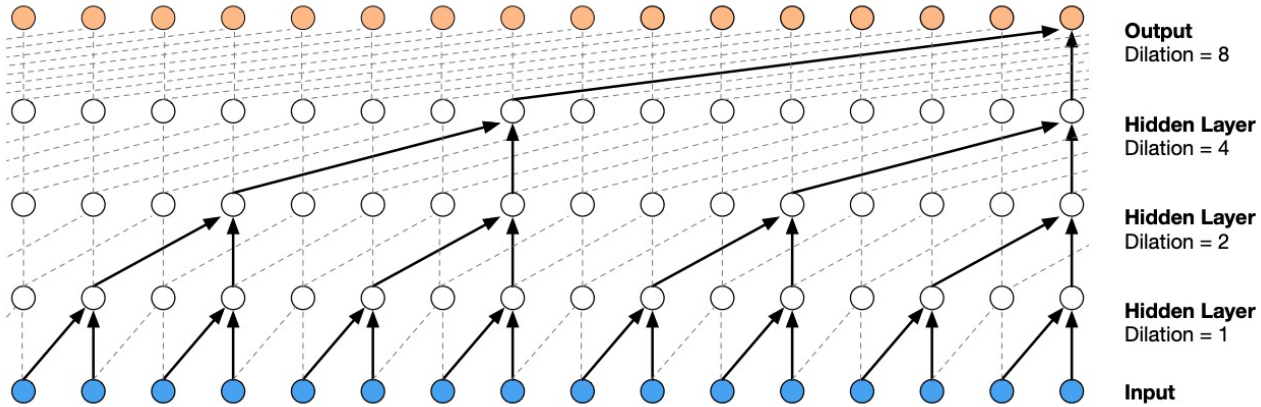
Can we do the same for audio?

- Predict the token sequence one by one.
- Later prediction is based on earlier prediction



WaveNet (Aaron et al. 2016)

- Model the waveform sequence?
- A second of 48kHz music has **48000** samples.

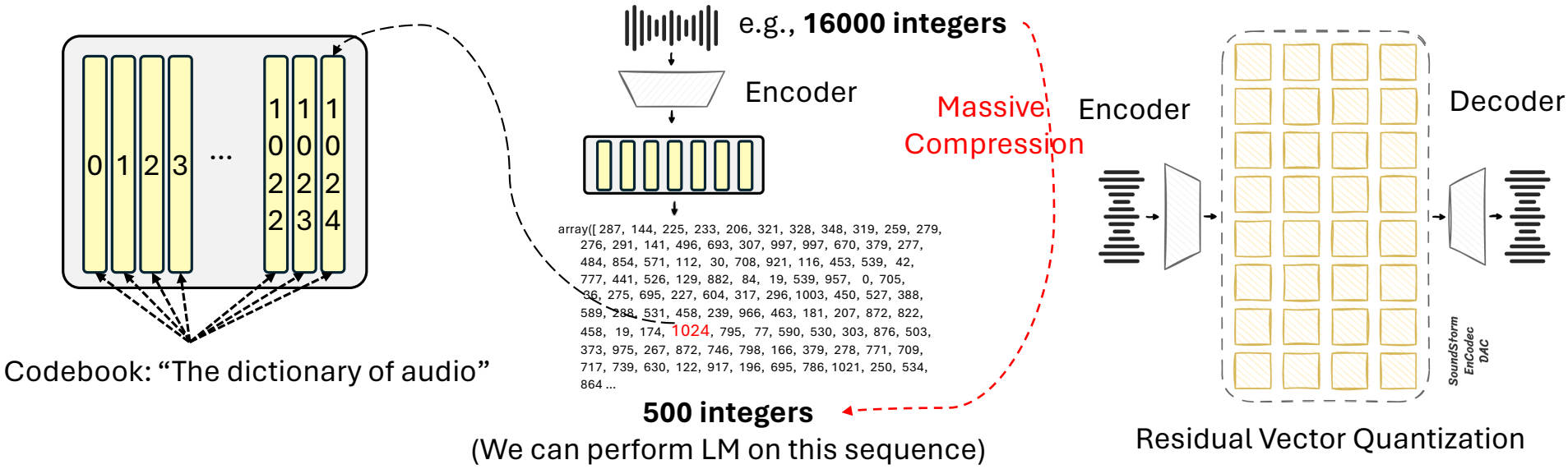


Dilated causal convolution

✅ Straightforward
✅ Nice result around 2017-2019
❌ Super heavy computation
❌ Slow inference
❌ Error propagation

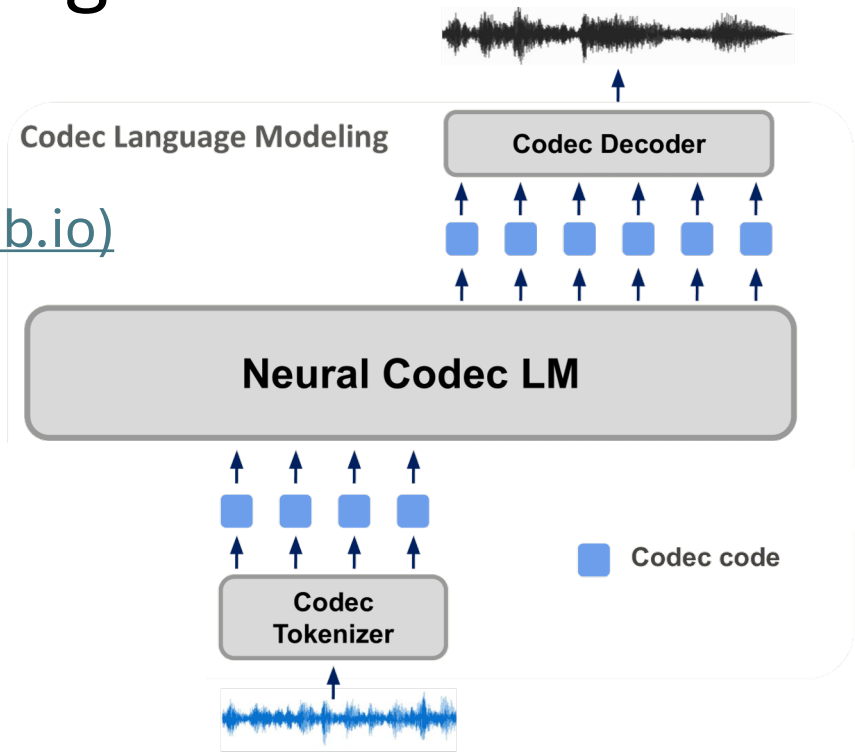
Neural Audio Codec

- The “language ” of audio
- Compression and discretization of audio signal



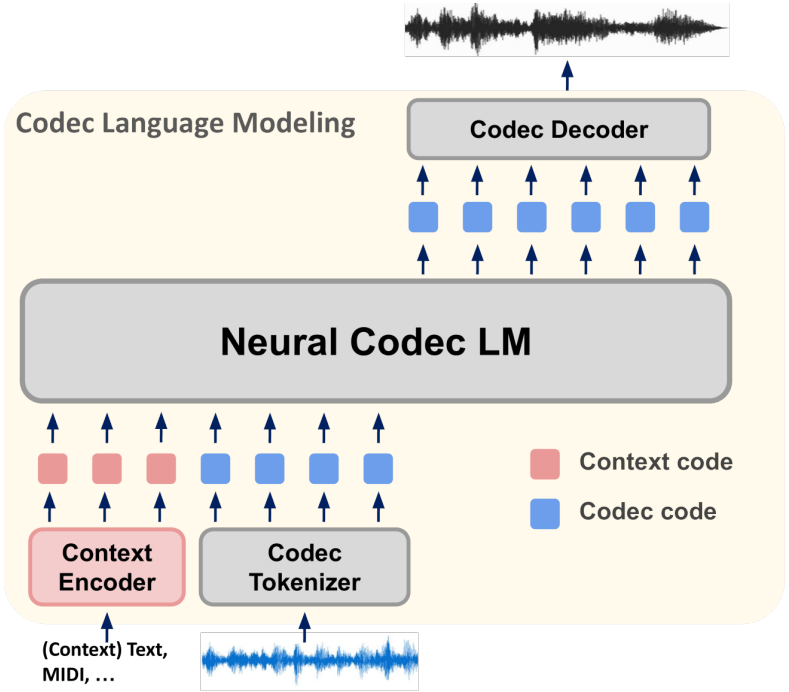
Audio Language Modeling

- Audio Continuation
- AudioLM (google-research.github.io) (Borsos *et al. 2022*)
- **We are more interested in conditional generation.**



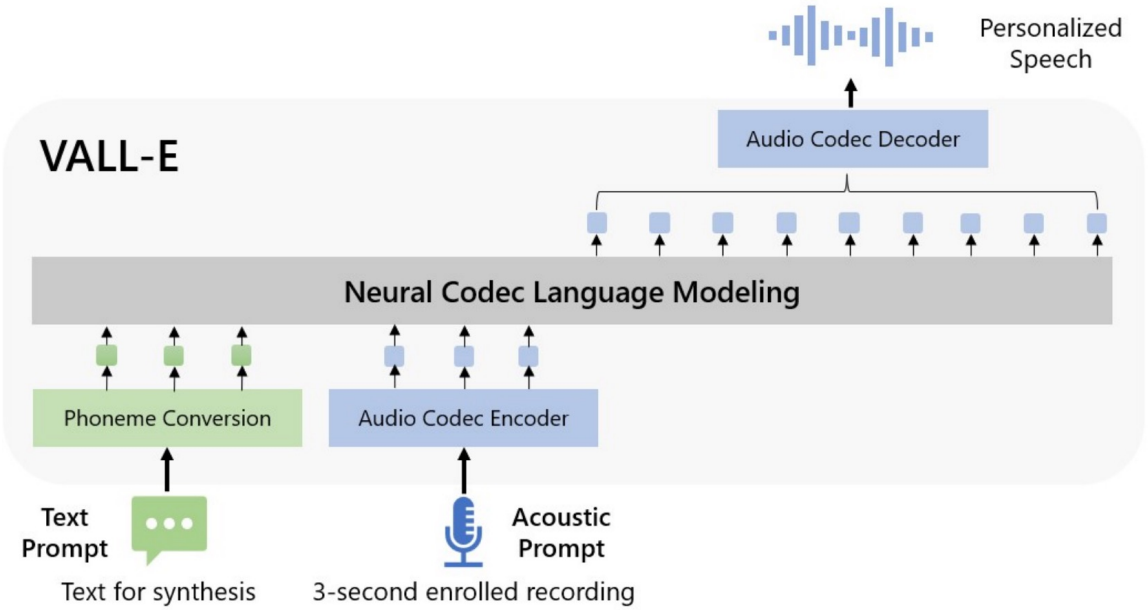
Conditional Audio Language Modeling

- Context Encoder
  - Text: Text-to-Audio
  - Video: Video-to-Audio
  - MIDI: Music Performance
  - Audio: Audio Editing
  - …



Example: VALL-E (*Wang et al. 2023*)

- Zero-shot Text-to-Speech Generation



Suggest Reading

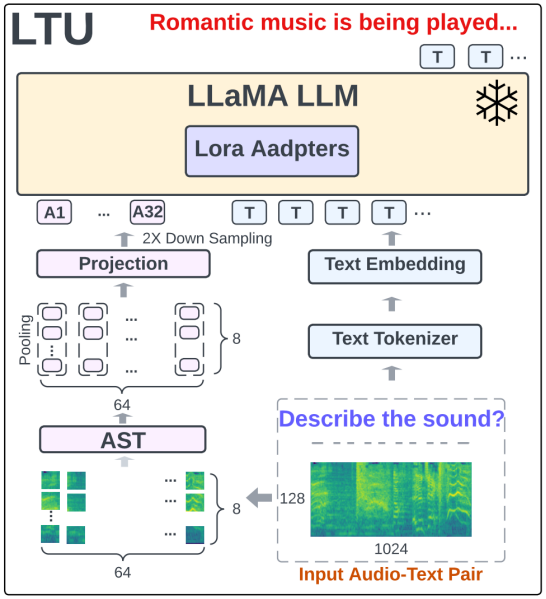
- **SoundStream:** Zeghidour, Neil, et al. "Soundstream: An end-to-end neural audio codec." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 495-507.
- **AudioLM:** Borsos, Zalán, et al. "AudioLM: a language modeling approach to audio generation." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).
- **VALL-E:** Wang, Chengyi, et al. "Neural codec language models are zero-shot text to speech synthesizers." arXiv preprint arXiv:2301.02111 (2023).
- **MusicGen:** Copet, Jade, et al. "Simple and controllable music generation." Advances in Neural Information Processing Systems 36 (2024).

Recent Works

What’s New

LTU (*Gong et al. 2024*)

- Teach language model to understand audio



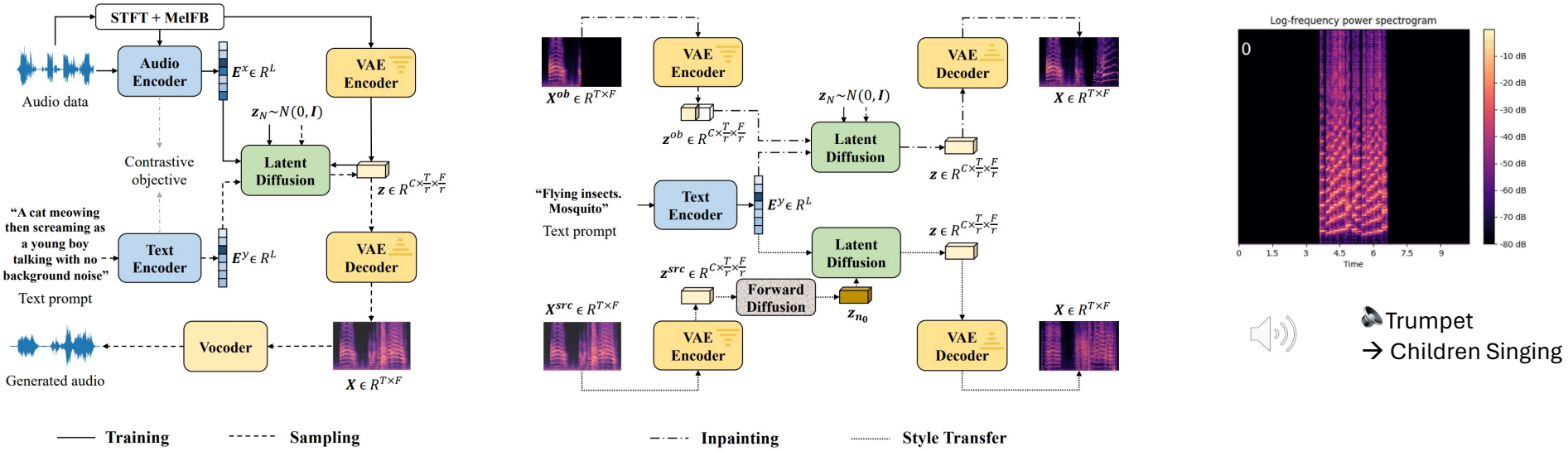
(Busy SIgnal Sounds in a Call)
Q: Classify the sound events in the audio clip. [listen]
LTU: Busy signal; Male speech, man speaking; Background noise; Tick; Speech.
Q: What makes you think it is a busy signal? [think]
LTU: The busy signal is characterized by a continuous tone, which is different from the sound of a ringing phone, which is a series of short pulses.
Q: What should the man do? [understand]
LTU: The man should wait for the busy signal to end before trying to call again or try a different phone number if necessary to reach the desired person or organization.

(Owls Night Calls)
Q: What can be inferred from the audio? Think step by step. [listen & think]
LTU: The audio clip contains the sound of wind, insects, and hooting. The wind is blowing and the insects are chirping. The hooting sound is likely from an owl or other bird of prey. The sound of the wind is high-pitched and hissing, while the insects are high-pitched and chirpy. The hooting sound is deep and resonant, suggesting a larger bird of prey or owl. The combination of these sounds suggests a natural outdoor environment, possibly at night. The hooting sound could also suggest a sense of danger or urgency, as owls are known to be predators.
Q: What type of video can this audio be used for dubbing? [understand]
LTU: This audio can be used for dubbing a horror movie.

YuanGongND/ltu: Code, Dataset, and Pretrained Models for Audio and Speech Large Language Model "Listen, Think, and Understand". (github.com)

AudioLDM 1&2 (*Liu et al. 2022 & 2023*)

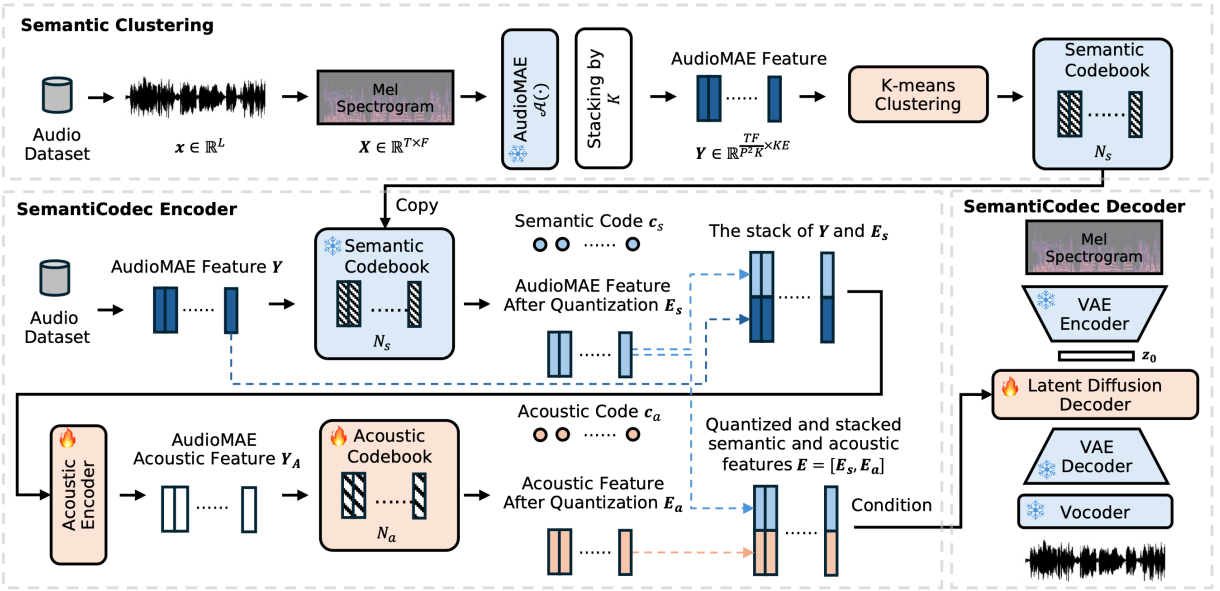
- Text to Audio Generation with Diffusion Model



🔗 Audioldm Text To Audio Generation - a Hugging Face Space by haoheliu

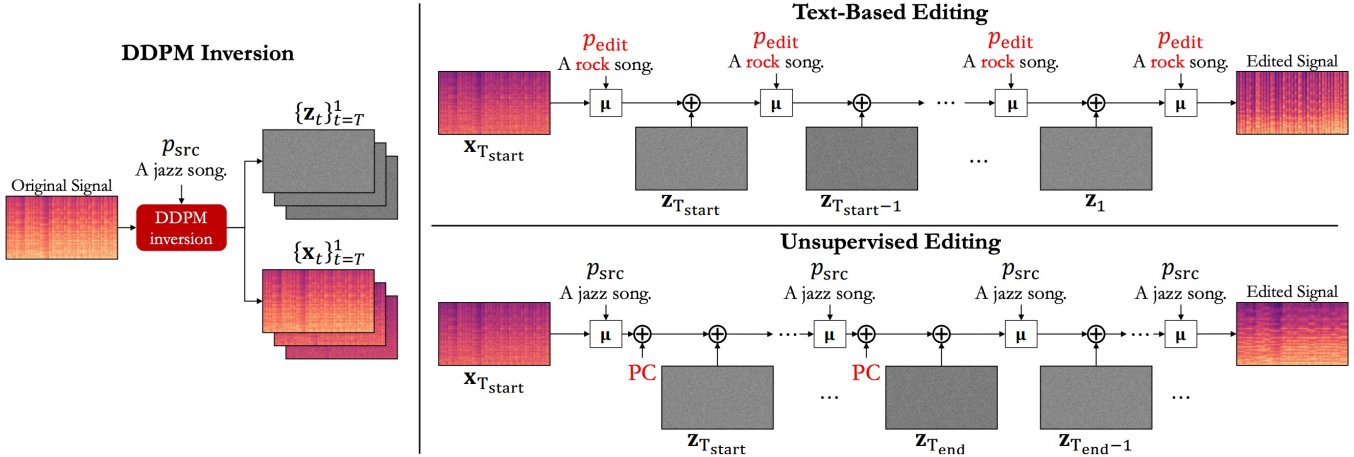
SemantiCodec (*Liu et al. 2024*)

- Compress a second of 16kHz audio into 50 integers



Demo: haoheliu.github.io/SemantiCodec/

Audio Editing (Manor & Tomer 2024)



Manor, Hila, and Tomer Michaeli. "Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion." *ICML 2024*.

Source
A recording of a happy upbeat **classical music piece**.

Target
A recording of a happy upbeat **arcade game soundtrack**.

[Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion (hilamanor.github.io)]

Thanks

- Audio and machine learning related Conferences: ICASSP, INTERSPEECH, ICML, ICLR, NeurIPS, …
  - Example: ICLR 2024 Papers
- *Understanding Deep Learning* - Book by Simon J. D. Prince
- Open source platforms:
  - Hugging Face – The AI community building the future.
  - GitHub
- Competitions:
  - DCASE2024 Challenge – DCASE
  - Kaggle: Your Machine Learning and Data Science Community