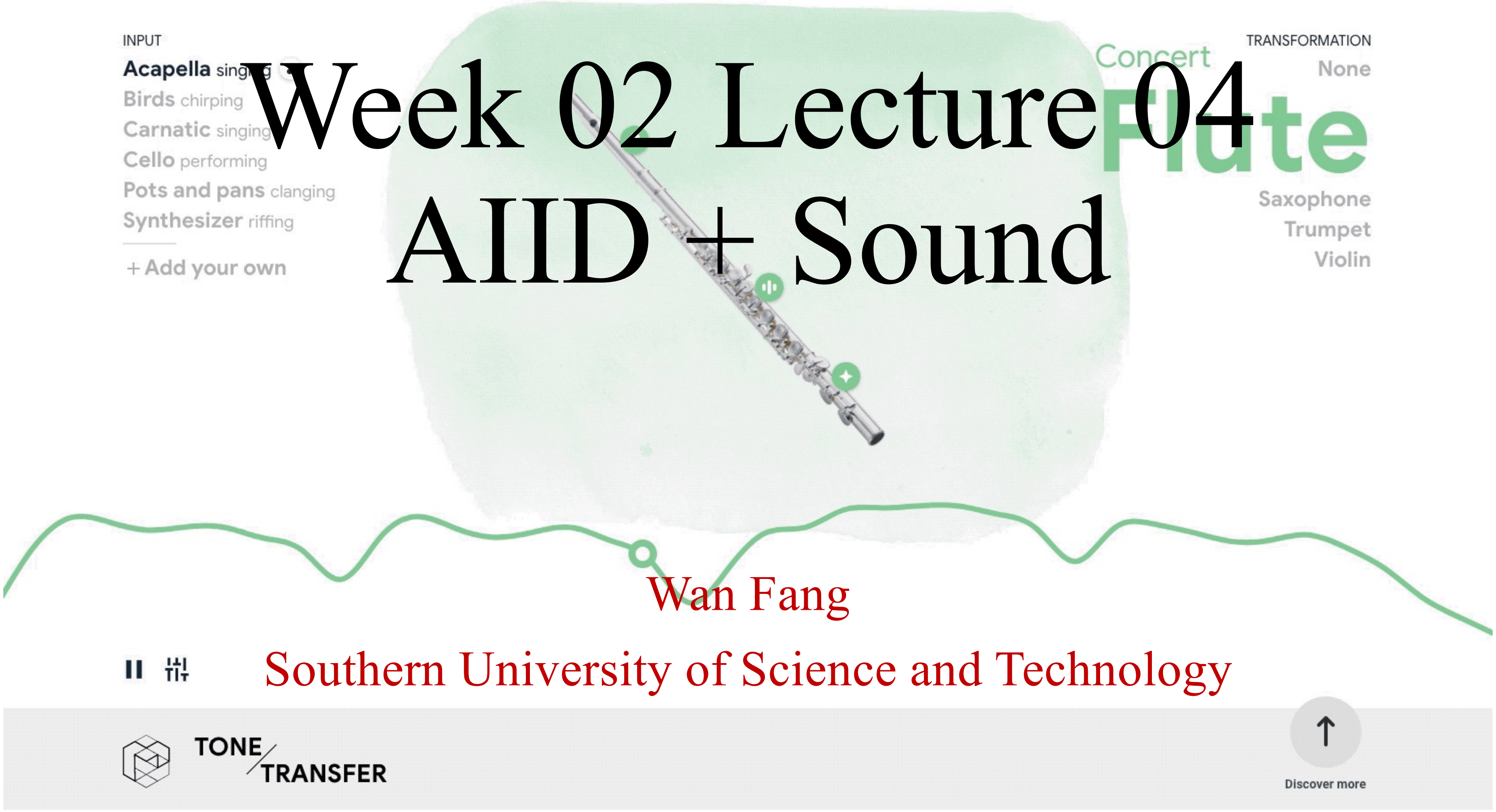
DS323: AI in Design (AIID)

Autumn 2023



Week 02 Lecture 04
AIID + Sound

Wan Fang

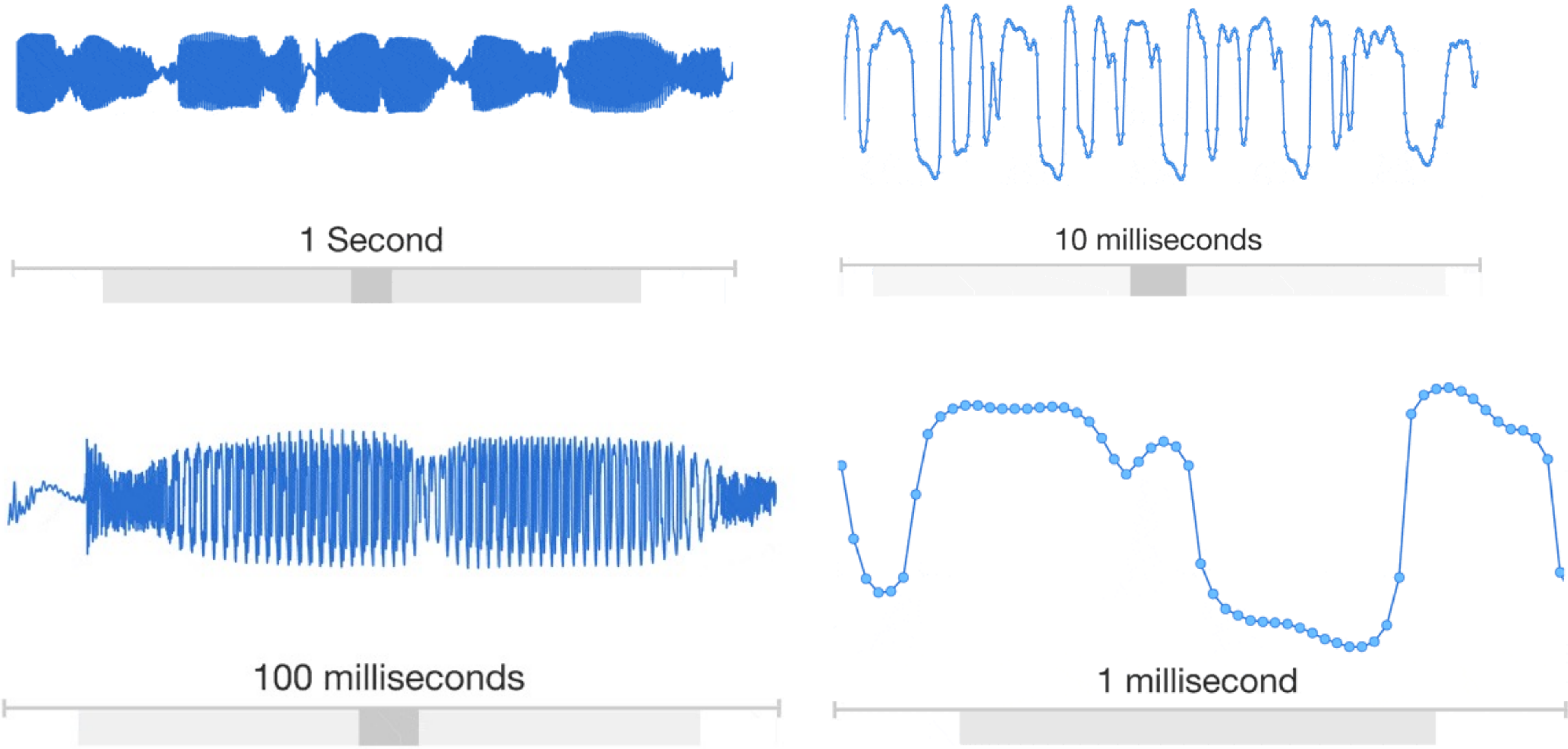
Southern University of Science and Technology

Source: https://magenta.tensorflow.org/tone-transfer

Agenda

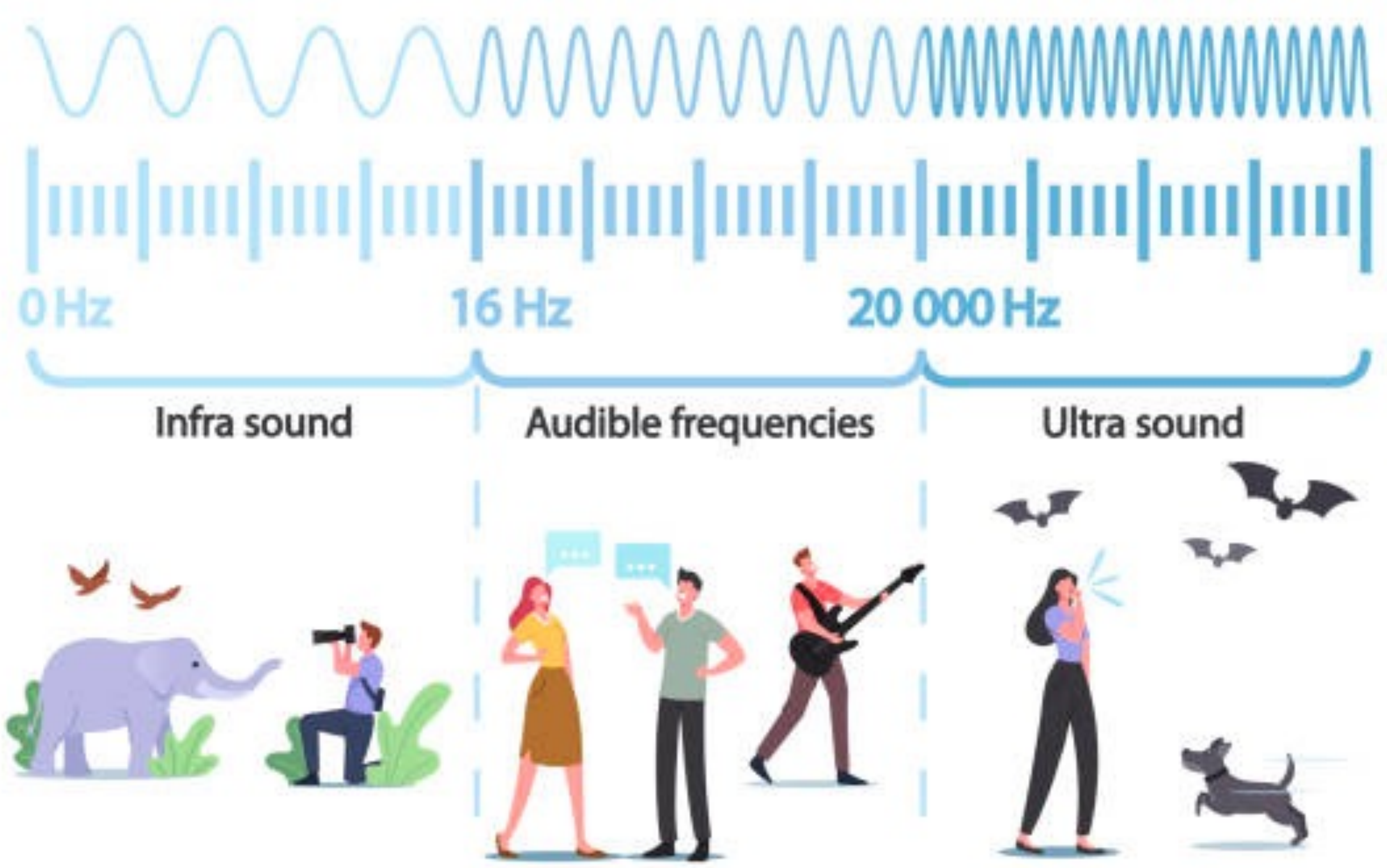
- Sound as Data
- Automatic Speech recognition (ASR)
- Voice Recognition
- Music Generation
    - Symbolic AI vs Audio AI systems
    - Tools to Make Your Own Generative Music
    - Concluding Thoughts and Further Questions

AIID + Sound

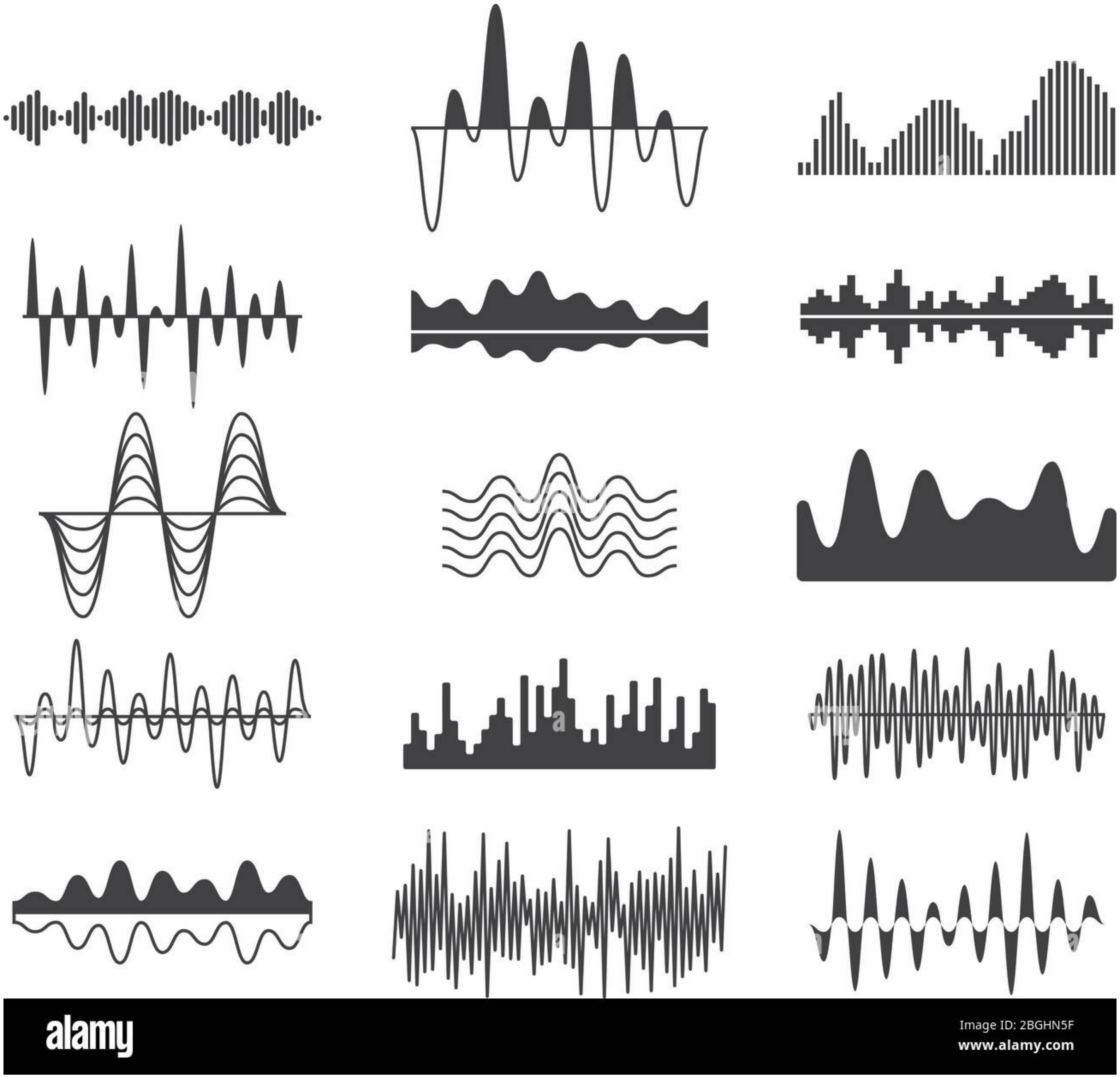


Audio samples from WaveNet paper (2018)

AIID + Sound


0 Hz
16 Hz
20 000 Hz
Infra sound
Audible frequencies
Ultra sound

AIID + Sound



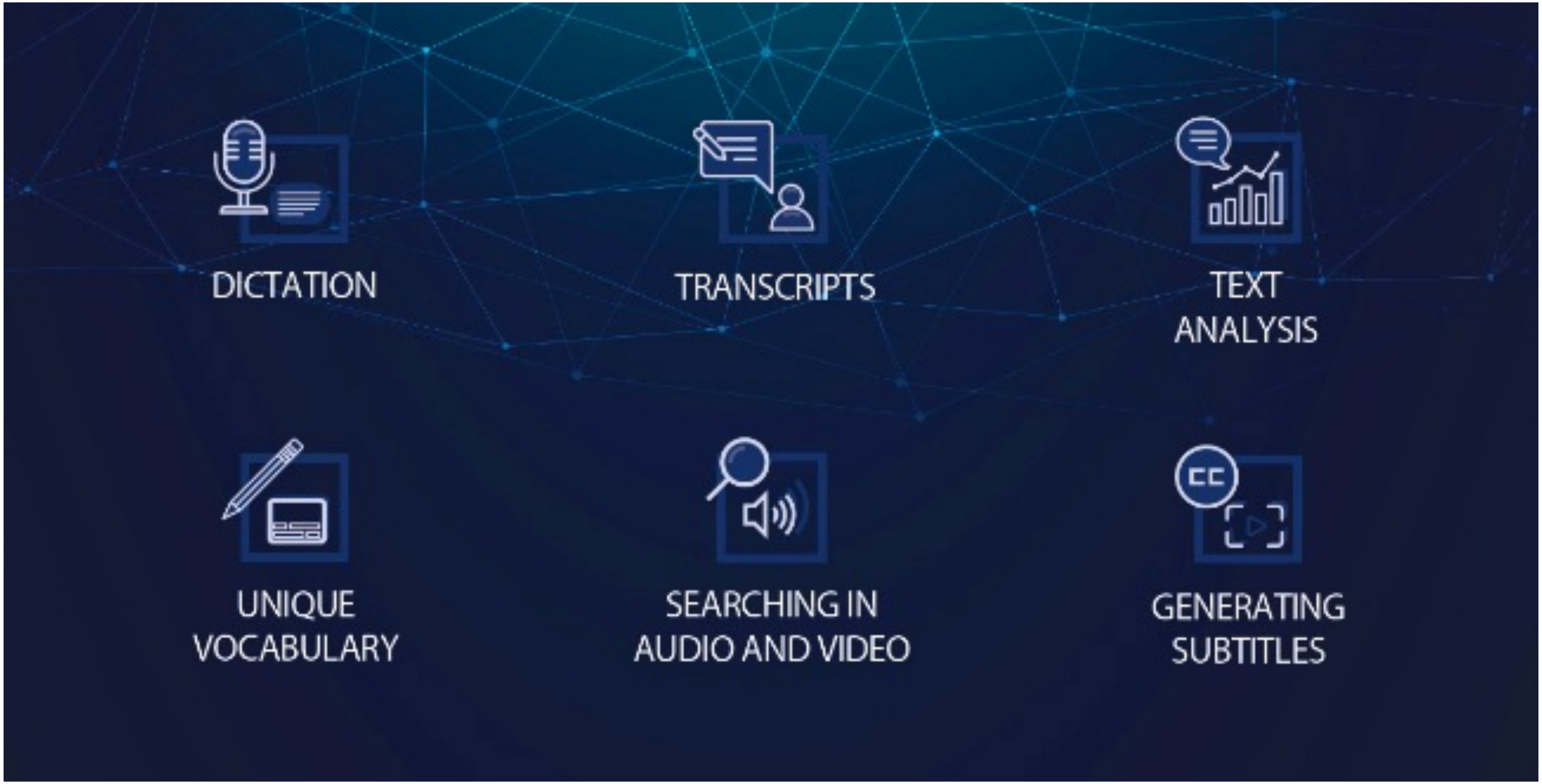
4 Properties of Sound

- Frequency
  pitch，音调
- Amplitude
  音强
- Timbre
  音色
- Duration
  音长

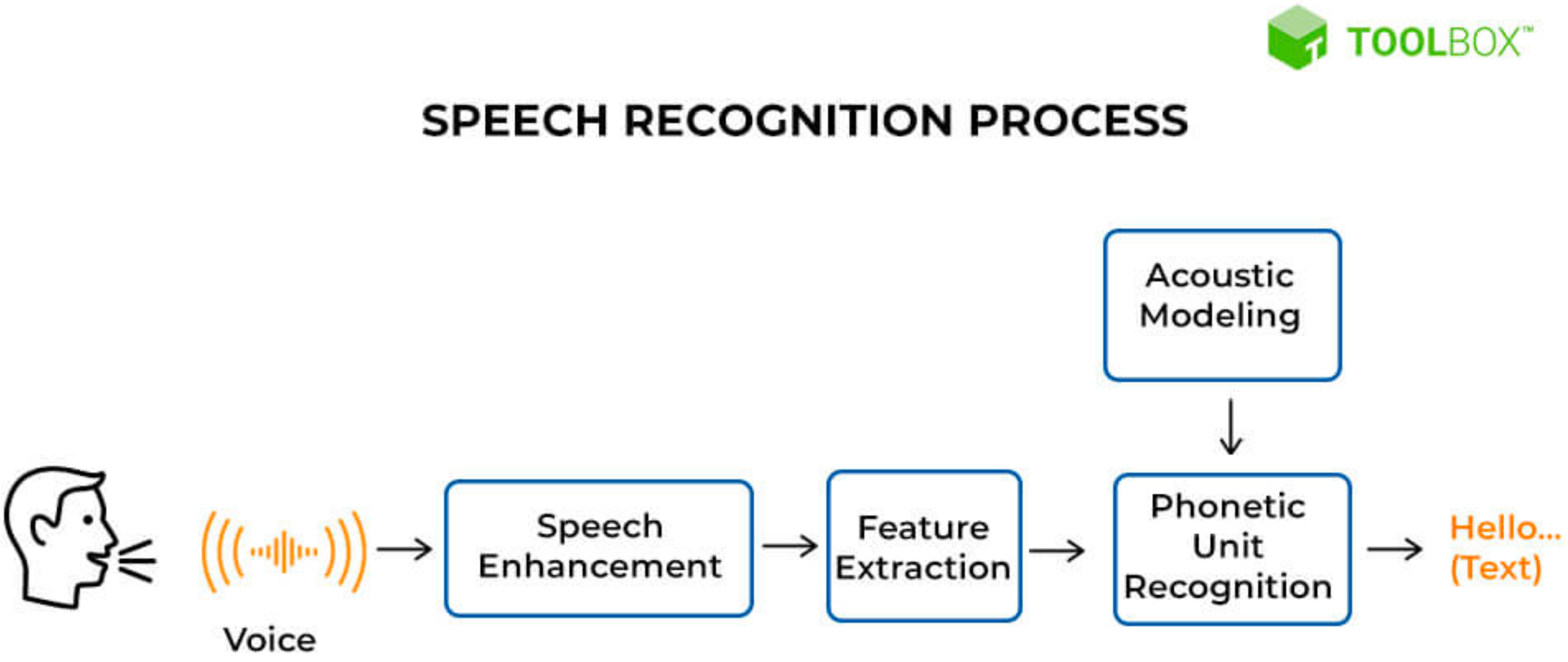
Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR)

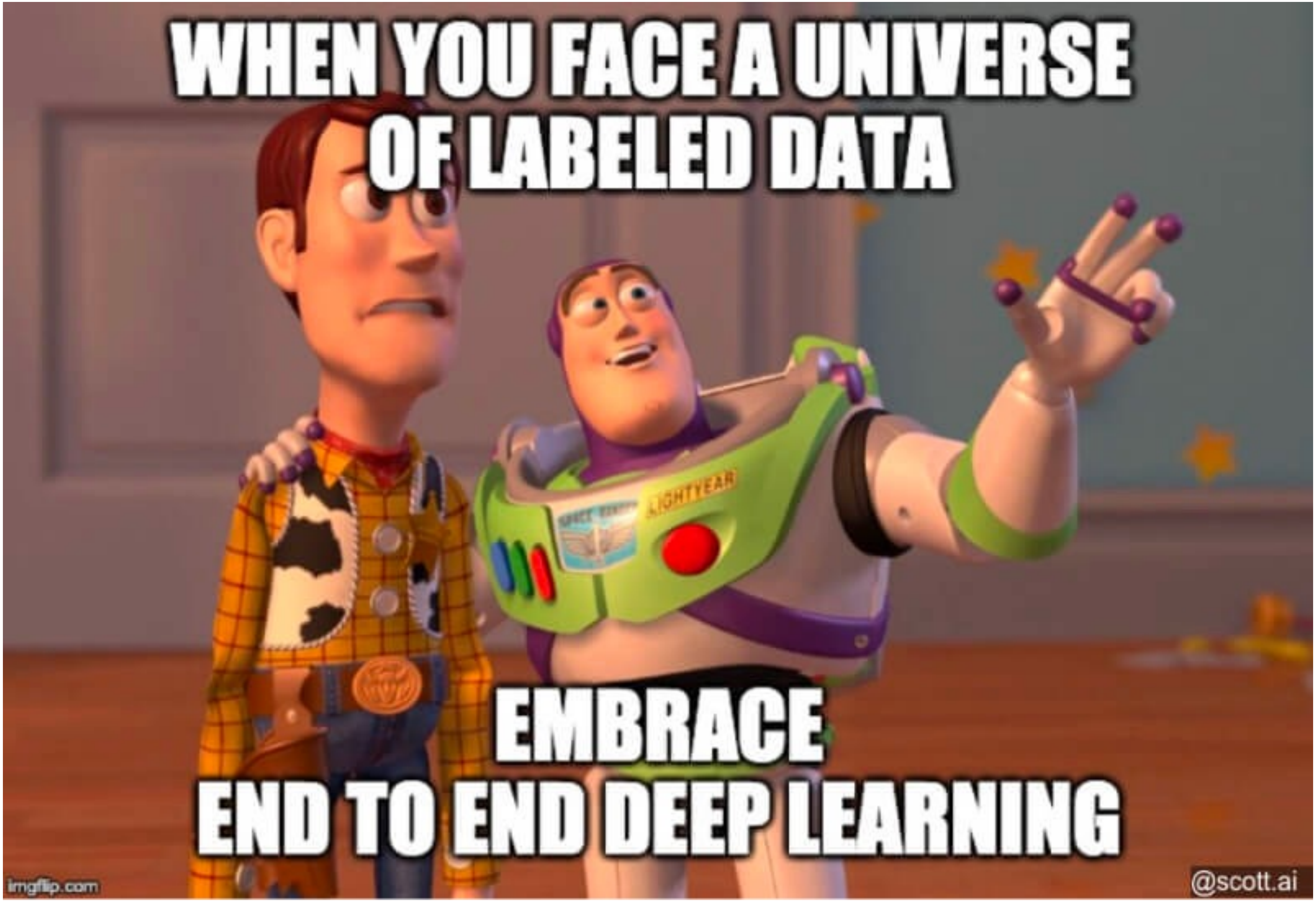
- Speech recognition is the ability of AI systems to identify spoken words and convert them into text.



Used to be like

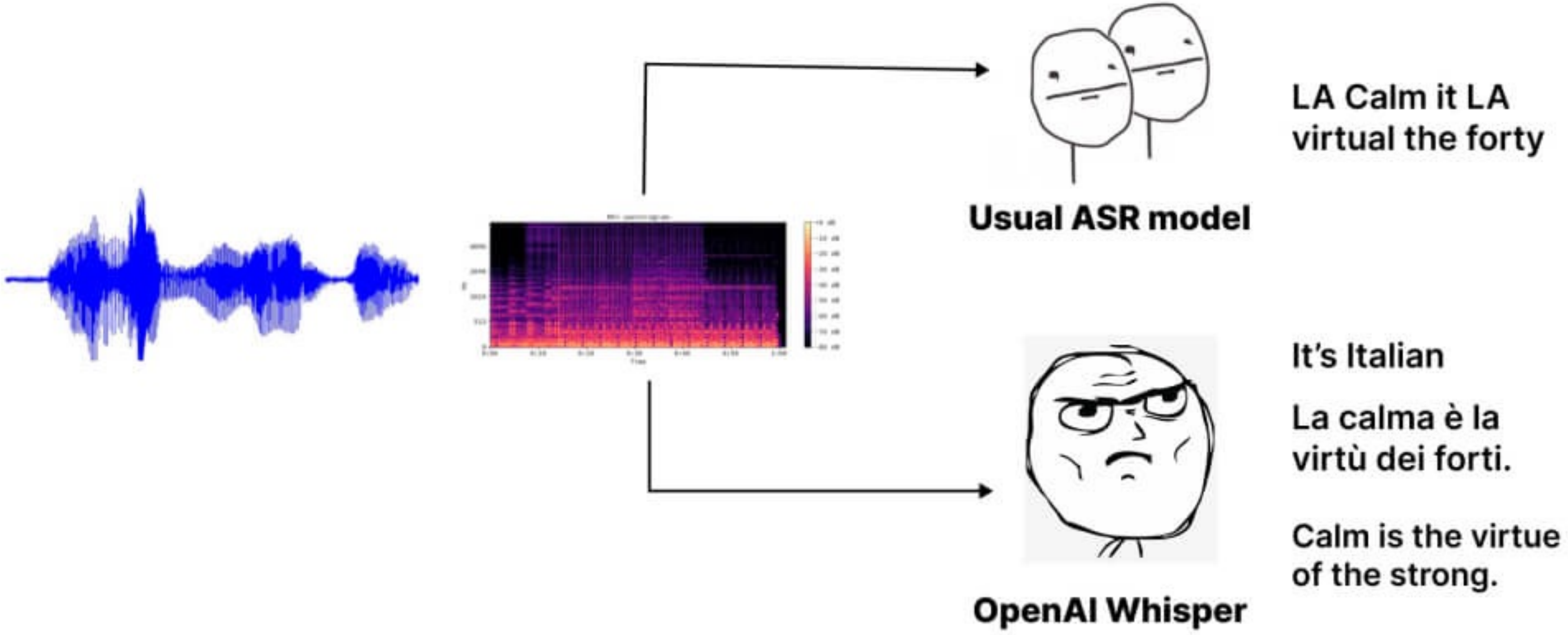


https://www.superannotate.com/blog/openai-whisper-automatic-speech-recognition-system


WHEN YOU FACE A UNIVERSE OF LABELED DATA
EMBRACE END TO END DEEP LEARNING
LIGHTYEAR
imgflip.com
@scott.ai

Whisper from OpenAI

- In 2022, this idea of training on **large data** to achieve cross-domain performance arrived in the world of speech recognition with OpenAI’s launch of Whisper.



LA Calm it LA
virtual the forty

Usual ASR model

It’s Italian

La calma è la
virtù dei forti.

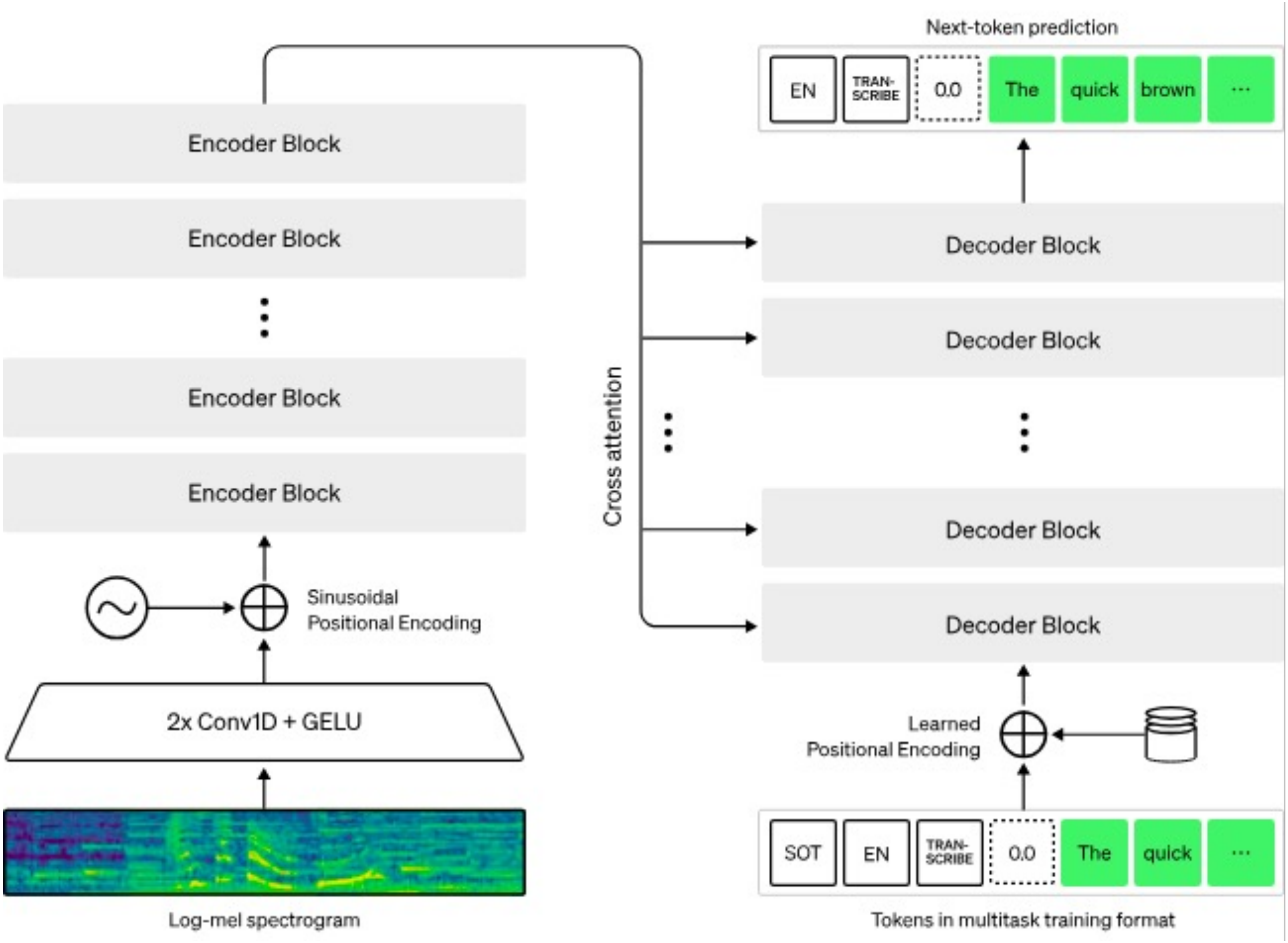
Calm is the virtue
of the strong.

OpenAI Whisper

Whisper from OpenAI

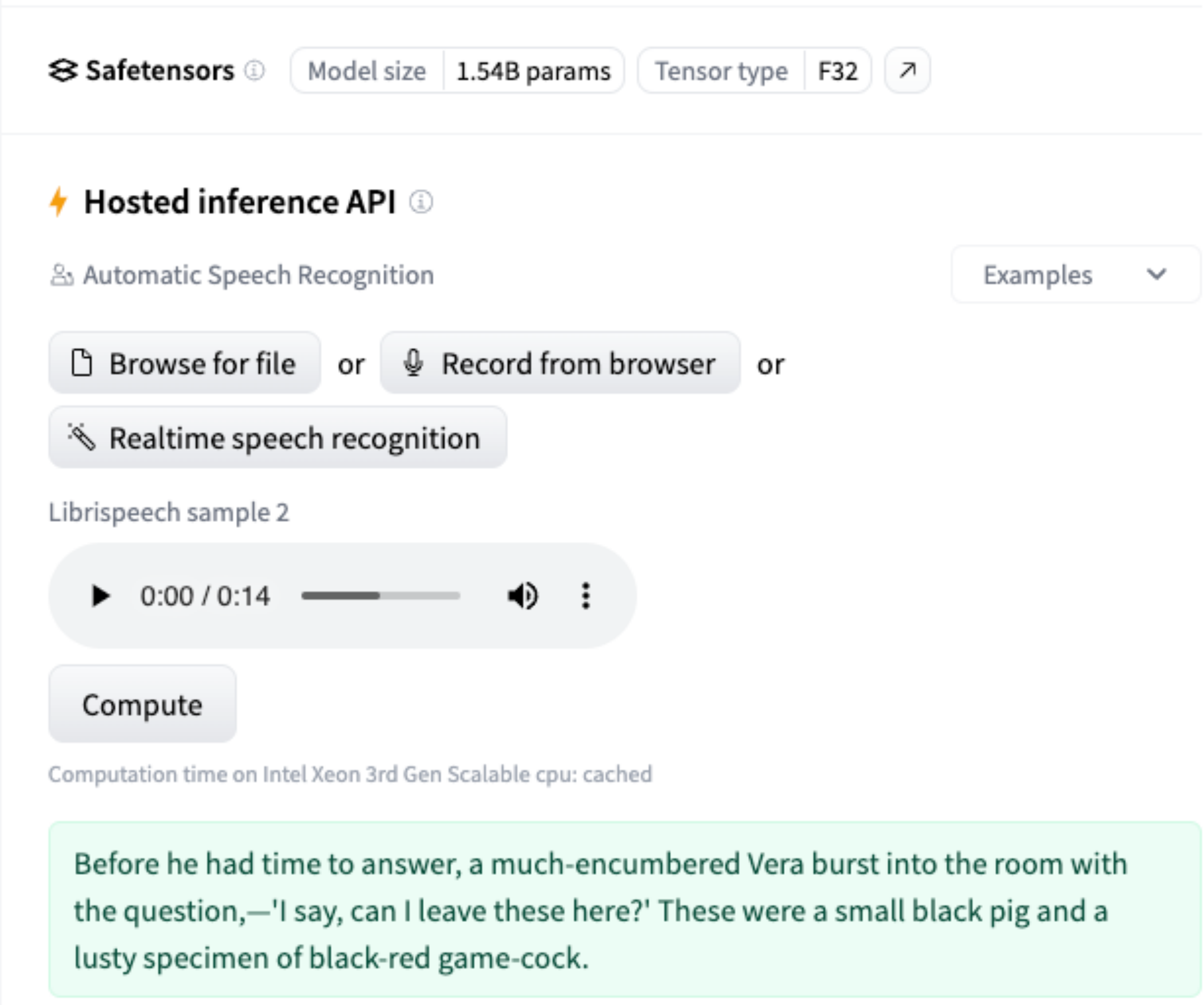
- 680,000 hours of multilingual and multitask supervised data
- A third of Whisper's audio dataset is non-English.

AIID + Sound


Next-token prediction
EN
TRAN-SCRIBE
0.0
The
quick
brown
···
Encoder Block
Encoder Block
Encoder Block
Encoder Block
Decoder Block
Decoder Block
Decoder Block
Decoder Block
Cross attention
Sinusoidal Positional Encoding
2x Conv1D + GELU
Log-mel spectrogram
Learned Positional Encoding
SOT
EN
TRAN-SCRIBE
0.0
The
quick
···
Tokens in multitask training format

Whisper from OpenAI

- https://huggingface.co/openai/whisper-large-v2



Safetensors

Model size | 1.54B params

Tensor type | F32

Hosted inference API

Automatic Speech Recognition

Examples

Browse for file or Record from browser or

Realtime speech recognition

Librispeech sample 2

0:00 / 0:14

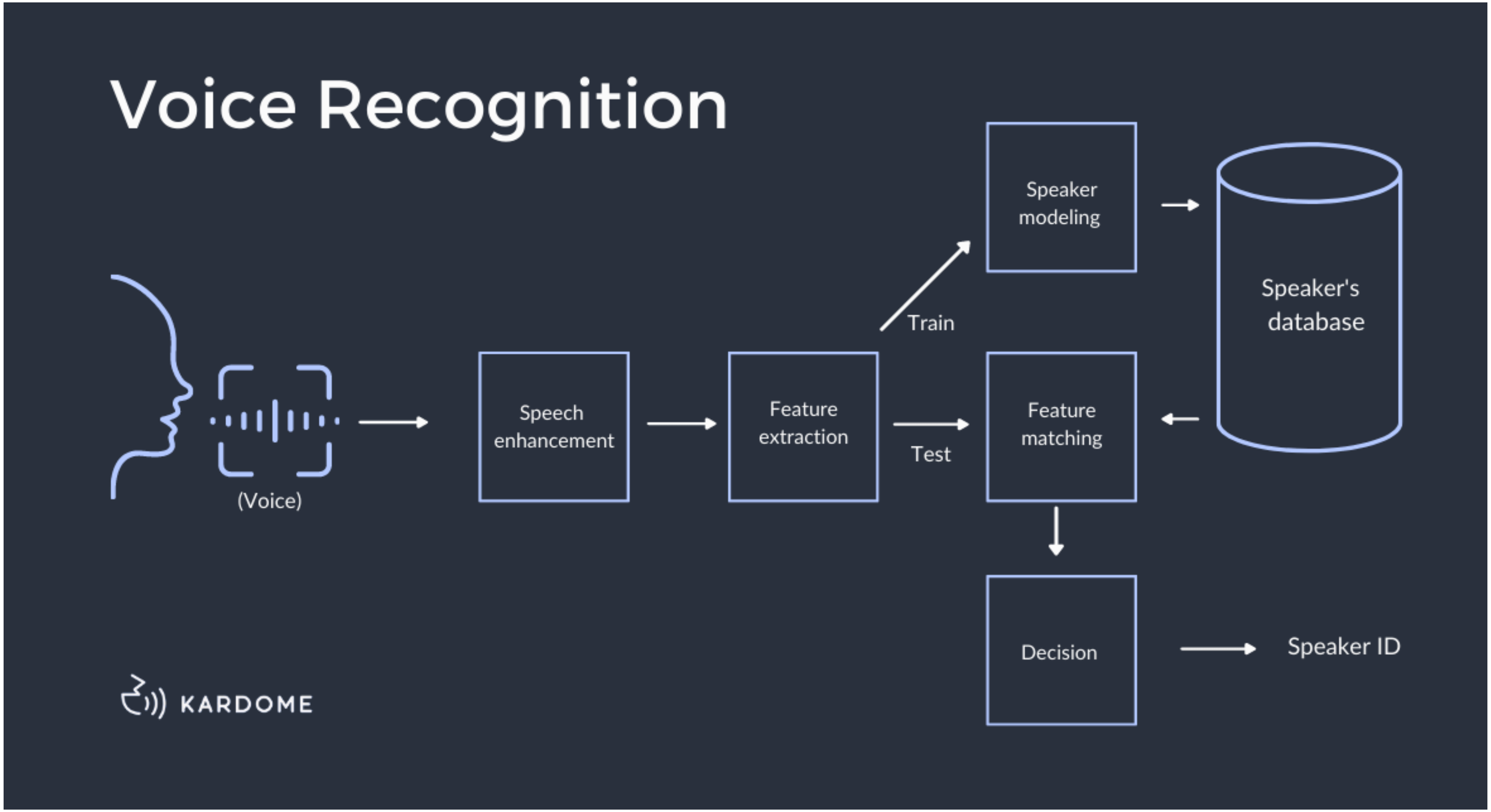
Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

Before he had time to answer, a much-encumbered Vera burst into the room with the question,—'I say, can I leave these here?' These were a small black pig and a lusty specimen of black-red game-cock.

Voice Recognition

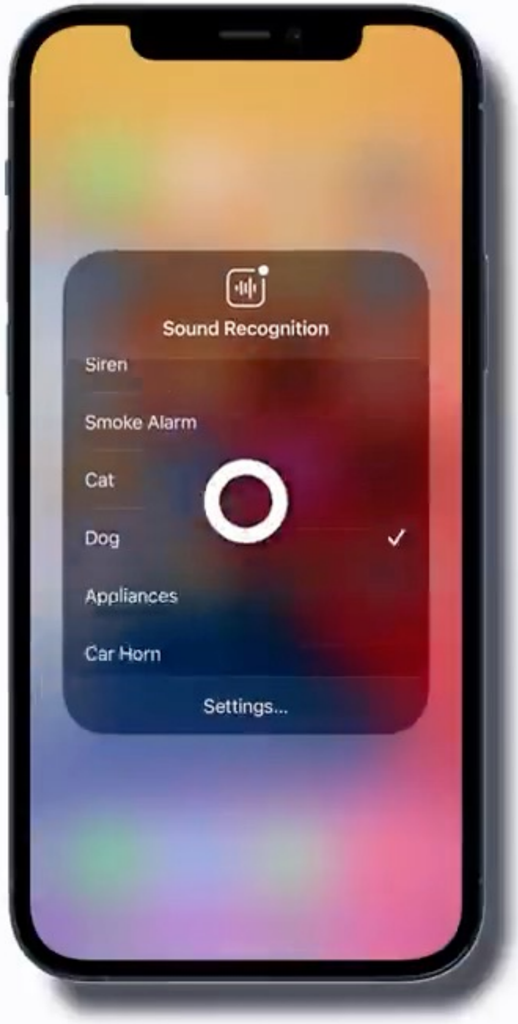
- Identify an individual user’s voice （Biometric）



Voice Recognition

Pay attention to the difference from speech recognition

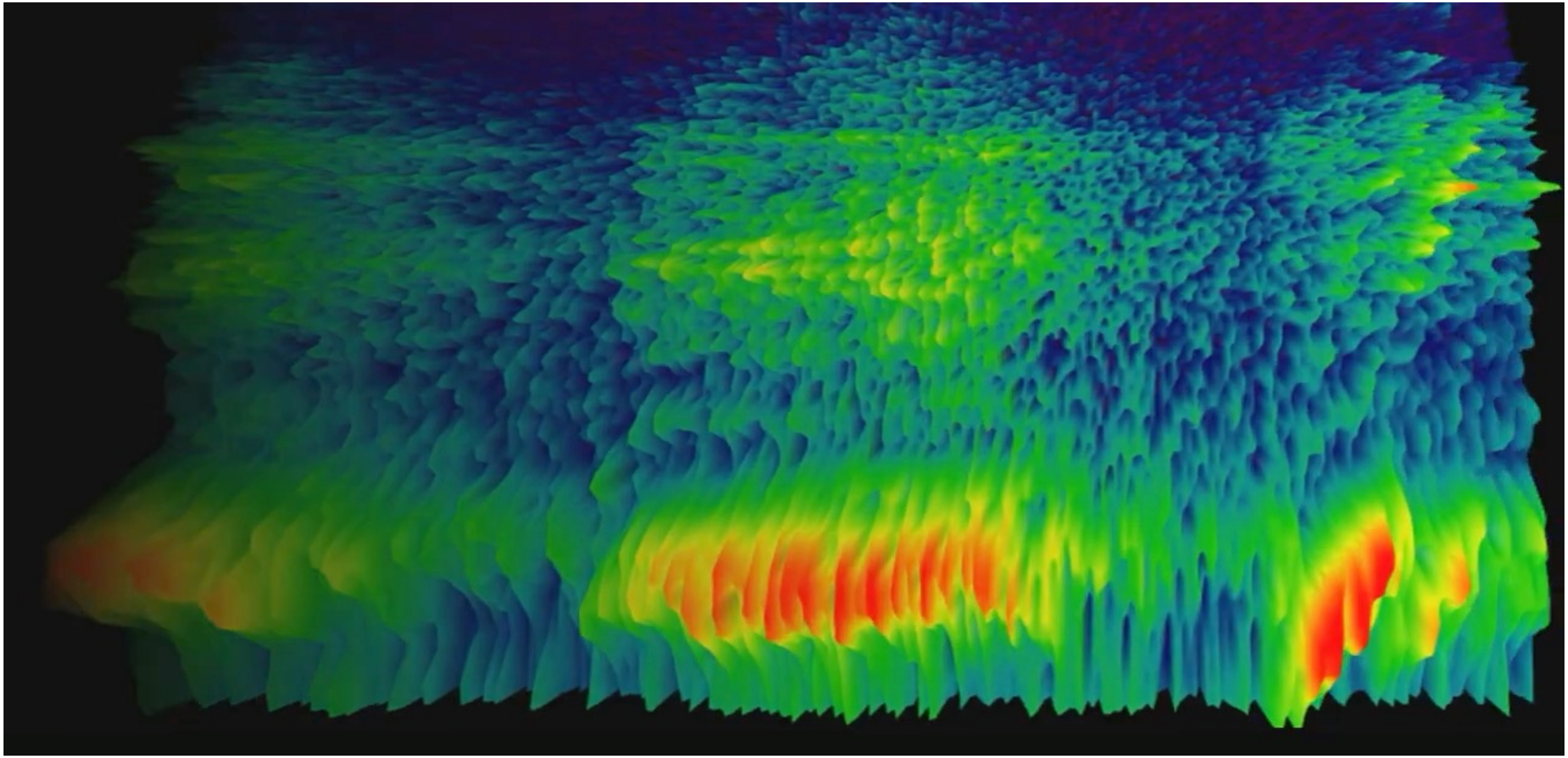
Recognize sounds in circumstances



Sound Recognition

- Siren
- Smoke Alarm
- Cat
- Dog ✓
- Appliances
- Car Horn

Using AI to listen to all of Earth's Species



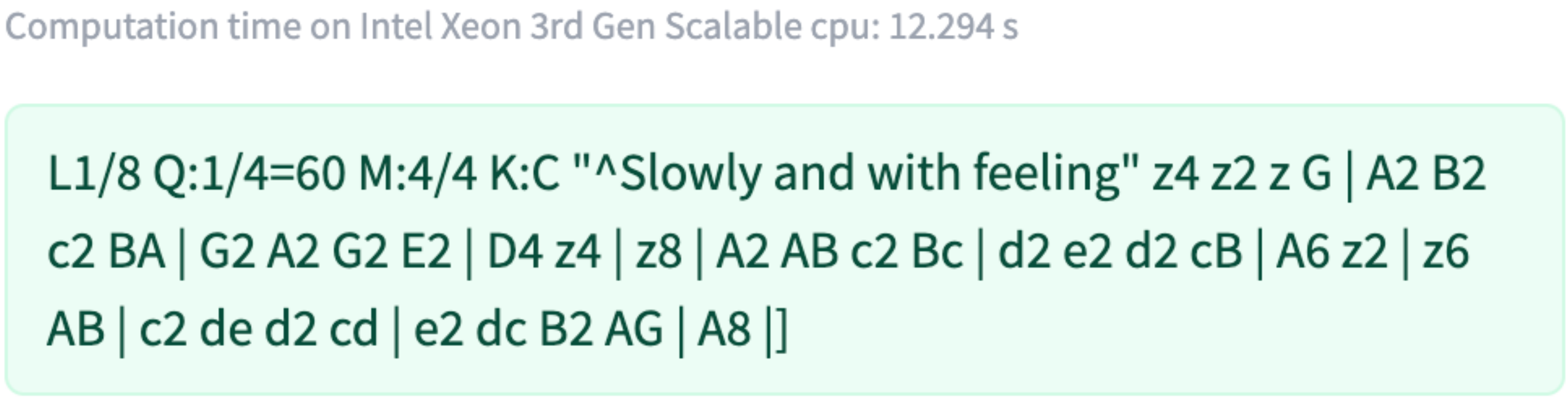
More resources

Using AI to listen to all of Earth's Species

Music Generation

Symbolic AI vs Audio AI systems

- Music AI generation into two broad camps: symbolic generation and audio generation
- A **symbolic** AI system generates the notes making up music
  - Exactly like a text generation model!
  - It requires a human to play the music notes, or additional music software to transform the notes into actual sound.



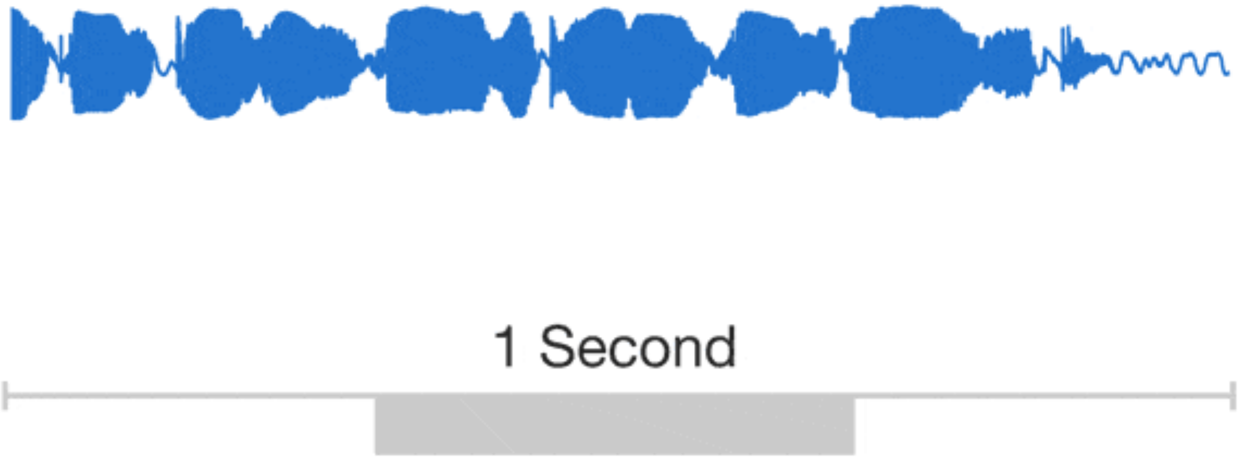
Computation time on Intel Xeon 3rd Gen Scalable cpu: 12.294 s

```
L1/8 Q:1/4=60 M:4/4 K:C "^Slowly and with feeling" z4 z2 z G | A2 B2
c2 BA | G2 A2 G2 E2 | D4 z4 | z8 | A2 AB c2 Bc | d2 e2 d2 cB | A6 z2 | z6
AB | c2 de d2 cd | e2 dc B2 AG | A8 |]
```

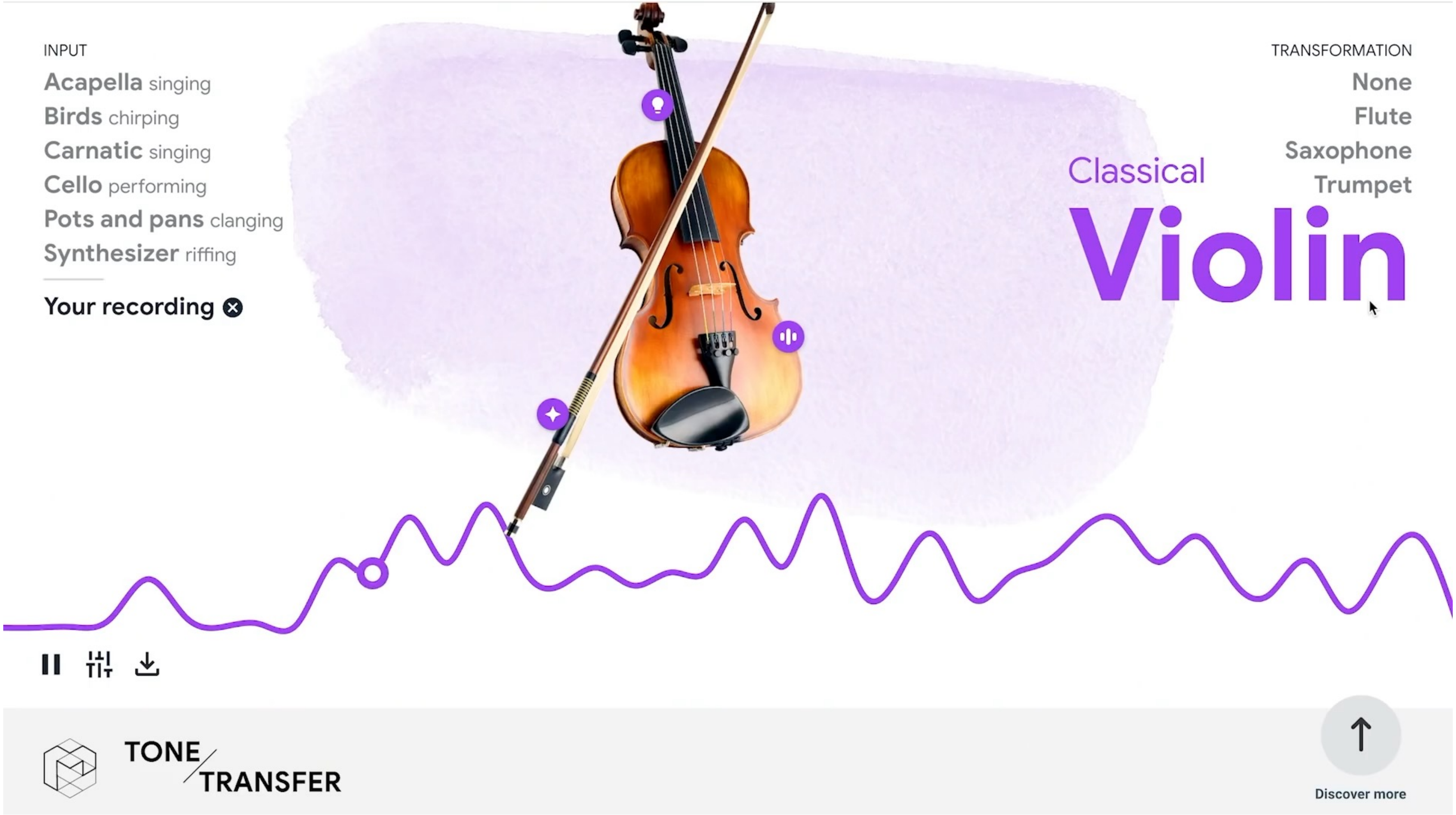
text-to-music: https://huggingface.co/sander-wood/text-to-music

Symbolic AI vs Audio AI systems

- An **audio** generation model synthesizes the waveform of the music directly!
  - *This is a very challenging for machine learning task!*
  - *A full 3-minute song in stereo puts us at over a billion samples. Keeping musical coherency across the first sample to the millionth sample is a difficult task.*



Tone Transfer



Source

MusicLM: Generating Music From Text

- Not open source, published 2023, join waitlist to use
- Dataset: The MusicCaps dataset contains **5,521** music examples, each of which is labeled with an English aspect list and a free text caption written by musicians.
- See Demo:
  - https://google-research.github.io/seanet/musiclm/examples/

Further questions

- What about music AI Copyright?
- Can a machine claim copyright if it is not a human?
- What does it mean to scrape data from artists who don’t want to be trained on?

Activity: Play with models

- MusicLM: Generating Music From Text: https://google-research.github.io/seanet/musiclm/examples/
- Magenta: https://magenta.tensorflow.org/demos/
- text-to-music: https://huggingface.co/sander-wood/text-to-music
- Neural Network Playground: https://playground.tensorflow.org/

Thank you~

Wan Fang

Southern University of Science and Technology