

Deep Learning for Audio

Lecture 10

Pavel Severilov

AI Masters

2024

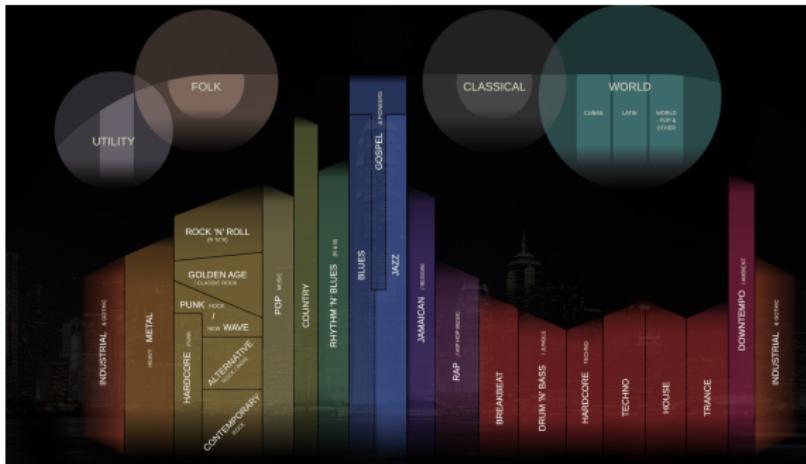
Outline

1. Music generation
2. Jukebox
3. Diffsound
4. AudioLM & MusicLM
5. AudioCraft: AudioGen & MusicGen

Outline

1. Music generation
2. Jukebox
3. Diffsound
4. AudioLM & MusicLM
5. AudioCraft: AudioGen & MusicGen

Problems with generating music



- ▶ Should NNs compose music by following the same logic and process as humans do?
- ▶ Are current evaluation methods good enough to compare and measure the creativity of the composed music?
- ▶ Difficult to collect data due to copyright
- ▶ Music is subjective
- ▶ Each music Genre has its own rules

Datasets

- ▶ JSB Chorales Dataset (chorales by Johann Sebastian Bach)
- ▶ Maestro Dataset (200 hours of virtuosic piano performances with fine alignment between note labels and audio waveforms)
- ▶ The Lakh MIDI Dataset (176,581 unique MIDI files)
- ▶ MetaMIDI Dataset (436,631 MIDI files)

Labs, apps, ideas

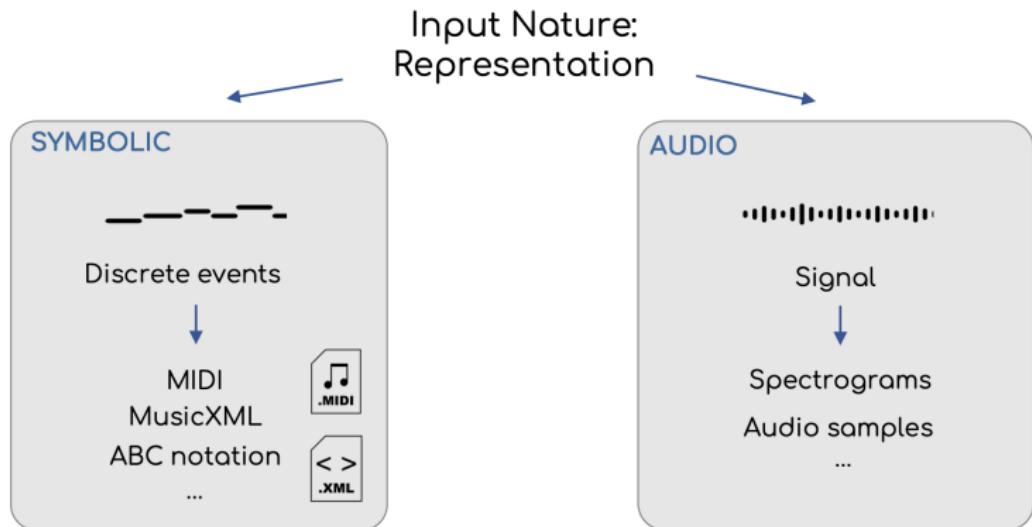
Research Groups and Labs:

- ▶ Google Magenta
- ▶ Audiolabs Erlangen
- ▶ Music Informatics Group
- ▶ Music and Artificial Intelligence Lab
- ▶ Metacreation Lab

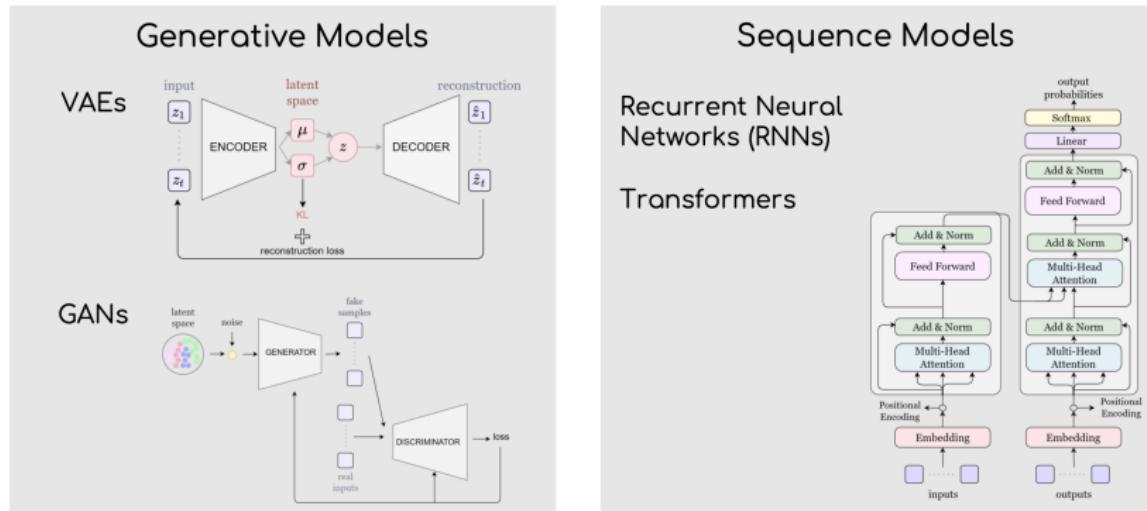
Apps:

- ▶ AIVA
- ▶ Amper Music
- ▶ Ecrett Music
- ▶ Humtap
- ▶ Amadeus Code
- ▶ Computoser
- ▶ Brain.fm

Input representations



History: before 2019



Let's use models from CV/NLP and not give context about data domain (music)!

MuseNet (OpenAI, 2019)



- ▶ Learn to predict the next token
- ▶ Uses GPT-2
- ▶ MIDI format
- ▶ Sparse Transformer to train a 72-layer network with 24 attention heads
- ▶ Long context: full attention over a context of 4096 tokens

Outline

1. Music generation
2. Jukebox
3. Diffsound
4. AudioLM & MusicLM
5. AudioCraft: AudioGen & MusicGen

Quantization

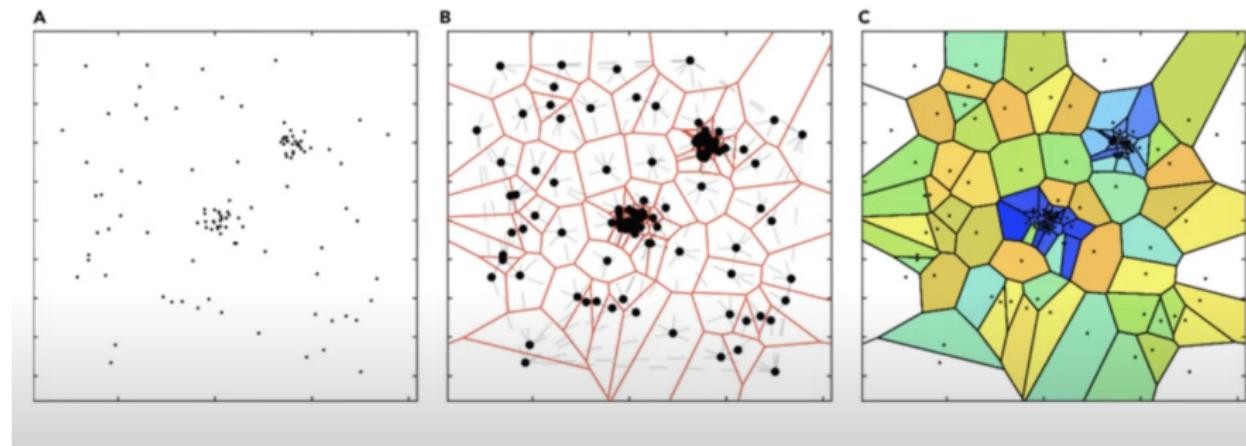
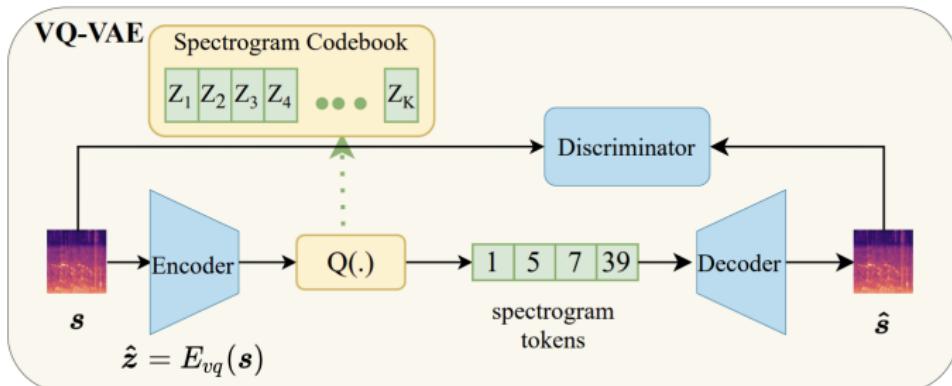


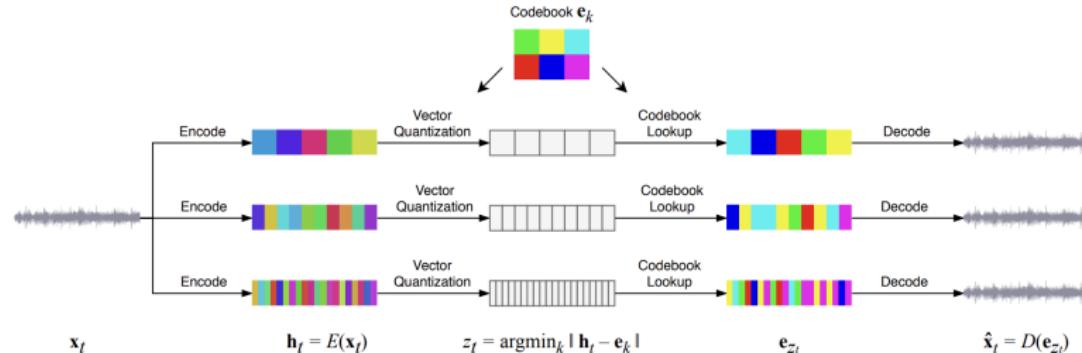
Figure: Recap: Voronoi diagram

VQ-VAE



- ▶ Encoder extracts the representation z from the mel-spectrogram
- ▶ Codebook contains a finite number of embedding vectors
- ▶ Decoder reconstructs the mel-spectrogram based on mel-spectrogram tokens
- ▶ Discriminator distinguishes the mel-spectrogram is original or reconstructed
- ▶ $Q(\cdot)$ denotes a spatial-wise quantizer that maps each features z_{ij} into its closest codebook entry z_k

Jukebox (OpenAI, 2020)

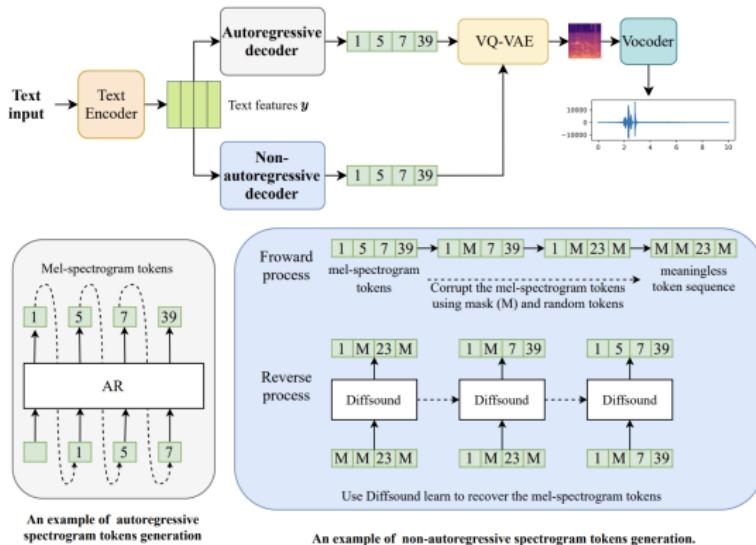


- ▶ Based on VQ-VAE
- ▶ Separated autoencoders with different temporal resolutions
- ▶ Spectral loss

Outline

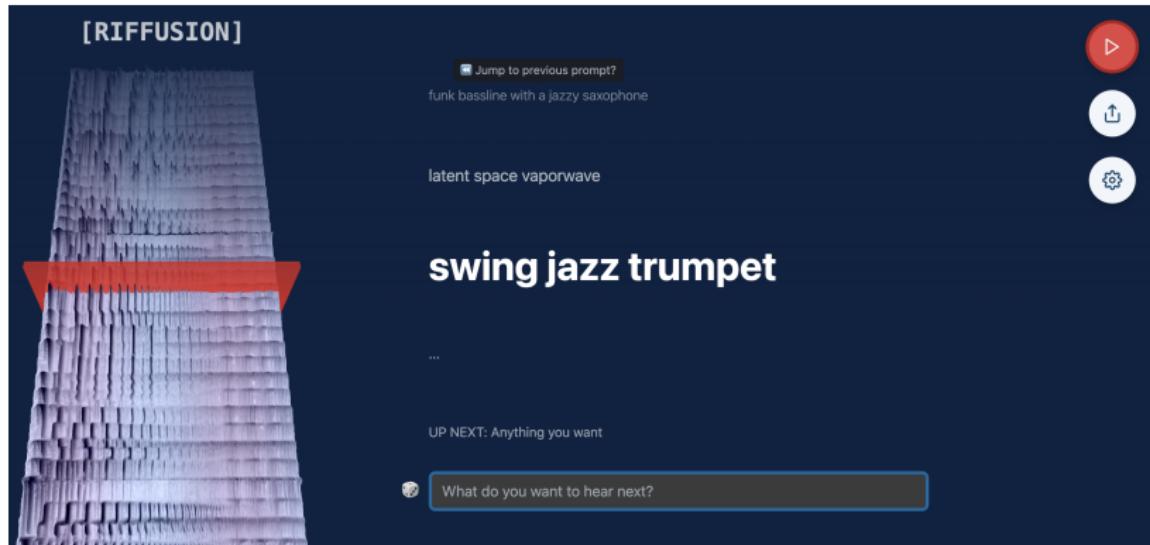
1. Music generation
2. Jukebox
3. Diffsound
4. AudioLM & MusicLM
5. AudioCraft: AudioGen & MusicGen

Diffsound



- ▶ Text encoder (CLIP) extracts text features from the text input
- ▶ Decoder generates mel-spectrogram tokens
- ▶ Pre-trained VQ-VAE transforms the tokens into mel-spectrogram
- ▶ Vocoder transforms mel-spectrogram into waveform

Riffusion (2022)



- ▶ Real-time music generation with stable diffusion
- ▶ Fine-tuned the model to generate images of spectrograms
- ▶ Combine short clips by sampling the latent space between a prompt with two different seeds, or two different prompts with the same seed

Riffusion: How to make smooth transition?

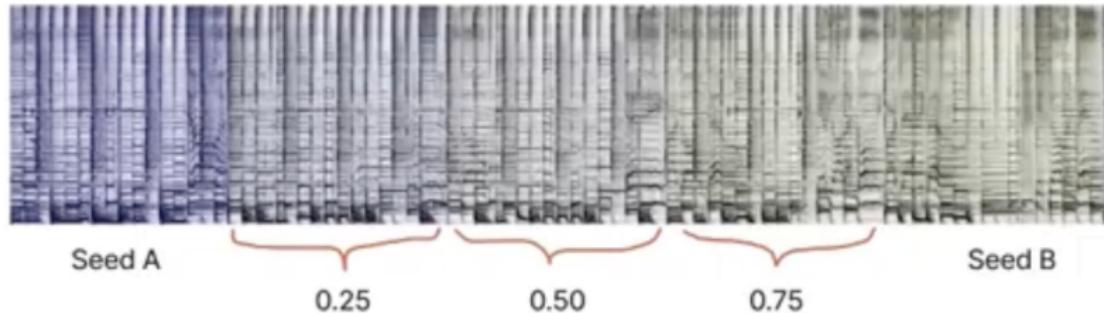


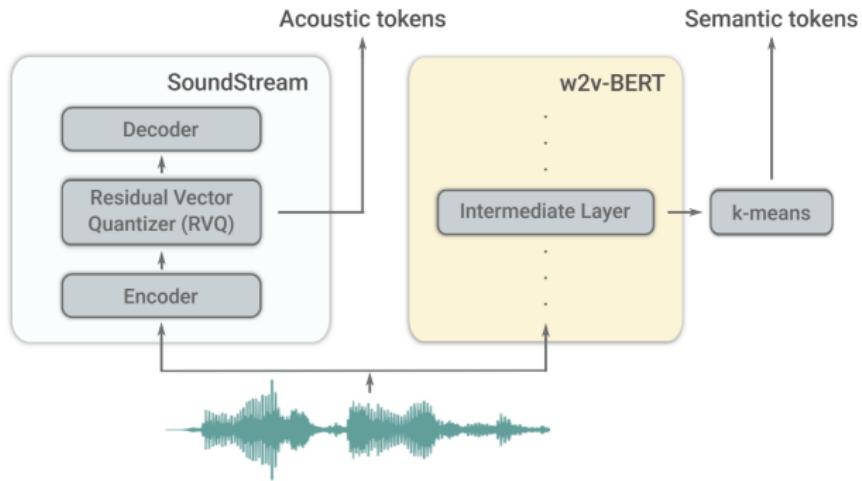
Figure: Latent space interpolation between two seeds of the same prompt

- ▶ laying the sequence is much smoother than just playing the two endpoints
- ▶ Interpolated clips are often diverse and have their own riffs and motifs come and go

Outline

1. Music generation
2. Jukebox
3. Diffsound
4. AudioLM & MusicLM
5. AudioCraft: AudioGen & MusicGen

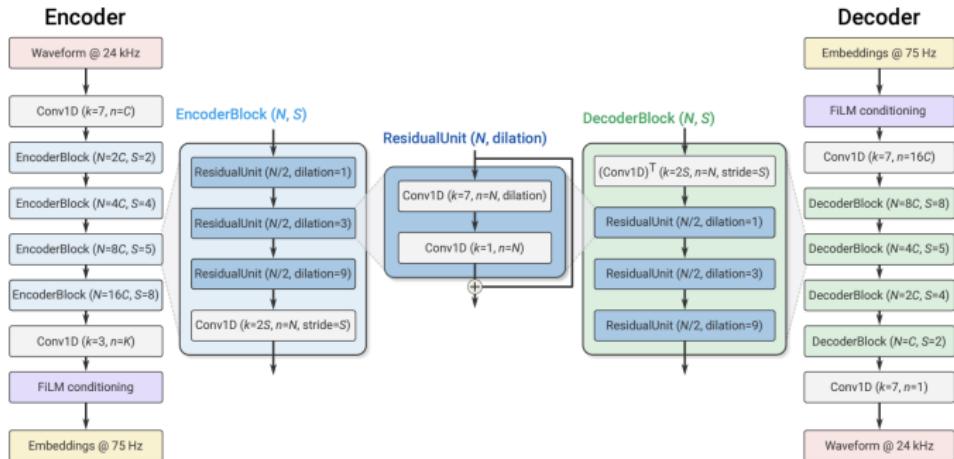
AudioLM (Google Research, 2022)



- ▶ SoundStream: 24 kHz audio to 50 Hz embeddings (Autoencoder codec with vector quantizers)
- ▶ w2v-BERT (wav2vec2 + MLM loss): extract embeddings from the 7th layer and quantize them using the centroids of a learned k-means over the embeddings

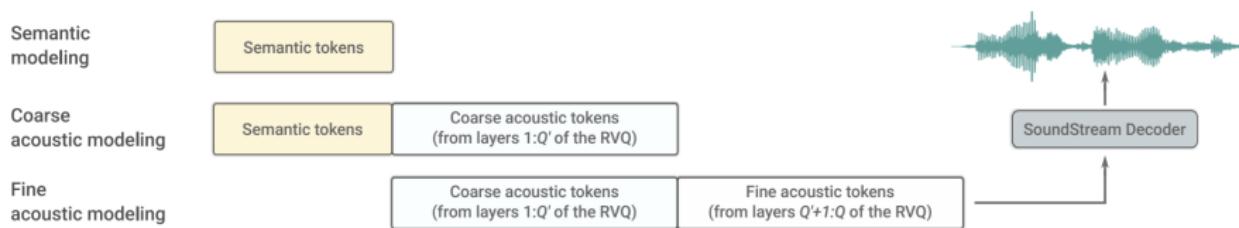
Borsos, Marinier et al., *AudioLM: a Language Modeling Approach to Audio Generation*, IEEE/ACM, 2022

Soundstream codec

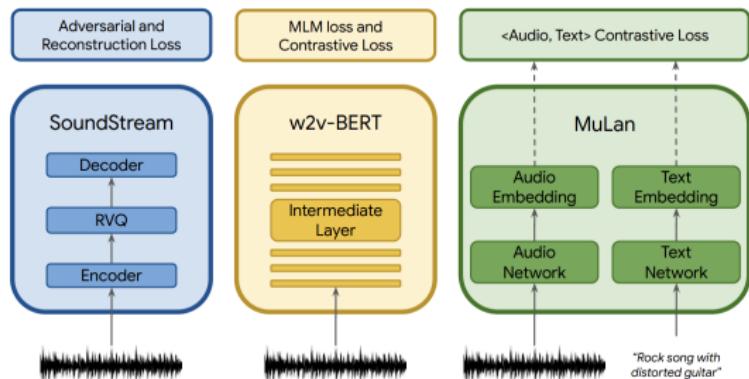


- ▶ A fully convolutional encoder/decoder network + a residual vector quantizer, trained jointly end-to-end
- ▶ Can efficiently compress speech, music and general audio
- ▶ Outperforms SOTA codecs at 3 kbps using audio at 24 kHz sampling rate

AudioLM: decoding



MusicLM (Google Research, 2023)



- ▶ Based on AudioLM: SoundStream + w2v-BERT + MuLan
- ▶ MuLan tokens
- ▶ Metrics Frechet Audio Distance (FAD), KL

MusicLM: train & generating

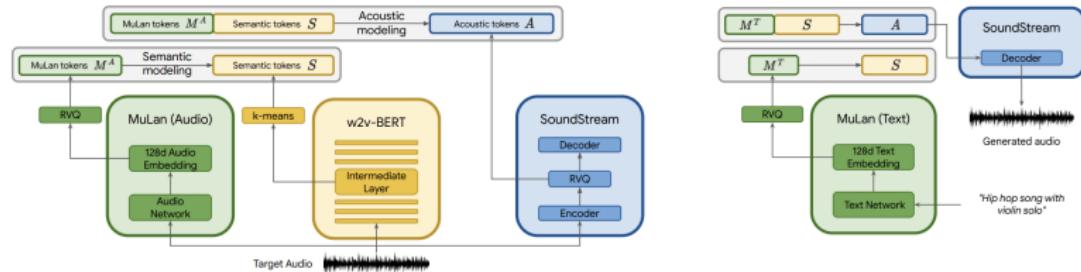


Figure: Training on the left, generating on the right

Suno: Bark

The screenshot shows the Suno app interface. At the top left is the Suno logo. Below it is a section titled "Explore new styles of music with Suno". A text box says "Here's a small taste (or whatever the listening equivalent is) of what's possible". In the center is a dark box with the text "Pick a style, or roll the dice...". To the right of this box is a grid of music styles. At the bottom is a pink button labeled "What will you create?".

grid of music styles:

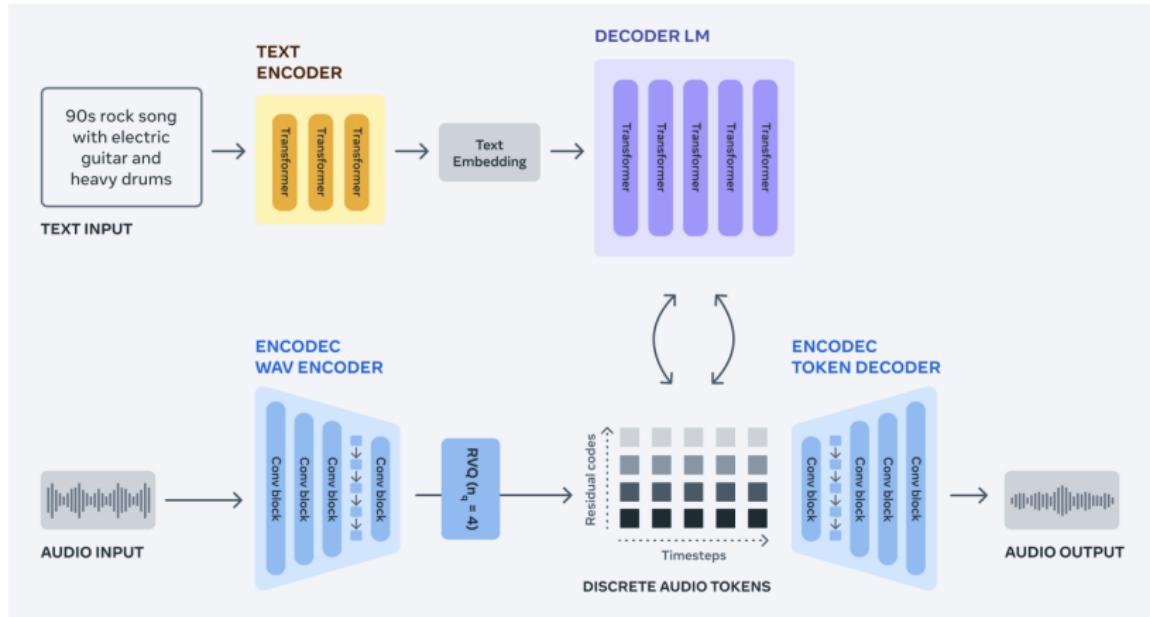
- french dubstep
- bedroom pop simba
- j-pop blues
- lo-fi afro house
- drum and bass
- afrobeat rockabilly
- dance
- disco chillsynth
- acoustic slushwave
- russian flamenco
- hip hop dubstep
- classical bohemian
- accordion drill
- drum and bass
- afrobeat rockabilly
- carnatic glitch hop
- psychedelic acid trance
- french dewbow
- japanese swing
- boogie caribbean
- carnatic
- classical
- bedroom pop new jack swing
- liquid drum and bass slushwave
- p-funk simba
- gospel disco
- harpsichord klezmer
- j-pop acid jazz
- drum and bass acoustic rock
- ambient tranc
- alternative rock
- dakar raga
- disco chillstep
- drum and bass acoustic rockabil
- afrobeat new jack swing
- hip hop techn
- sitar sertanejo
- accordion delta blues
- acoustic rockabil
- russian dubstep
- drill sertanejo
- psybient new wave
- rumbe
- french big band
- japanese surf rock
- classical
- rumba cape verdean
- bedroom pop
- boogie
- carnatic acid tranc
- drill sertanejo
- psybient hip hop
- j-pop acid breaks
- liquid drum and bass rockabilly
- p-funk mariachi
- gospel acid breaks
- harpsichord hip hop
- dakar new wave
- disco alternative r&b
- drill sertanejo
- goa
- ambient techno mento
- accordion ambient techn
- acoustic rock chillsynth
- afrobeat griot
- ambient techno mento
- french afro-rock
- japanese surf
- russian celtic
- hip hop
- sitar rumba
- barbershop breakbeat
- blues rock american primitivism
- carnatic
- city pop symphonic metal
- indie grunge
- liquid drum and bass new jack swing
- p-funk jazz
- psybient griot
- roots reggae house
- dakar math rock
- dirty south boom bap
- drill raga
- goa
- trance afro-cuban jazz
- harpsichord g-funk
- accordion afro trap
- acoustic rock afro-jazz
- afrobeat garage
- ambient techno hyphy
- life and drum blues p-funk
- japanese shou

- ▶ Uses Bark model
- ▶ Bark is architecturally very similar to Google's AudioLM

Outline

1. Music generation
2. Jukebox
3. Diffsound
4. AudioLM & MusicLM
5. AudioCraft: AudioGen & MusicGen

AudioCraft (Meta, 2023)



- ▶ AudioGen & MusicGen