The Future Of Voice

The State Of Generative AI For Audio

Shift from AI Analysis to AI Synthesis

Analysis

Analyzing data generated by humans

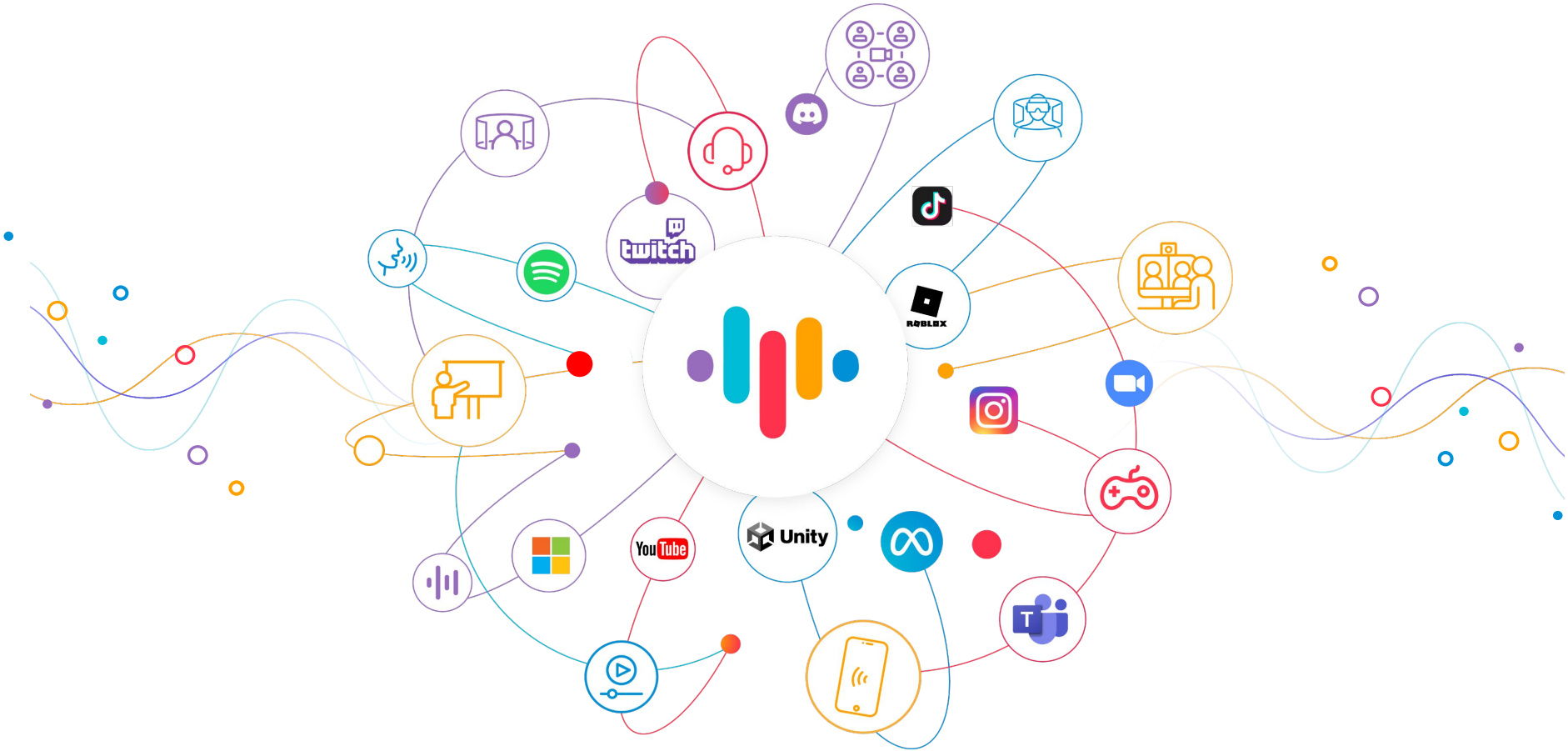
Used as an example for object recognition, text analysis and transcription

Synthesis

Generating new data to benefit humans and AI

Used to generate new Art (Dall–E), content such as blog posts & articles (LLM) and more...

Voice is Everywhere



Some Background

Why should we care about Voice?

Interacting with Machines

ChatGPT created a new way to interact with machines.

In order to do that, we need:

1. **Speech Recognition**: To understand what we are saying
2. **Speech Synthesis (TTS):** For us to understand what is machine is saying.

Content Creation

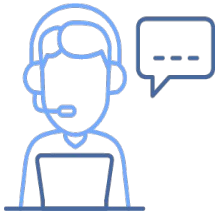
Music and audio are also being considered in the same category.

Videos

Games

New Type of Music

Type of Speech Applications

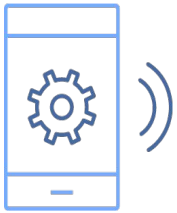


Speech Recognition

The ability for a computer/machine to analysis a human voice - usually converting human speech into text

Speech Profiling

Metadata information that can be extracted from speech. Speaker Recognition, Emotion Detection, Language Recognition and Age Estimation



Speech Synthesis

The ability for a computer/machine to generate human voice

Types of Speech and Audio Generation

What I will cover

1. Audio and Speech Generation / Generative AI for Speech

2. LLM for Speech

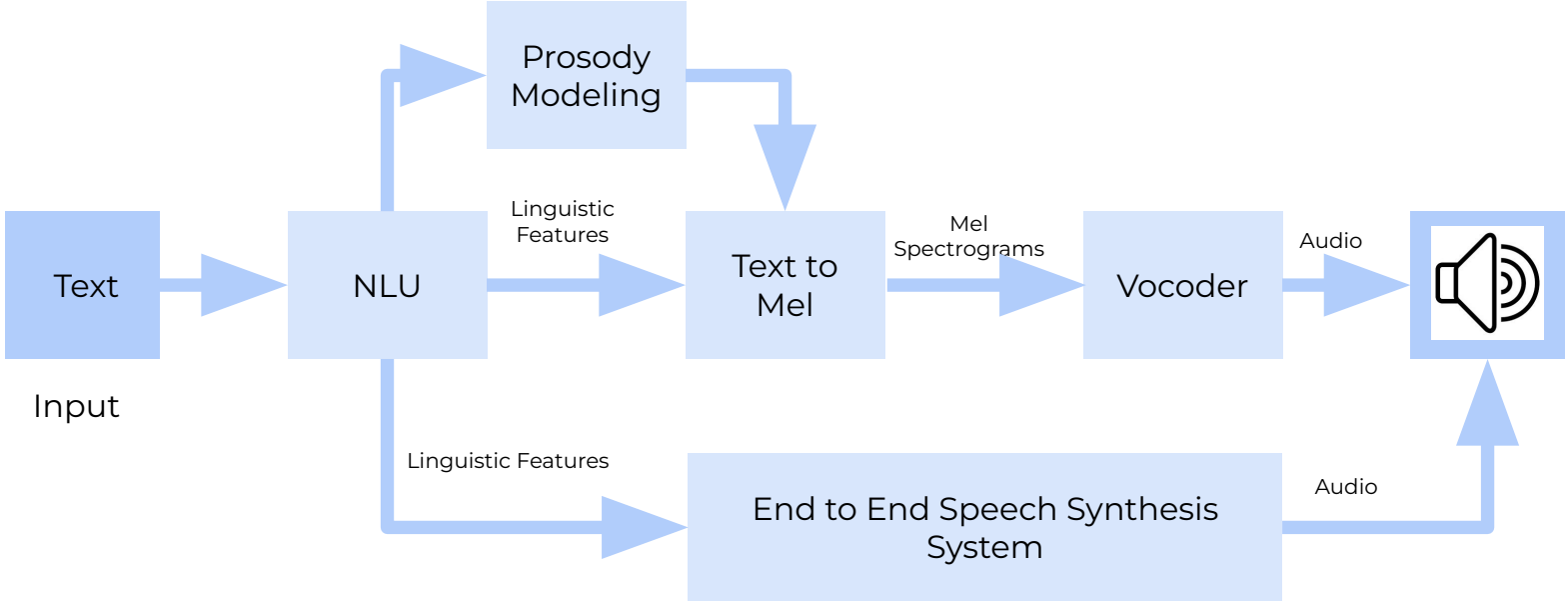
3. Ethics and Safety

Generative AI for Speech

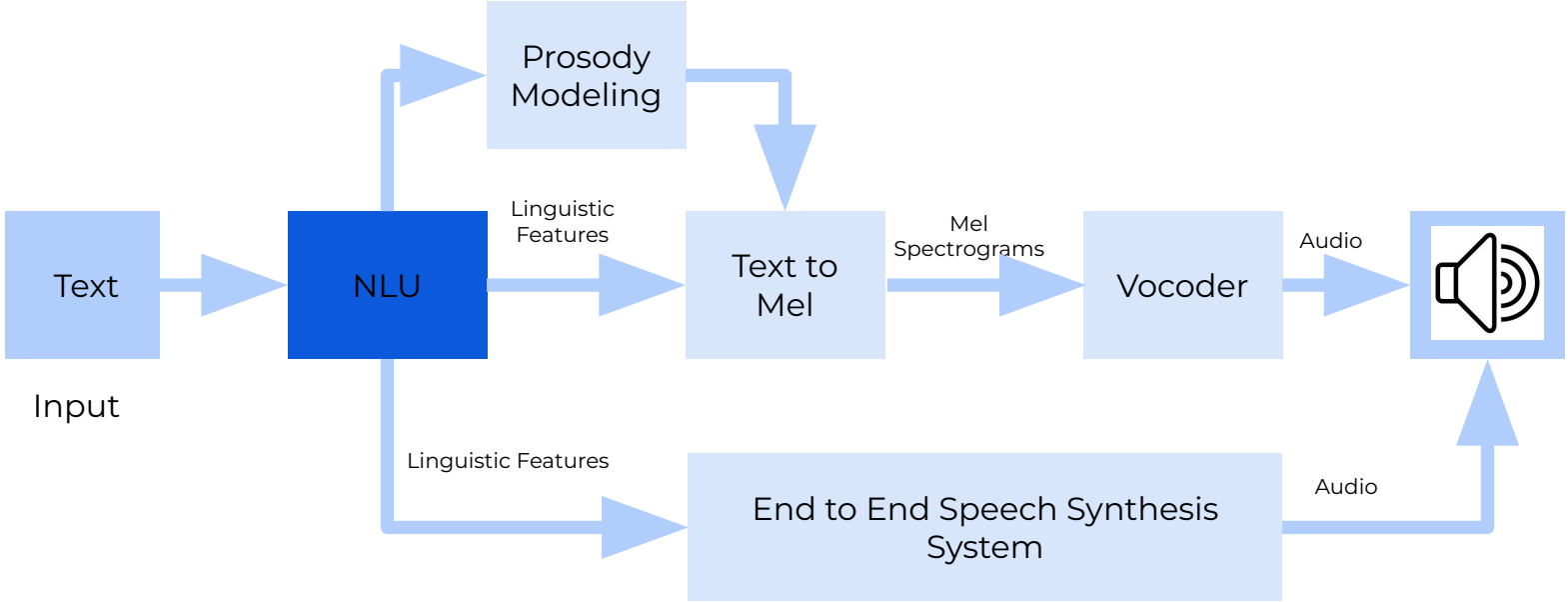
Different Type of Generation

Input	Output	Example
Text	Speech	Text to Speech
Speech	Speech	Voice Conversion
Text + Speech	Speech	Style Transfer
Text	Music	Music Generation
Speech	Music + Speech	Singing Synthesis

TTS Block Diagram



TTS NLU

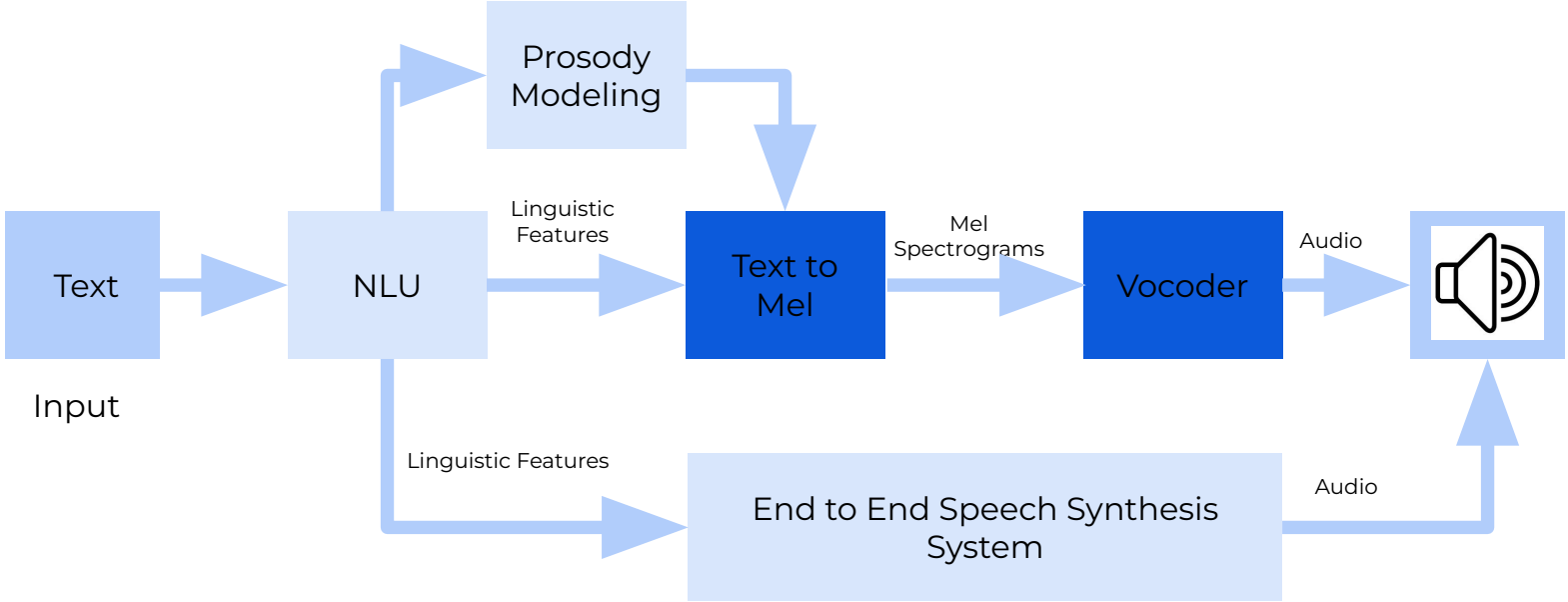

Prosody Modeling
Text
NLU
Linguistic Features
Text to Mel
Mel Spectrograms
Vocoder
Audio
Input
Linguistic Features
End to End Speech Synthesis System
Audio

TTS NLU

Converting the text into a format TTS can process

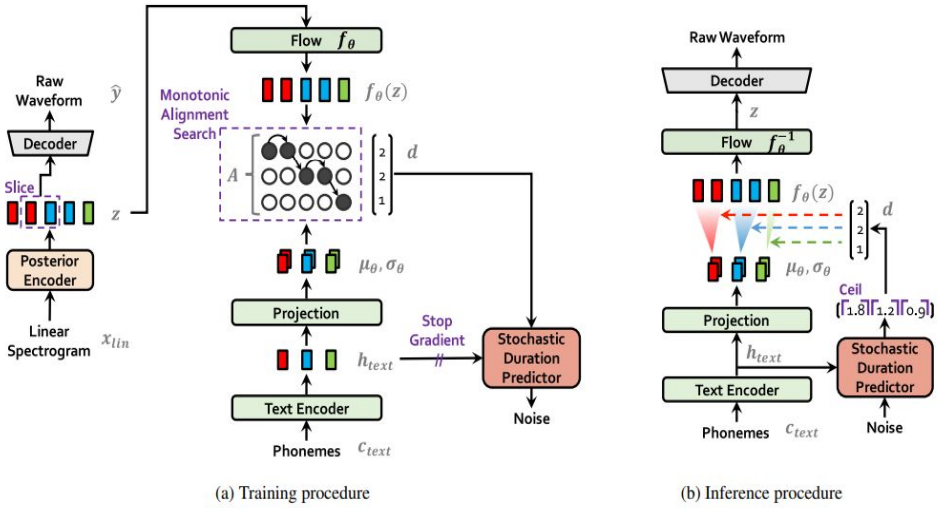
- My email is abc@gmail.com My e mail is a b c at g mail dot com
- My father was born in 1939 (“nineteen thirty-nine”)
- Please press 1939 (“one-nine-three-nine”)
- This computer costs $1939 (“one thousand nine hundred thirty-nine”)

TTS Speech Generation



TTS Speech Generation

- **LingiusticEncoder**: Phoneme encoder
- **PosteriorEncoder** - Encodes audio into hidden representation
- **Duration predictor and Alignment search:** Uses dynamic programming to align outputs of prosody encoder and linguistic encoder (compare to attention in Taco2).
- The duration predictor learns phoneme durations based on the alignment and is used at inference time
- Integrated GAN-based vocoder trained jointly with the rest of the model



Source: VITS - Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech
Link: https://arxiv.org/pdf/2106.06103.pdf

State of the art of TTS

What is it good for?

- Avatar Dubbing , Podcast
- Voice inpainting , editing
- Video dubbing with lip sync

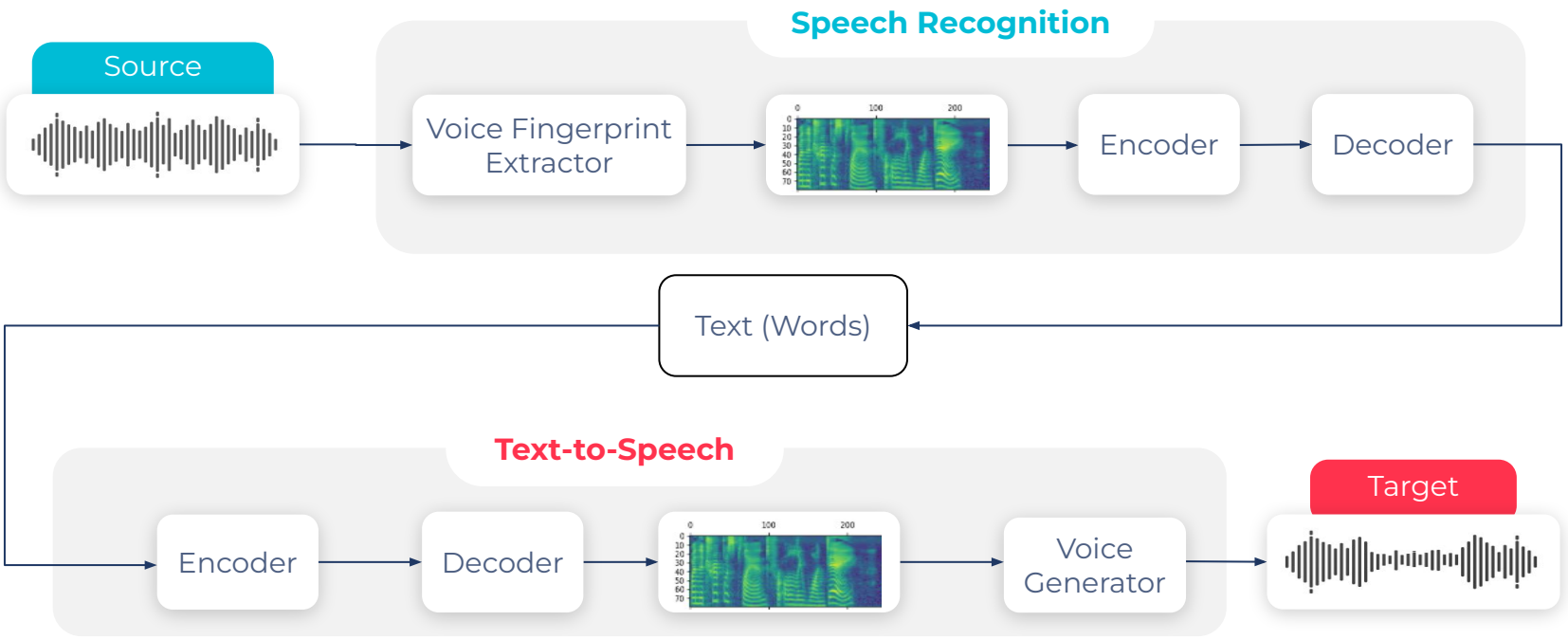
What needs to be solved?

- Conversational Speech
- High expressiveness
- Emotion controllability

Speech to Speech

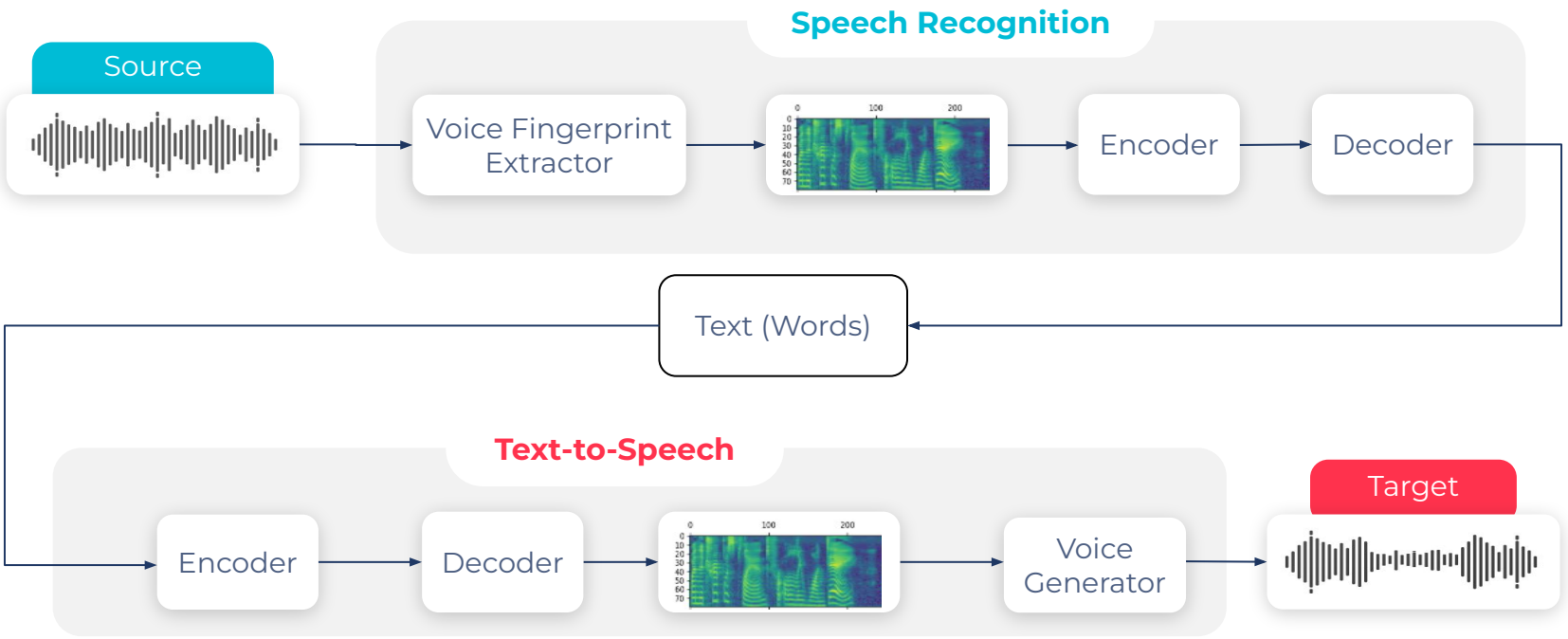
Tech

The Classical Approach



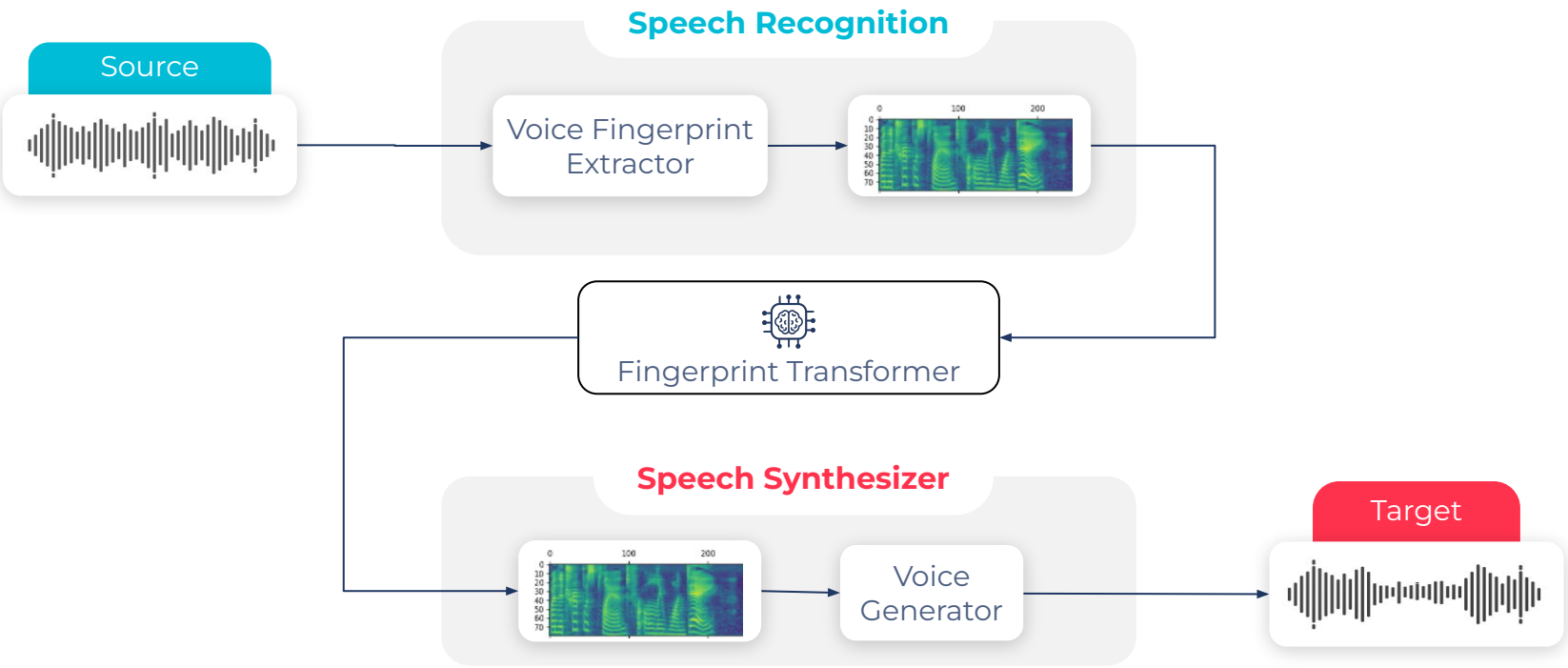
Tech

The Classical Approach

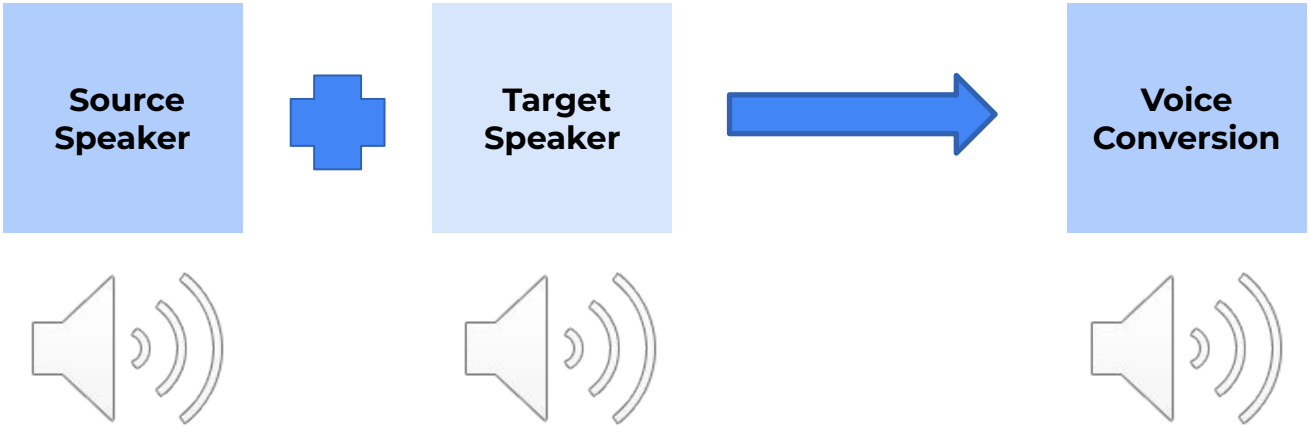


Tech

Direct Speech-to-Speech Conversion



Example – Voice Conversion



Example – Meta VoiceBox

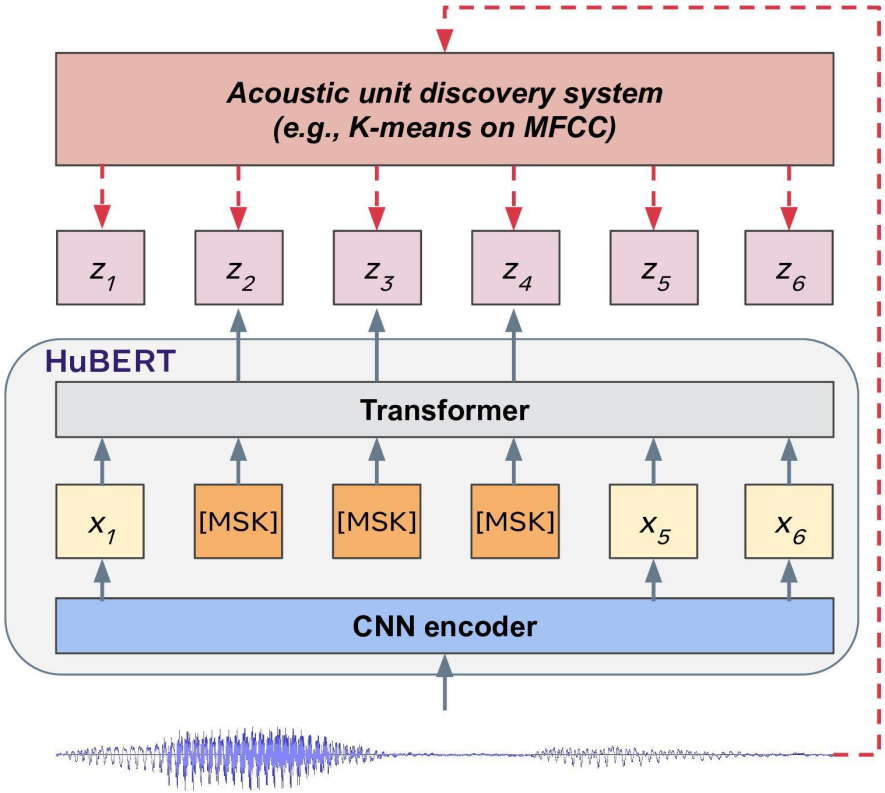
Introducing Voicebox:

https://ai.meta.com/blog/voicebox-generative-ai-model-speech/

- **Multi-Speaker TTS**: Ability to adapt to new speakers.
- Multi-Lingual Support: Capability to move speakers to other languages.
- **Style Transfer**: Conversion to new target speakers based on a new speaker and a text sample.
- **Editing and Noise Removal**: Enhancing the audio quality by editing and removing noise.

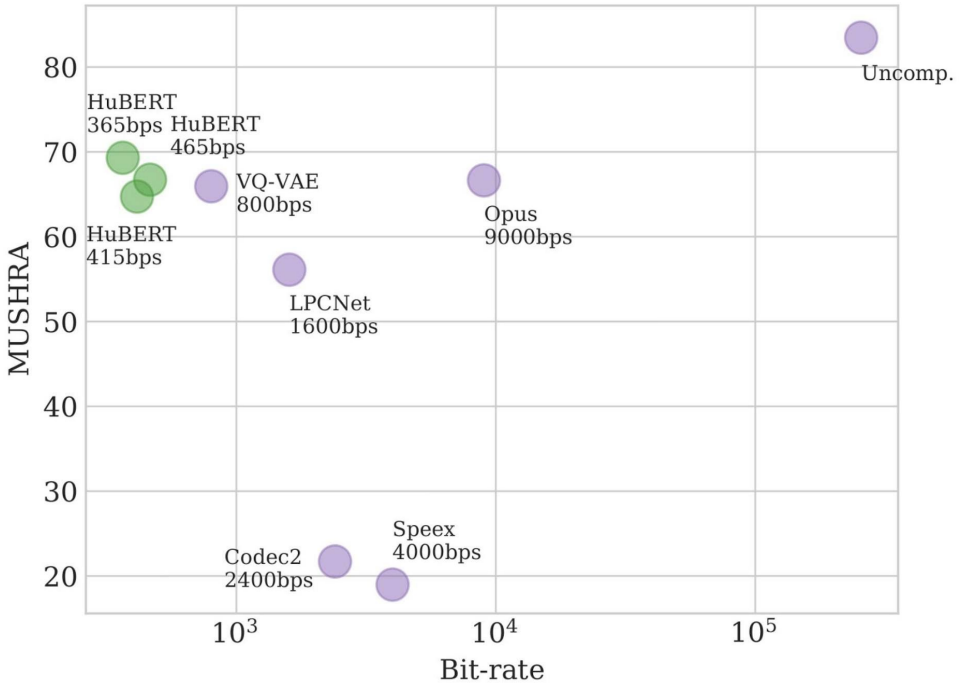
LLM in Speech

HuBERT – BERT Like LM for Speech



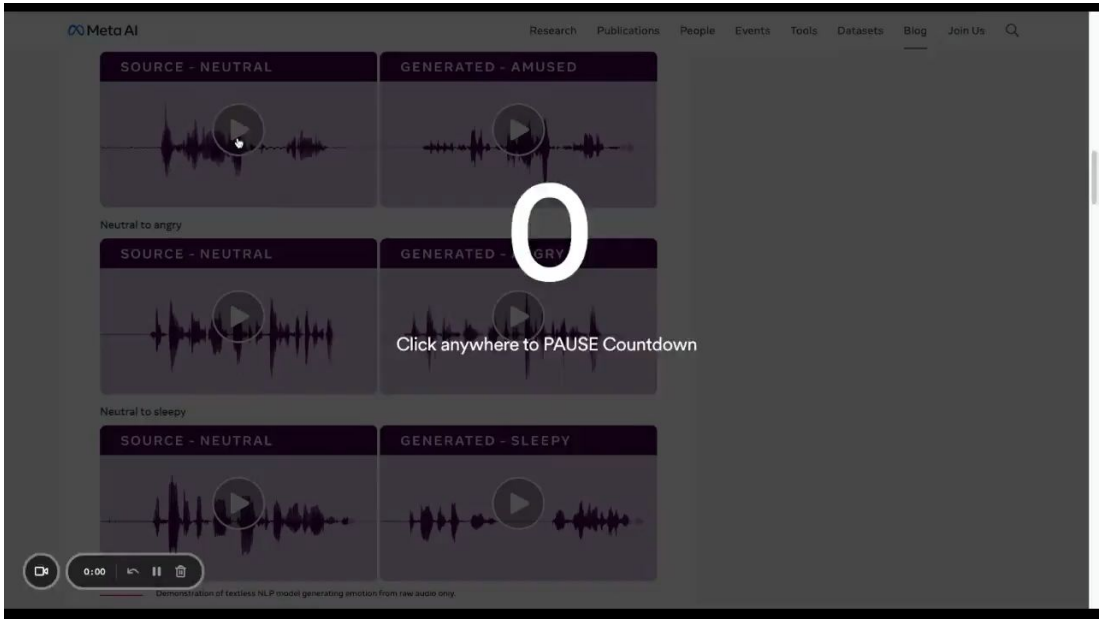
Meta AI Blog: HuBERT: Self-supervised representation learning for speech recognition, generation, and compression

Applications – Voice Compression



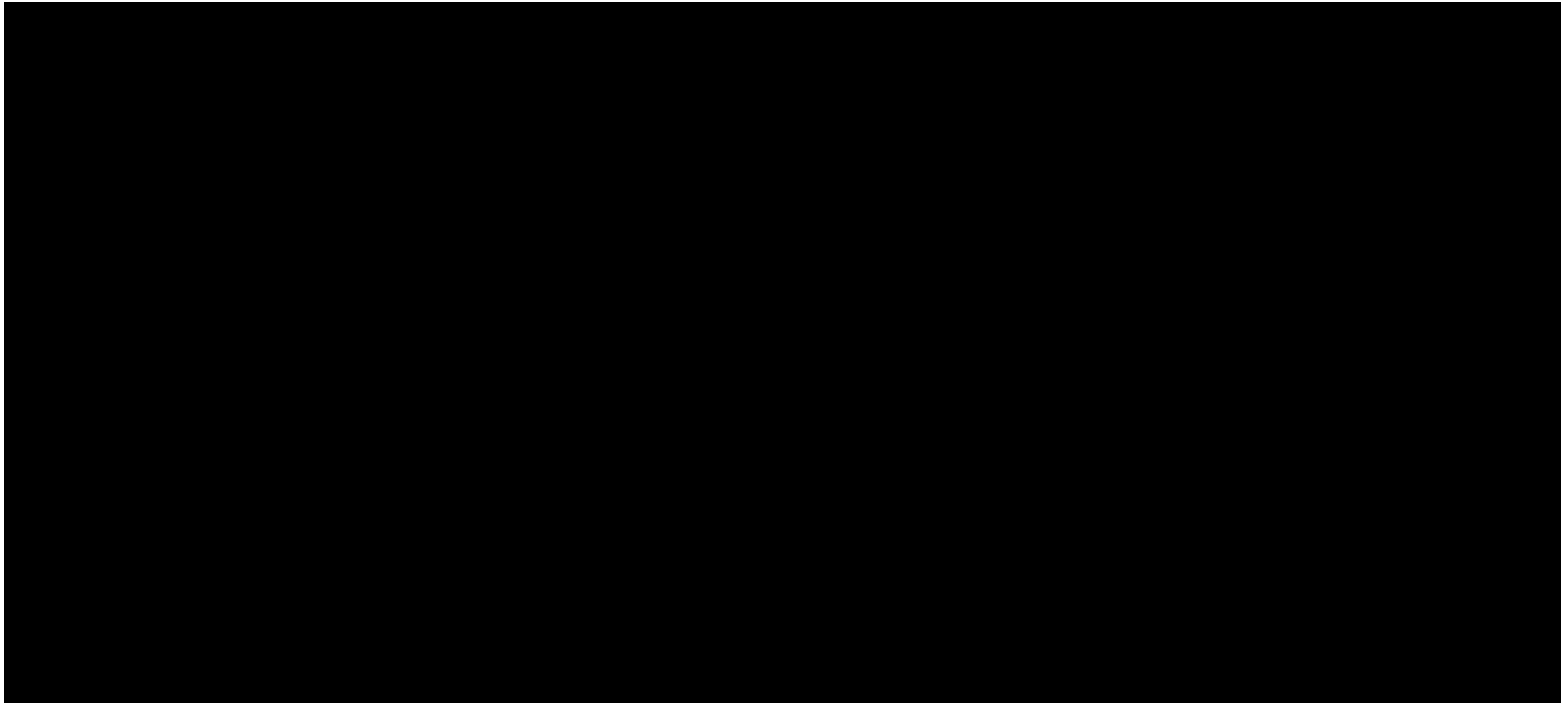
Meta AI Blog: HuBERT: Self-supervised representation learning for speech recognition, generation, and compression

Applications – Emotion Transfer



Meta AI Blog: Generating chit-chat including laughs, yawns, 'ums,' & other nonverbal cues from raw audio

Example – Google MusicLM and Meta AudioCraft



Meta AI Blog: Introducing AudioCraft: A Generative AI Tool For Audio and Music

Ethics and Safety

Voice Cloning and Audio Deep Fakes

As we can adapt TTS system, we can use it for Voice Cloning and Audio Deep Fakes

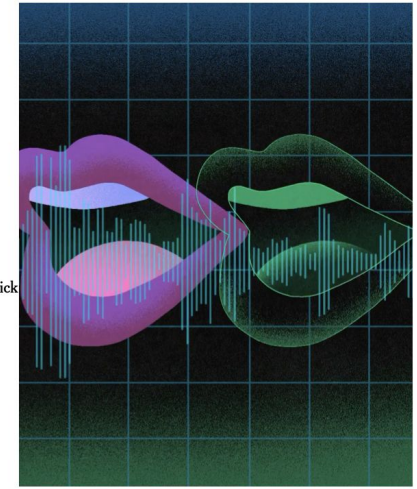
1. VALL – E : Can clone your voice in only 5s
2. VALL – E X : Does not have to be in English

Don’t believe your ears: voice deepfakes

Audio deepfakes that can mimic anyone’s voice are already being used for multi-million dollar scams. How are deepfakes made and can you protect yourself from falling victim?

Voice Deepfakes Are Coming for Your Bank Balance

Artificial intelligence tools have given scammers a potent weapon for trying to trick people into sending them money.



Audio Fake Detection

- Understanding the risk factors: Enterprise vs. Consumer.
- Understanding the attack vectors: Which voice cloning software will be used?
- Generating suitable datasets.
- Building proper classifiers and detectors:

1. General purpose classifiers: Synthetic vs. Real.
2. User Specific Model

Let’s talk some more

info@meaning.team