Audio Generation

Bo Kang, Thomas Demeester, Tijl De Bie

Outline

- Introduction
- An overview of models
- Demo: Finetune a TTS model
- References

Outline

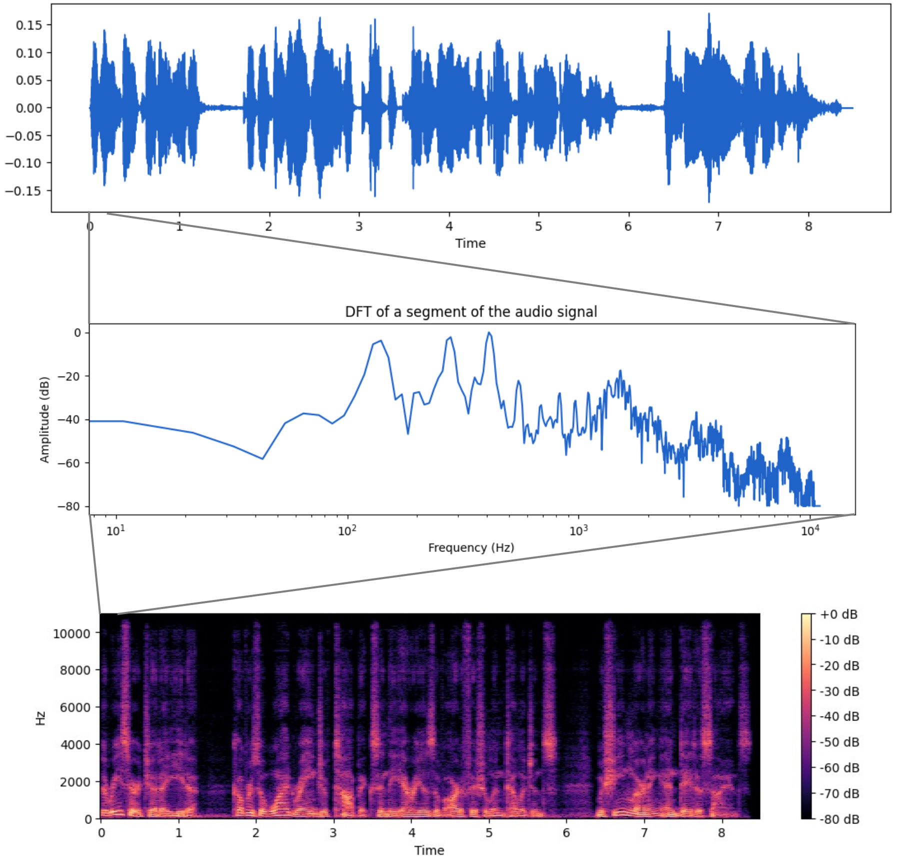
- Introduction
- An overview of models
- Demo: Finetune a TTS model
- References

Introduction

- Generate or manipulate audio with neural network models
- Applications
  - Classification
  - Speech recognition
  - Text to speech generation
  - Voice Cloning
  - Music generation
  - …
- Tools
  - Commercial: ElevenLabs[1], OpenAI TTS[2]
  - Open source: HF transformers[3], Tortoise[4], Bark[5], Coqui(XTTS)[6]

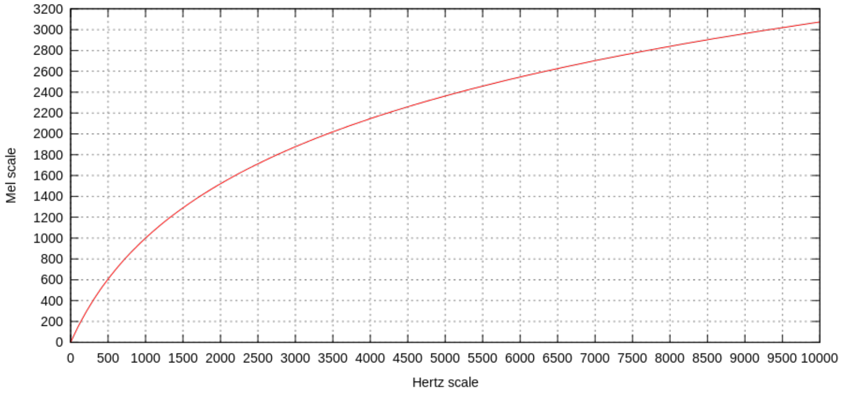
Basic Data Concepts

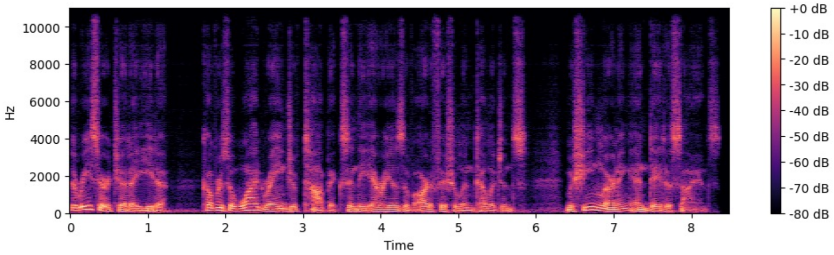
- Waveform
- Fourier transform
- Spectrogram
- Log-mel spectrogram

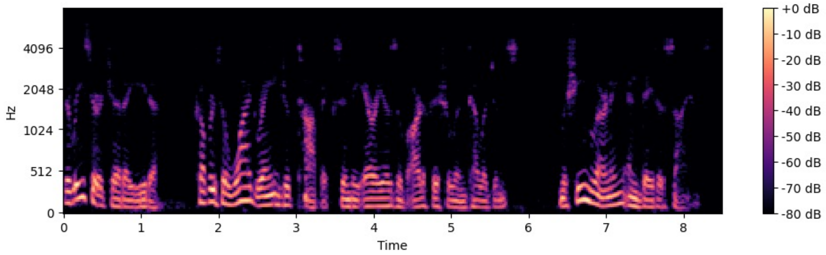


Basic Data Concepts

- Waveform
- Fourier transform
- Spectrogram
- Log-mel spectrogram

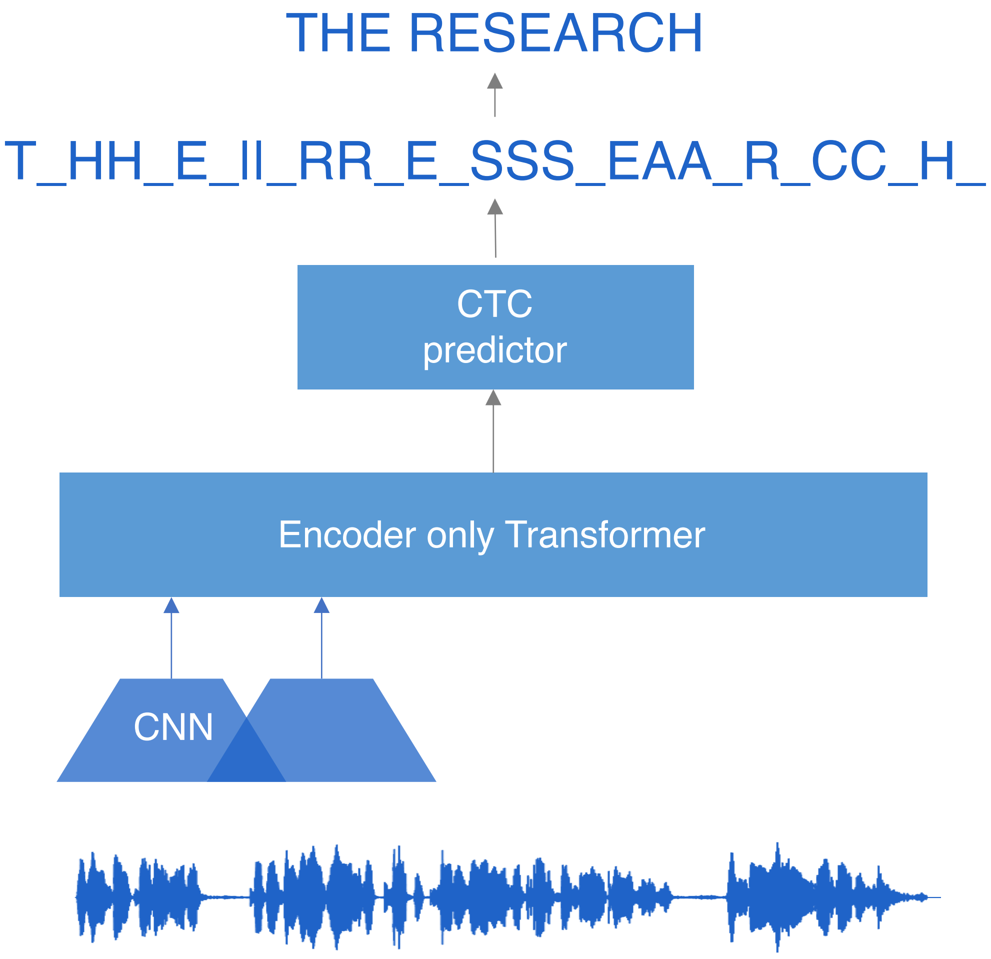






Basic Architectures for Speech Recognition

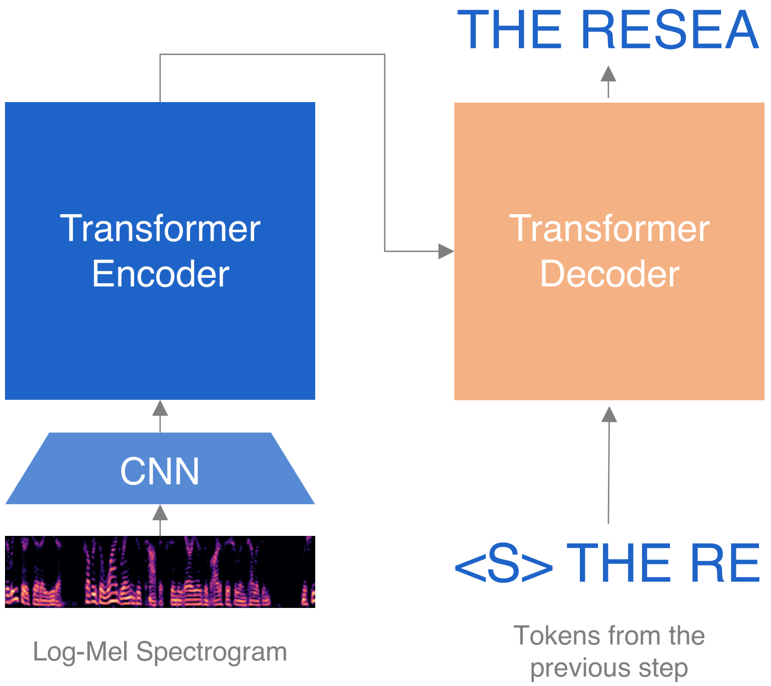
- CTC (Connectionist Temporal Classification)
    - Encoder only transformer model
    - Two phase generation



Basic Architectures for Speech Recognition

- CTC (Connectionist Temporal Classification)
  - Encoder only transformer model
  - Two phase generation

- Seq2Seq
  - Encoder-decoder transformer model



Outline

- Introduction
- An overview of models
- Demo: Finetune a TTS model
- References

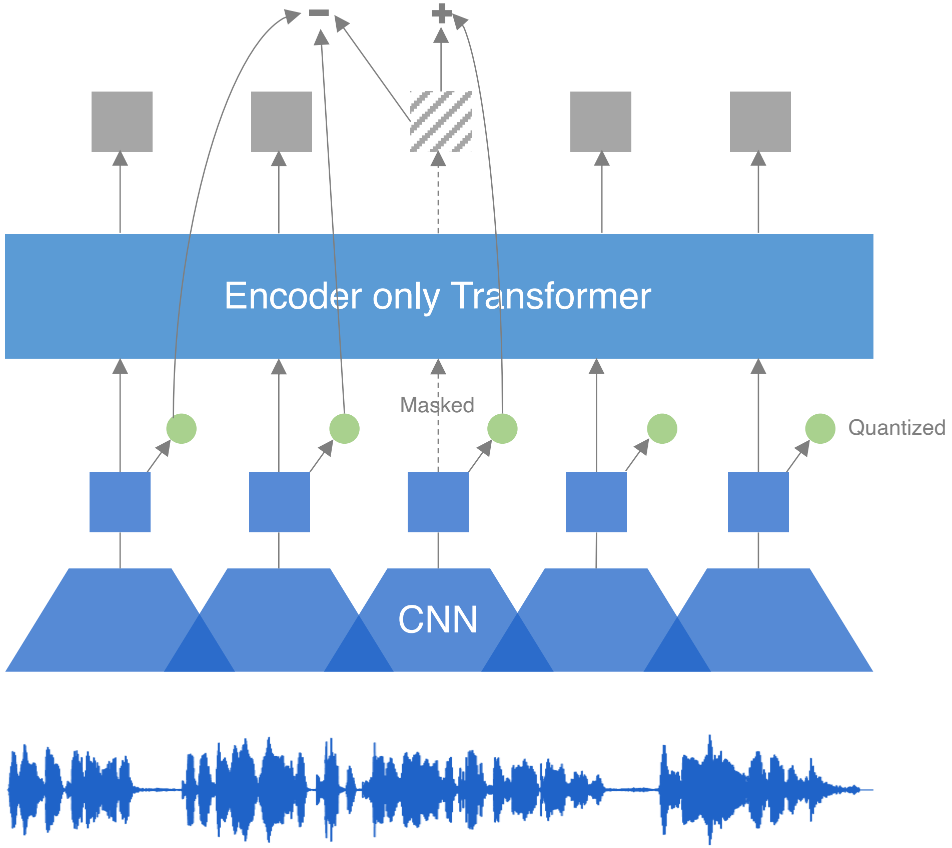
An Overview

Classification and Recognition

wav2vec2, 2020	HuBERT, 2021	SpeechT5, 2021	Whisper 2022	Whisper v3 2023
HiFi-GAN, 2020	UnivNet, 2021	Soundstream, 2021	EnCodec, 2021	
Audio LLM 2022	VALLE 2023	Bark, 2023	Tortoise, 2023	XTTS, 2023
				OpenAI TTS, 2023

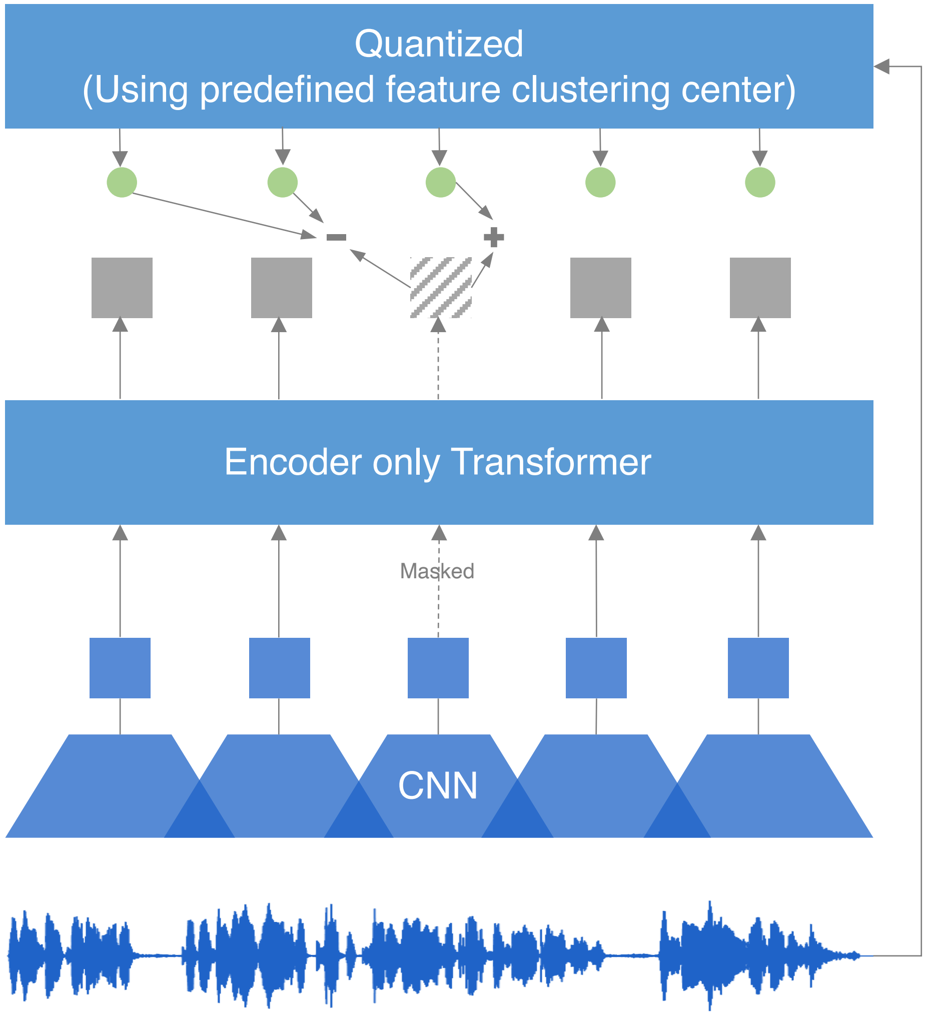
Wav2Vec2

- Idea
  - 1D CNN extract feature
  - Encoder-only transformer represents contexts
  - Self-supervised pretraining from context embedding and quantized embedding pairs
  - Finetune with CTC loss



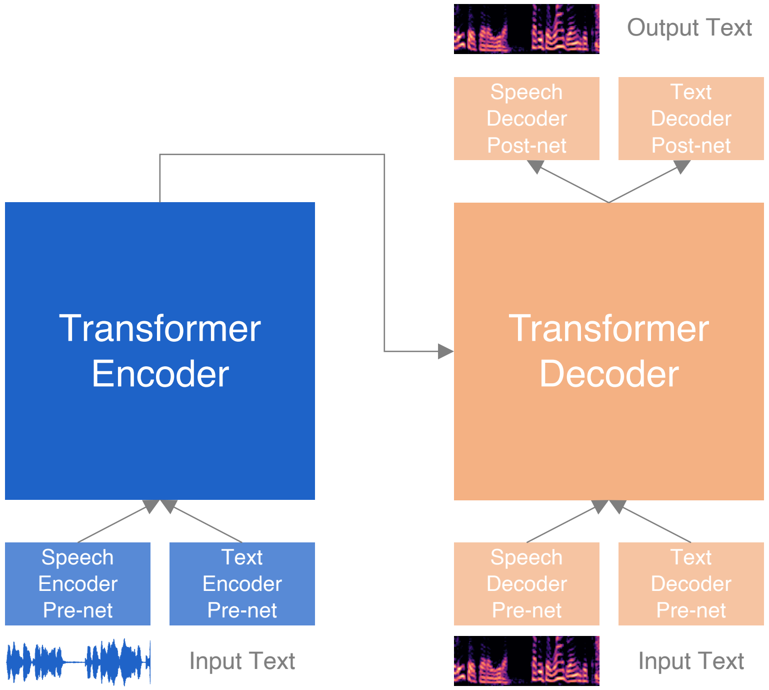
HuBERT

- Idea
  - Similar to Wave2Vec
  - Except the quantization is according to predefined features



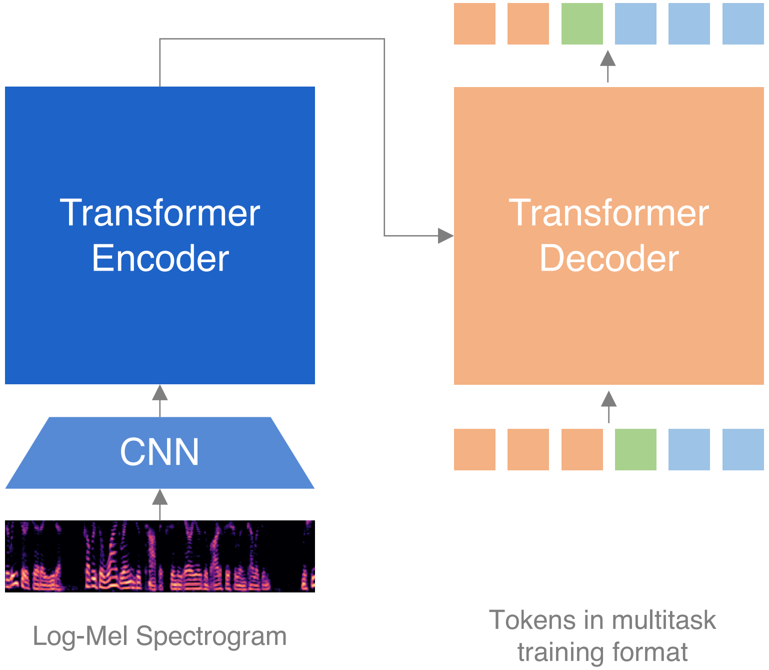
SpeechT5

- Idea
  - Standard encoder-decoder transformer architecture
  - Multitask and Multimodal enabled by pre-net/post-net
  - Within/cross modal pretraining
    - Hubert like self-supervised pretraining
    - Decoder speech reconstruction
    - BART like masked text token prediction
    - Share quantization codebook between modals at decoder level

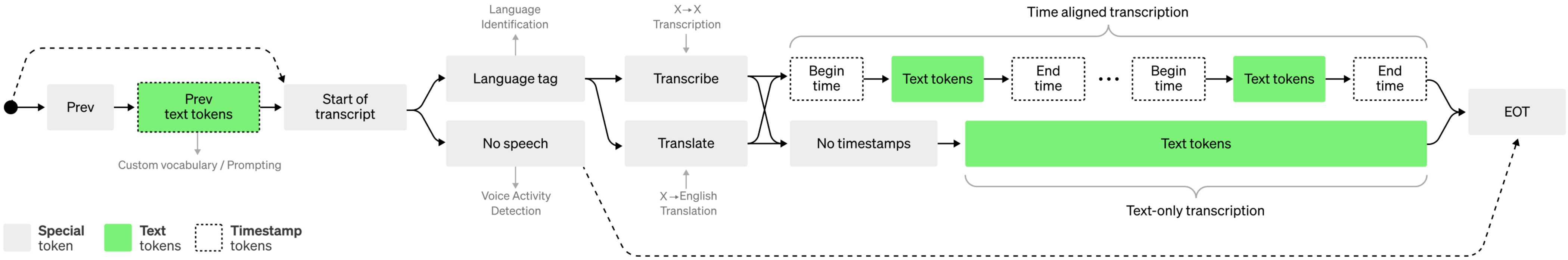


Whisper

- Idea
  - Off-the-shelf encoder decoder transformer architecture
  - Scale with large training dataset
  - Multitask enabled by decoder input format

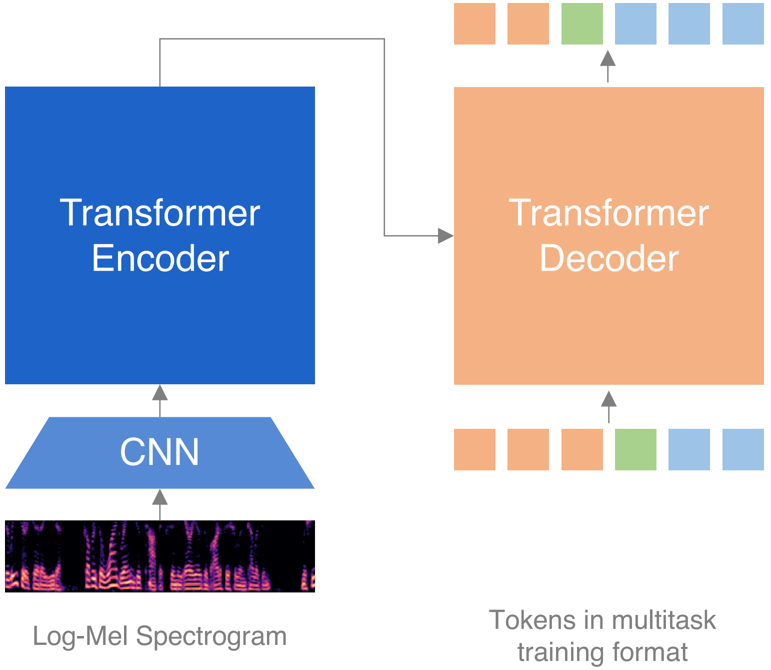


Whisper



- Idea
  - Off-the-shelf encoder decoder transformer architecture
  - Scale with large training dataset
  - Multitask enabled by decoder input format
    - Speech detection
    - Speech transcript
    - Translation
    - Language identification

- Whisper v3, more fine-grained Mel-scale discretization, more data, support more languages



An Overview

wav2vec2, 2020

HuBERT, 2021

SpeechT5, 2021

Whisper 2022

Whisper v3 2023

Waveform encode decode

HiFi-GAN, 2020

UnivNet, 2021

Soundstream, 2021

EnCodec, 2021

Audio LLM 2022

VALLE 2023

Bark, 2023

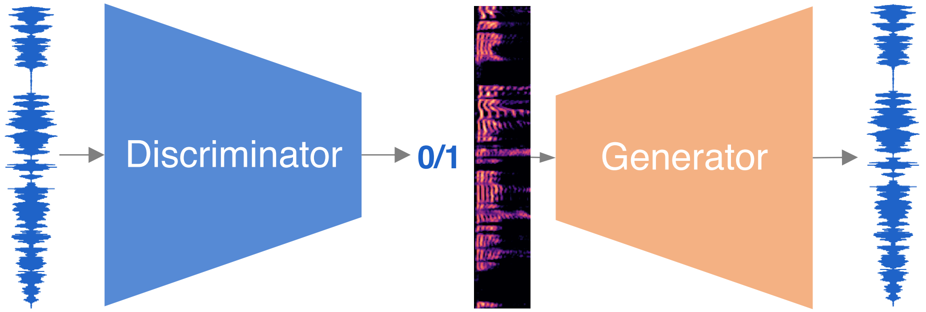
Tortoise, 2023

XTTS, 2023

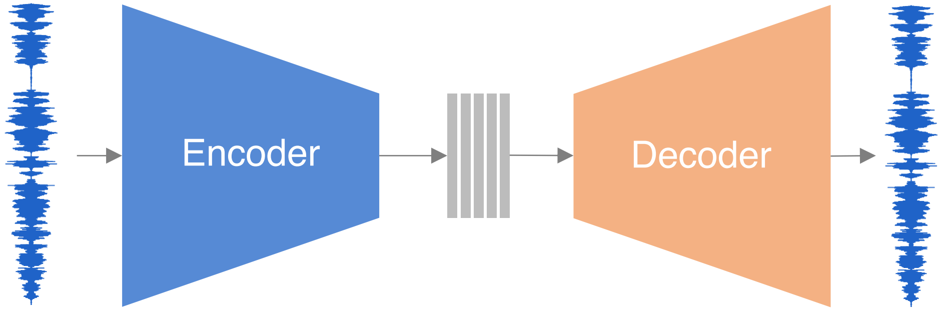
OpenAI TTS, 2023

Neural Vocoder and Audio Codec Handles Model In/output

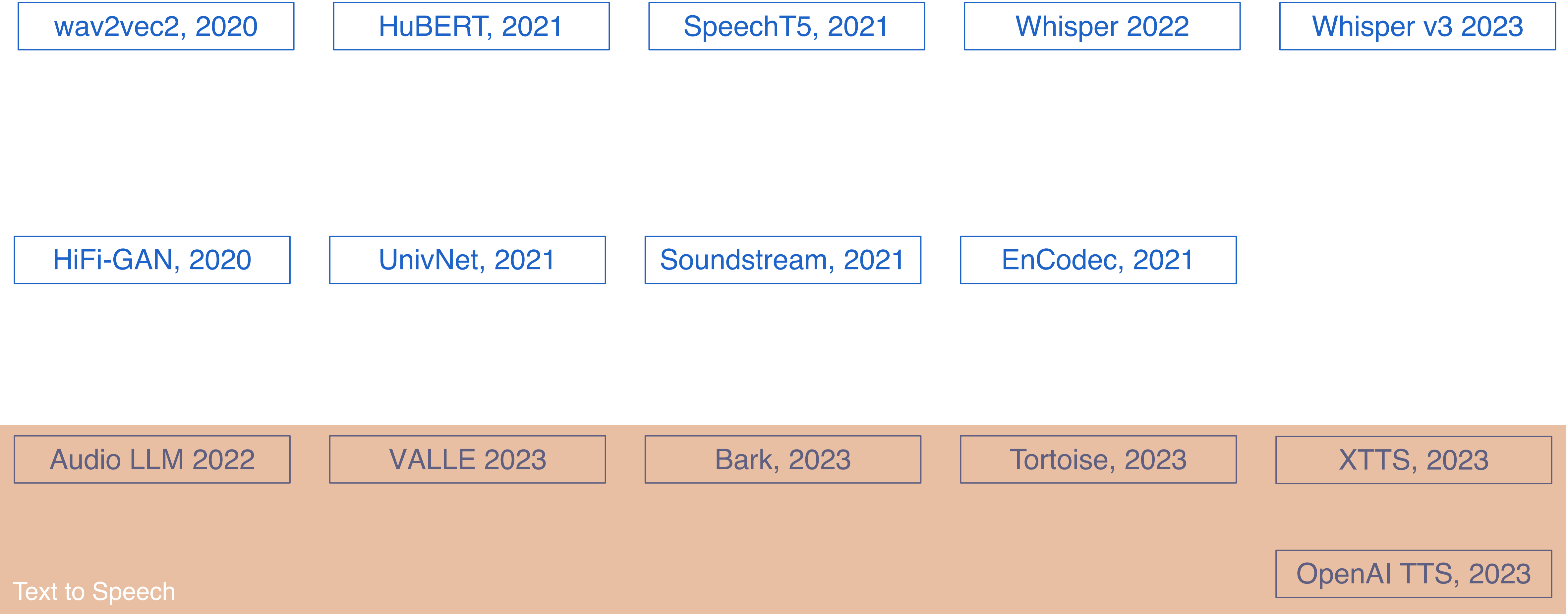
- Neural Vocoder takes mel-spectrograms as input and generates waveforms (HiFi-GAN, UnivNet)



- Audio Codec encodes (quantize) waveform into codes (acoustic tokens), decodes codes back to waveform (Soundstream, Encodec)



An Overview



wav2vec2, 2020

HuBERT, 2021

SpeechT5, 2021

Whisper 2022

Whisper v3 2023

HiFi-GAN, 2020

UnivNet, 2021

Soundstream, 2021

EnCodec, 2021

Audio LLM 2022

VALLE 2023

Bark, 2023

Tortoise, 2023

XTTS, 2023

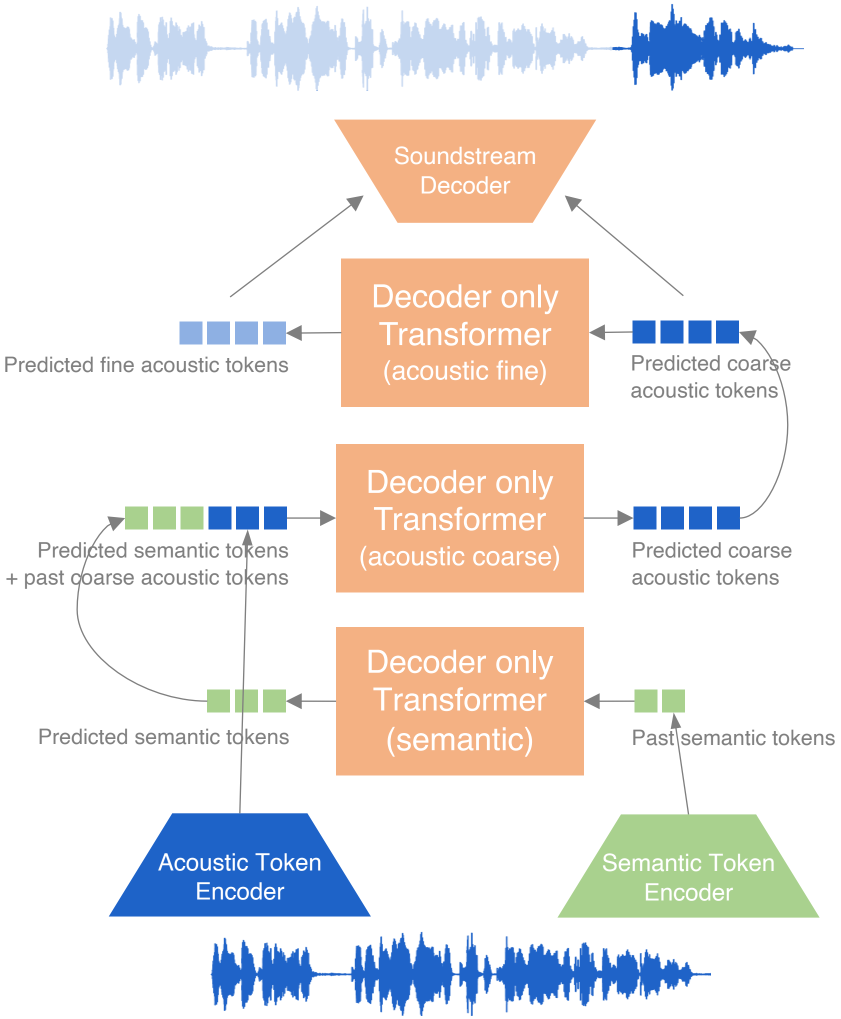
OpenAI TTS, 2023

Text to Speech

AudioLM

- Idea
  - Semantic token encoder
  - Acoustic token encoder
  - Decoder only transformer model for predicting next token
  - Cascading next acoustic token prediction

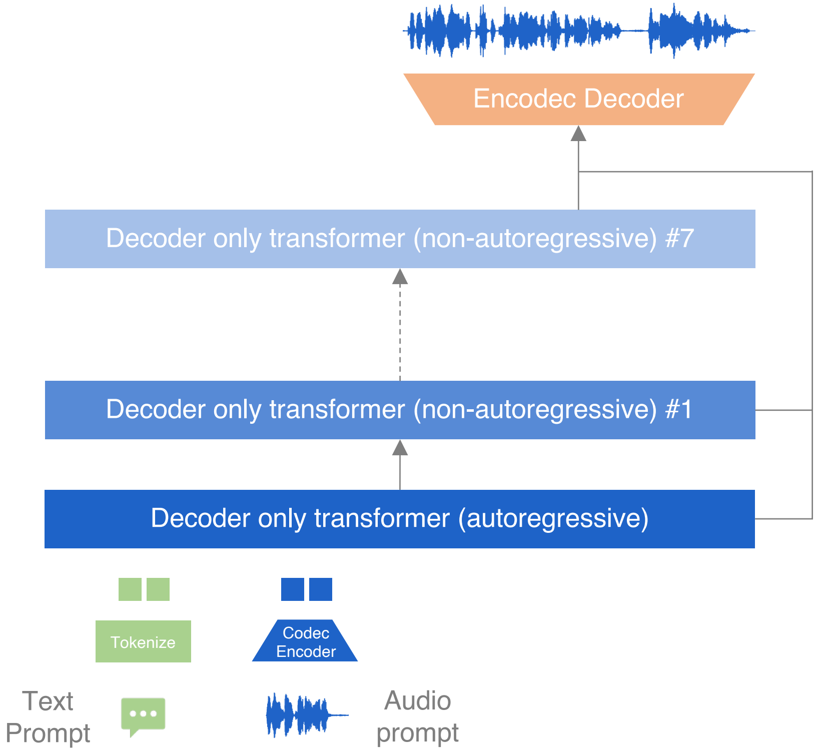
- Music Continuation Example
  - Prompt: 🔊
  - Original: 🔊
  - Continuation: 🔊


Soundstream Decoder
Decoder only Transformer (acoustic fine)
Predicted fine acoustic tokens
Predicted coarse acoustic tokens
Decoder only Transformer (acoustic coarse)
Predicted semantic tokens + past coarse acoustic tokens
Predicted coarse acoustic tokens
Decoder only Transformer (semantic)
Predicted semantic tokens
Past semantic tokens
Acoustic Token Encoder
Semantic Token Encoder

VALLE

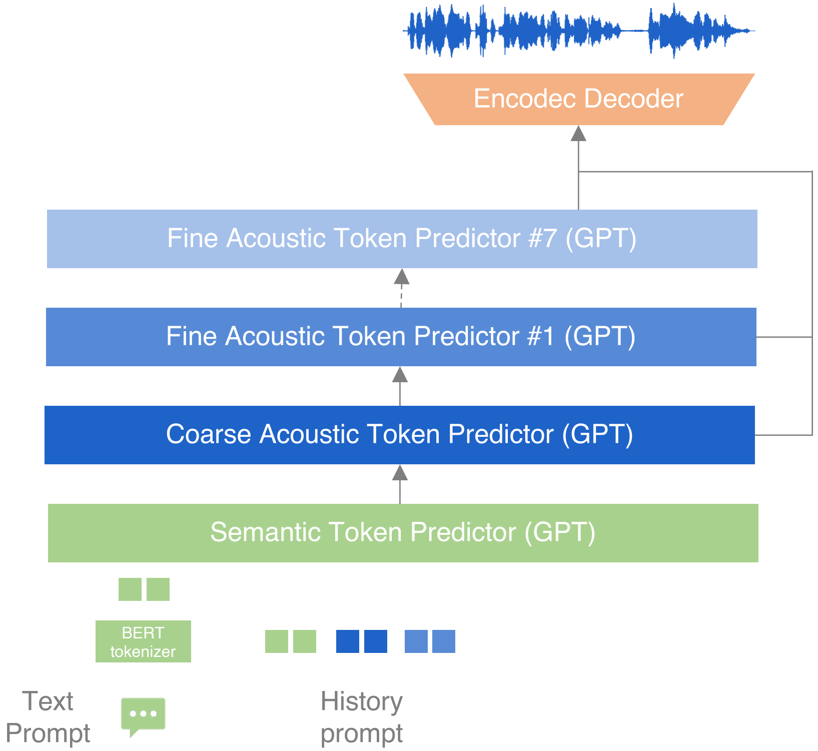
- Idea
  - Audio prompt for voice cloning, use Vocoder for acoustic tokenization
  - Cascading of acoustic token prediction

- Official Example
  - Text prompt: “and lay me down in thy cold bed and leave my shining lot.”
  - Audio Prompt: 🔊
  - Original: 🔊
  - TTS: 🔊


Encodec Decoder
Decoder only transformer (non-autoregressive) #7
Decoder only transformer (non-autoregressive) #1
Decoder only transformer (autoregressive)
Tokenize
Codec Encoder
Text Prompt
Audio prompt

Bark

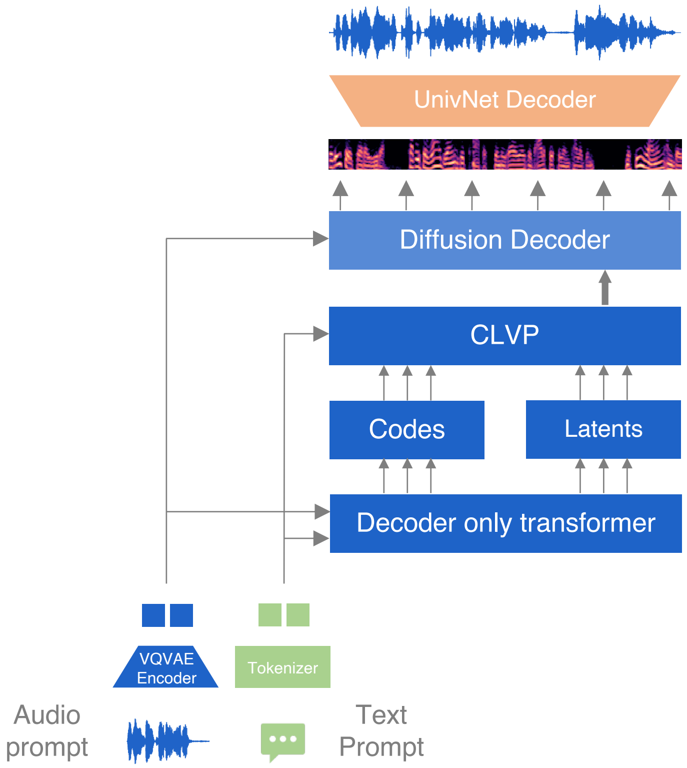
- Idea
  - Cascading of semantic token, coarse acoustic token, fine acoustic token three stage prediction (AudioLM)
  - 8 Layers of acoustic token prediction (VALLE)
  - Use Vocoder to generate wave form (Encodec)
- Example
  - Text prompt: “I have a silky smooth voice, and today I will tell you about the exercise regimen of the common sloth.”
  - TTS(build-in): 🔊
  - Speaker: 🔊
  - Clone: 🔊


Encodec Decoder
Fine Acoustic Token Predictor #7 (GPT)
Fine Acoustic Token Predictor #1 (GPT)
Coarse Acoustic Token Predictor (GPT)
Semantic Token Predictor (GPT)
BERT tokenizer
Text Prompt
History prompt

Tortoise TTS

- Idea
  - Autoregressive models convert between unaligned domains
  - CLIP like model to rank autoregressive model's output
  - Diffusion models captures expressive modalities

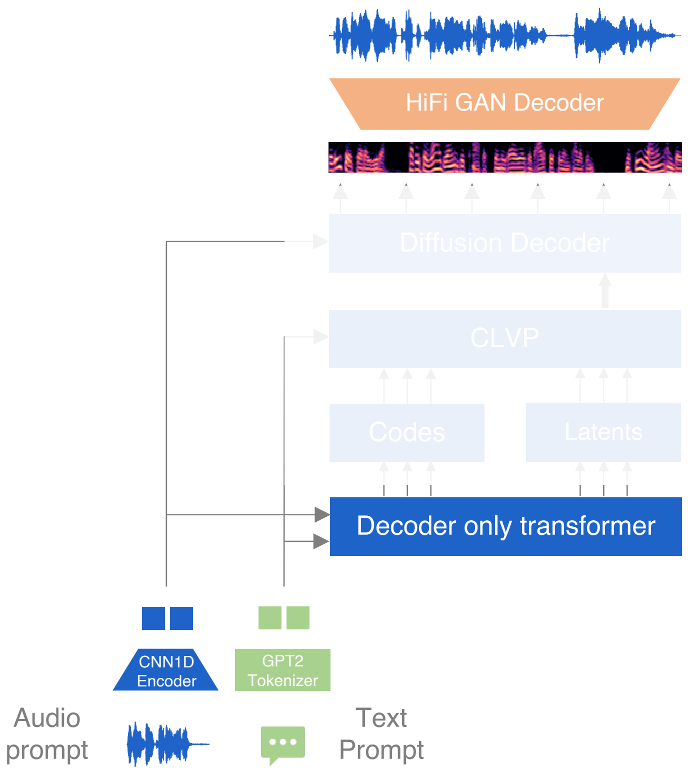
- Example
  - Text prompt: "I have a silky smooth voice, and today I will tell you about the exercise regimen of the common sloth."
  - TTS (build-in):
  - Speaker:
  - Clone:



XTTS

- Idea
  - Adapted from Tortoise TTS
- The inference code does not include CLVP and diffusion decoder Module, only the GPT2 and HiFi GAN Decoder

- Example
  - Text prompt: “I have a silky smooth voice, and today I will tell you about the exercise regimen of the common sloth.”
  - TTS(build-in): 🔊
  - Speaker: 🔊
  - Clone: 🔊



OpenAI TTS

- No details about the model yet

- Example:
  - Text prompt: “I have a silky smooth voice, and today I will tell you about the exercise regimen of the common sloth.”
  - TTS: 🔊

Outline

- Introduction
- An overview of models
- Demo: Finetune a TTS model
- References

Demo: Finetune a TTS Model

- Data preparation
  - Record sentences from Harvard Sentences
  - Example:
    - Text: There are more than two factors here.
    - Record: 🔊
- Download and Lunch training tool: mrq/ai-voice-cloning
- Load and transcribe the training data
- Validate configuration
- Train
- Test : 🔊

Outline

- Introduction
- An overview of models
- Demo: Finetune a TTS model
- References

References

1. https://elevenlabs.io
2. https://platform.openai.com/docs/guides/text-to-speech
3. https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer
4. https://github.com/neonbjb/tortoise-tts
5. https://github.com/suno-ai/bark
6. https://github.com/coqui-ai/TTS
7. Kong et al., HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, 2020
8. Jang et al., UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation, 2021
9. Zeghidour et al., SoundStream: An End-to-End Neural Audio Codec, 2021
10. Défossez et al., High Fidelity Neural Audio Compression (Encodec), 2022
11. Baevski et al., wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020
12. Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, 2020
13. Ao et al., SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing, 2021
14. Radford et al., Robust Speech Recognition via Large-Scale Weak Supervision (Whisper), 2022
15. Boros et al., AudioLM: a Language Modeling Approach to Audio Generation, 2022
16. Wang et al., Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers (VALLE), 2023
17. Betker, Better speech synthesis through scaling (Tortoise), 2023

Thanks for your attention