Deep Learning for Audio

Lecture 10

Pavel Severilov

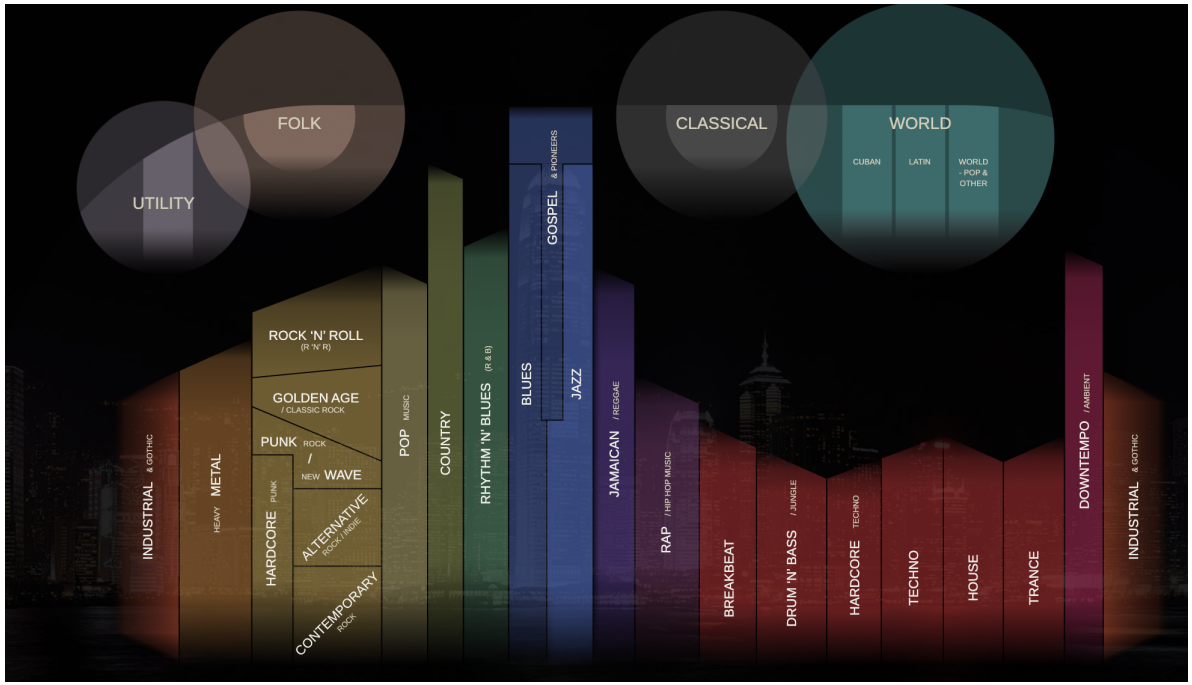
AI Masters

2024

Outline

Outline

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

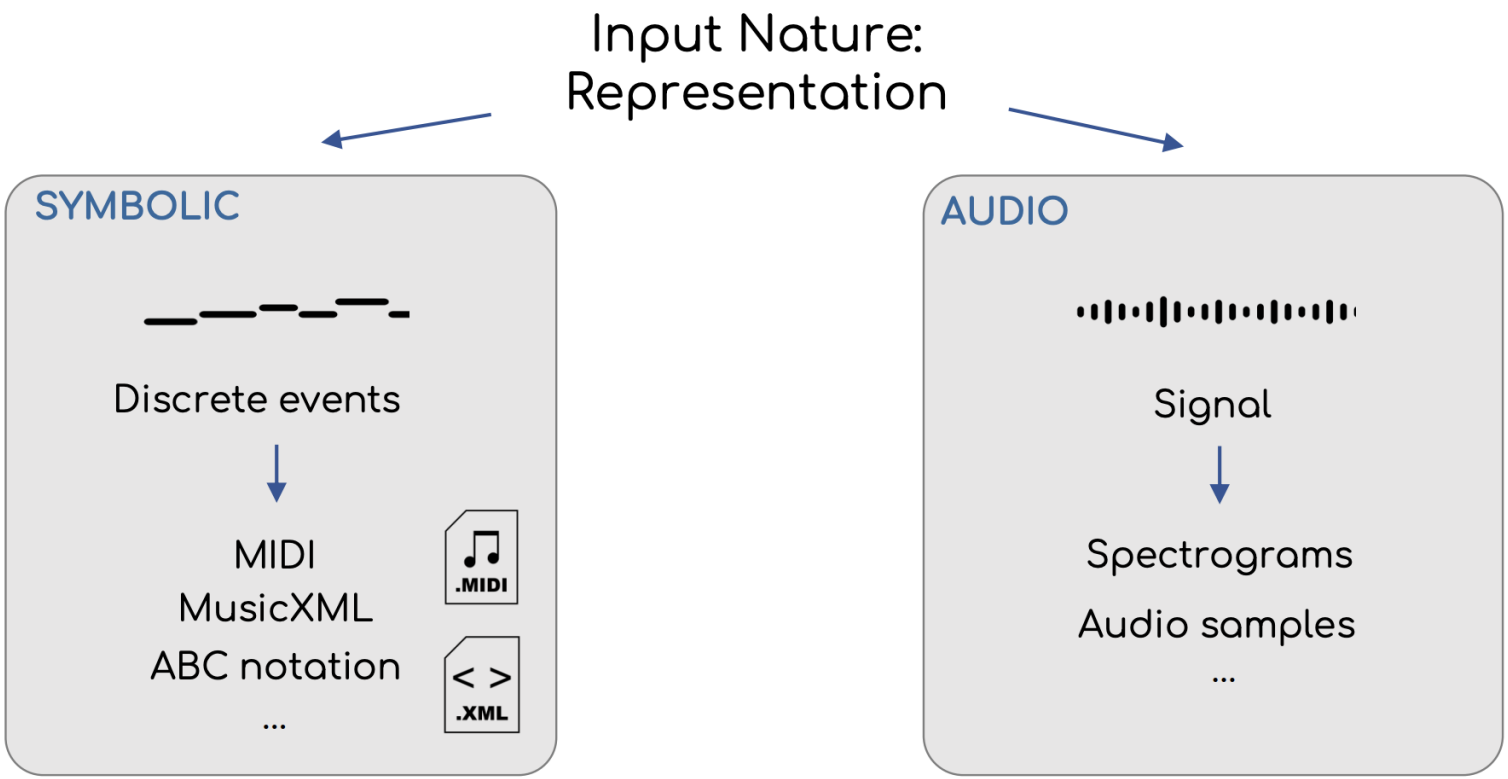
Research Groups and Labs:

- 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

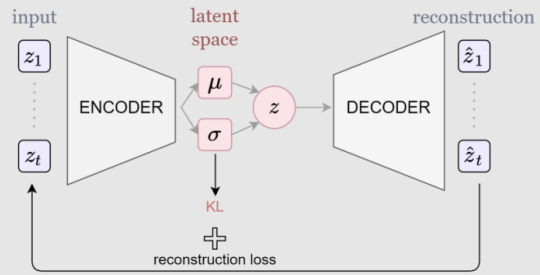
Input representations



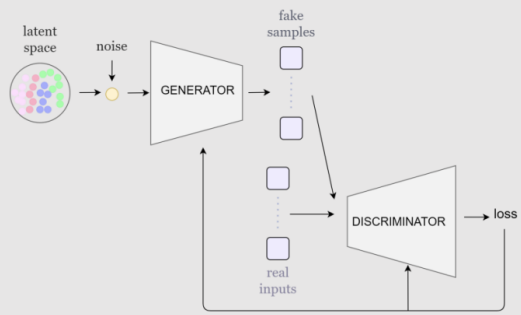
History: before 2019

Generative Models

VAEs



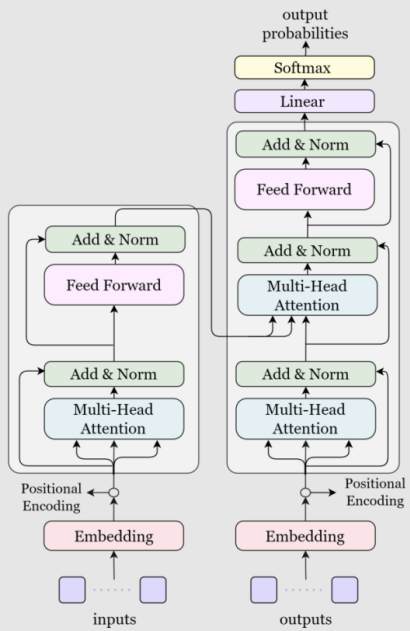
GANs



Sequence Models

Recurrent Neural Networks (RNNs)

Transformers



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

Quantization

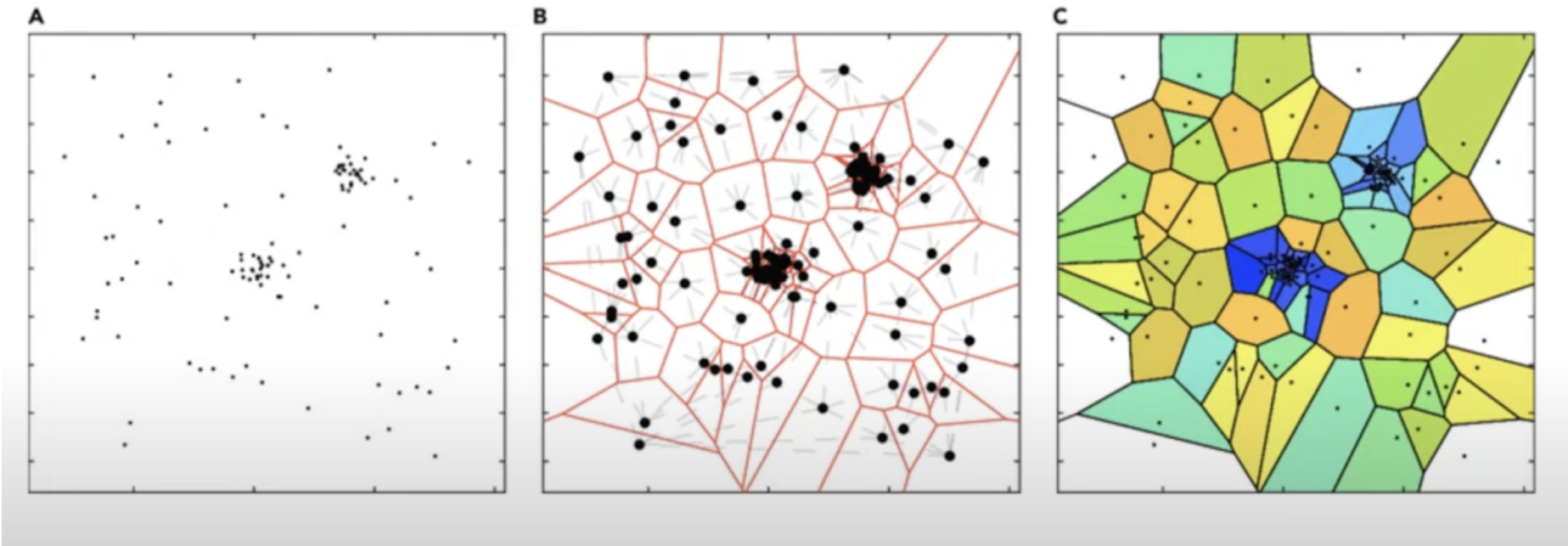
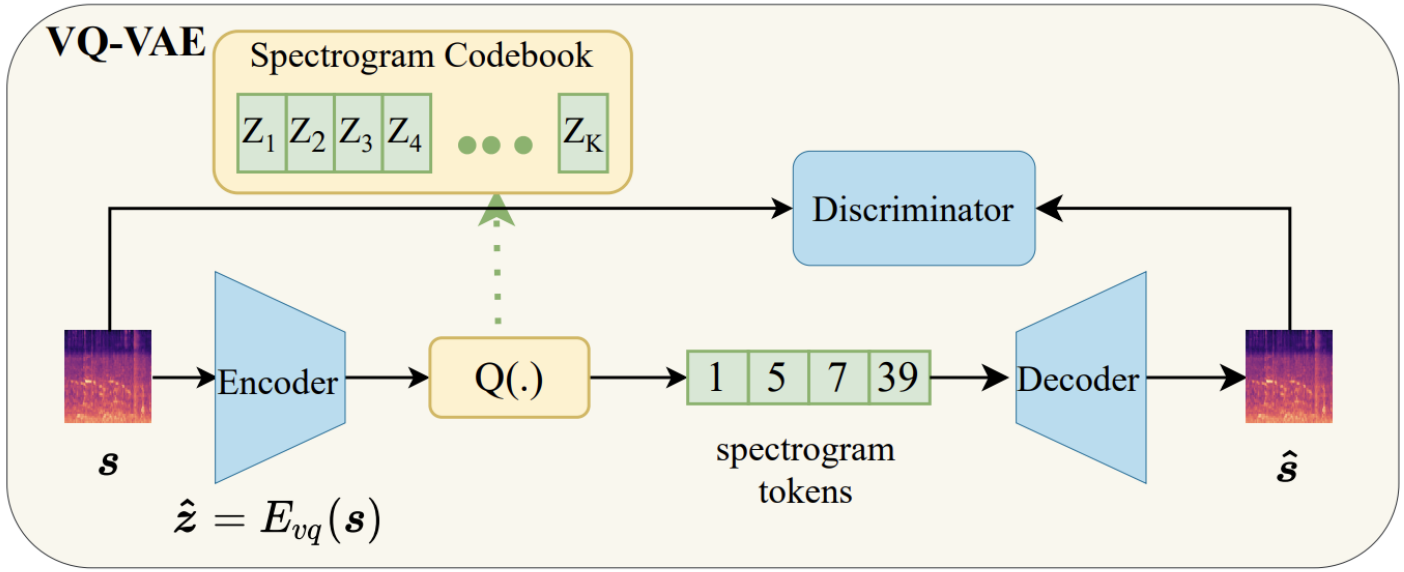


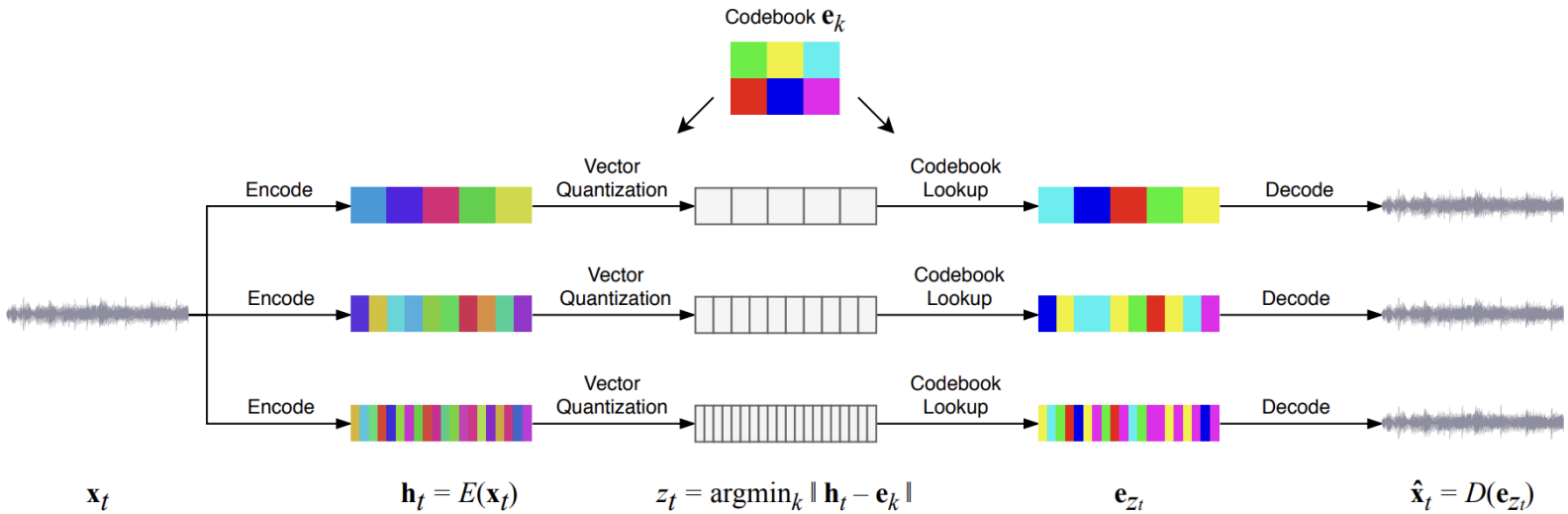
Figure: Recap: Voronoi diagram

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

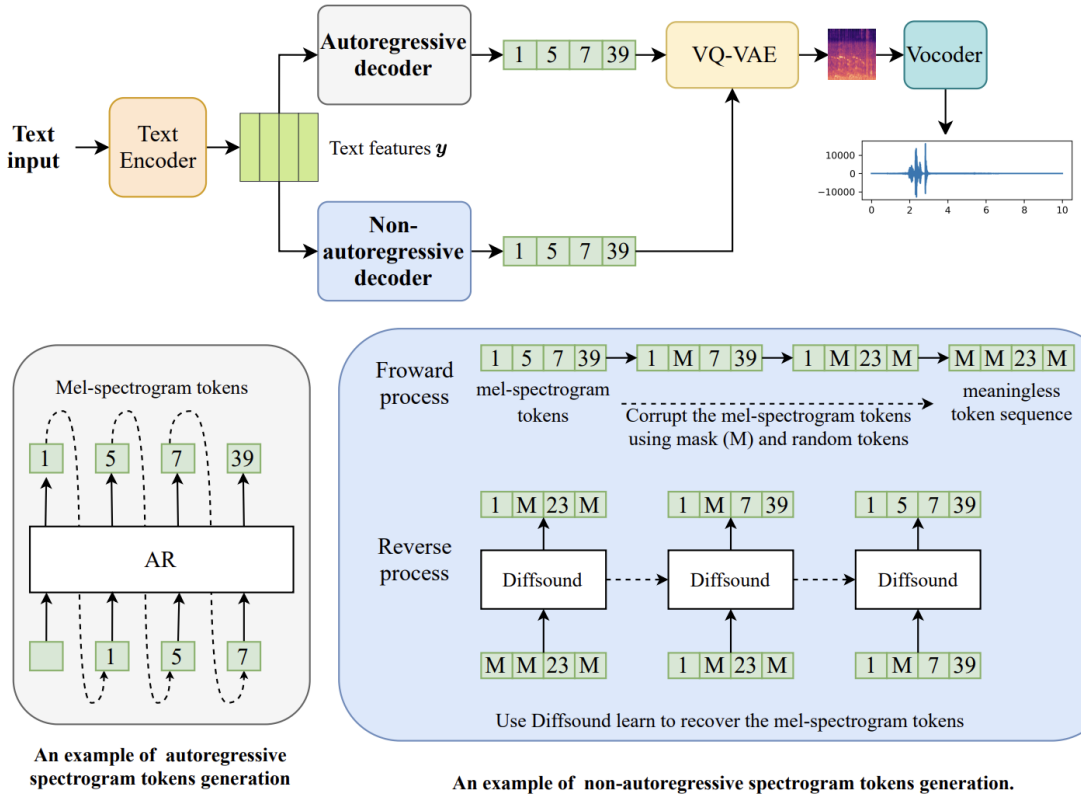
Jukebox (OpenAI, 2020)



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

Outline

Diffsound

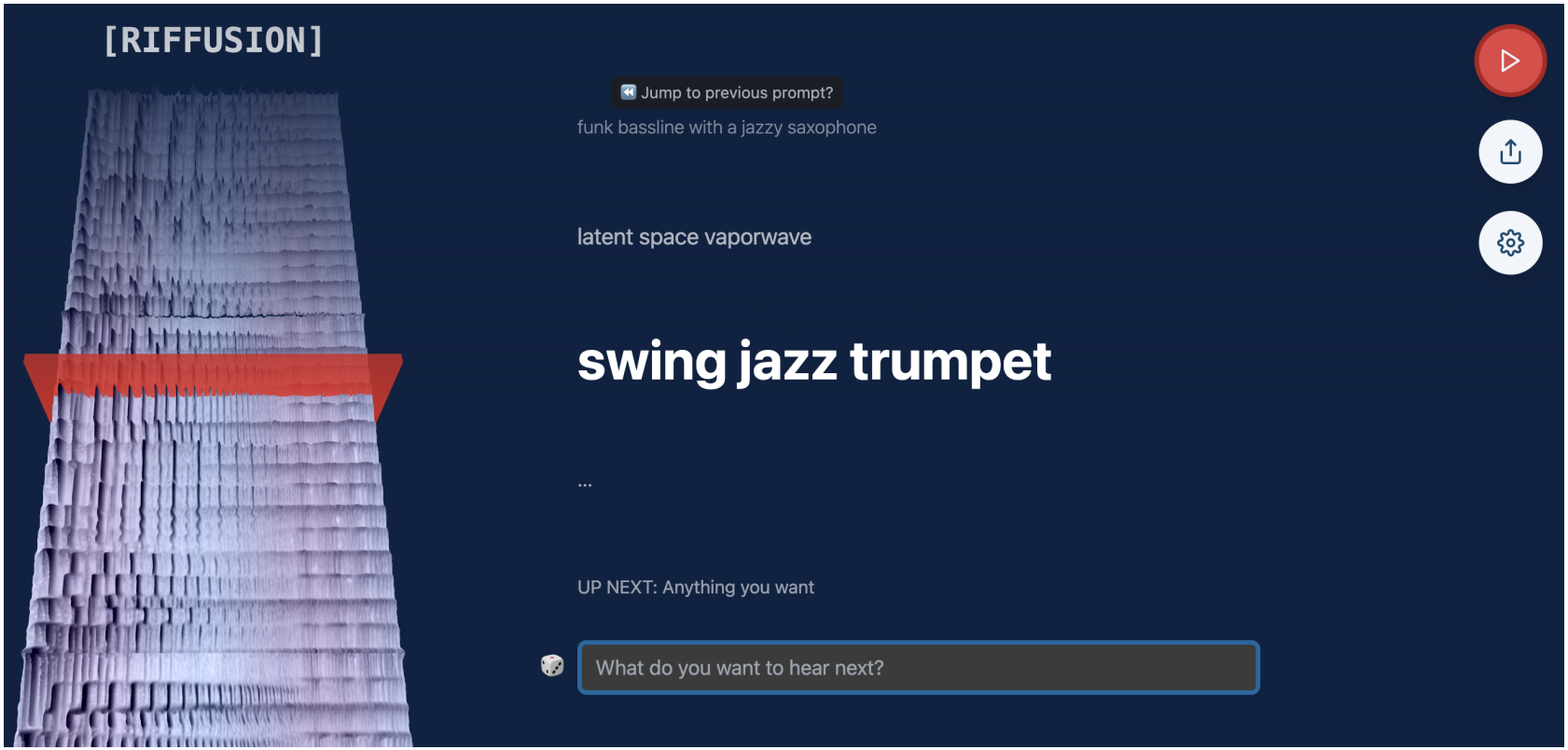


An example of autoregressive spectrogram tokens generation

An example of non-autoregressive spectrogram tokens generation.

- 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?

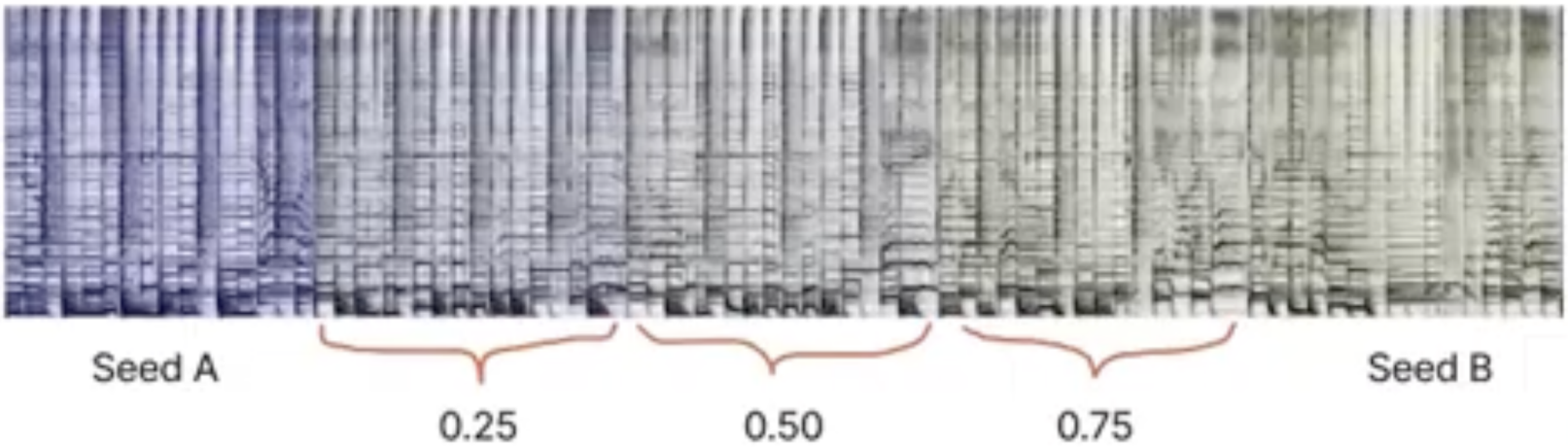
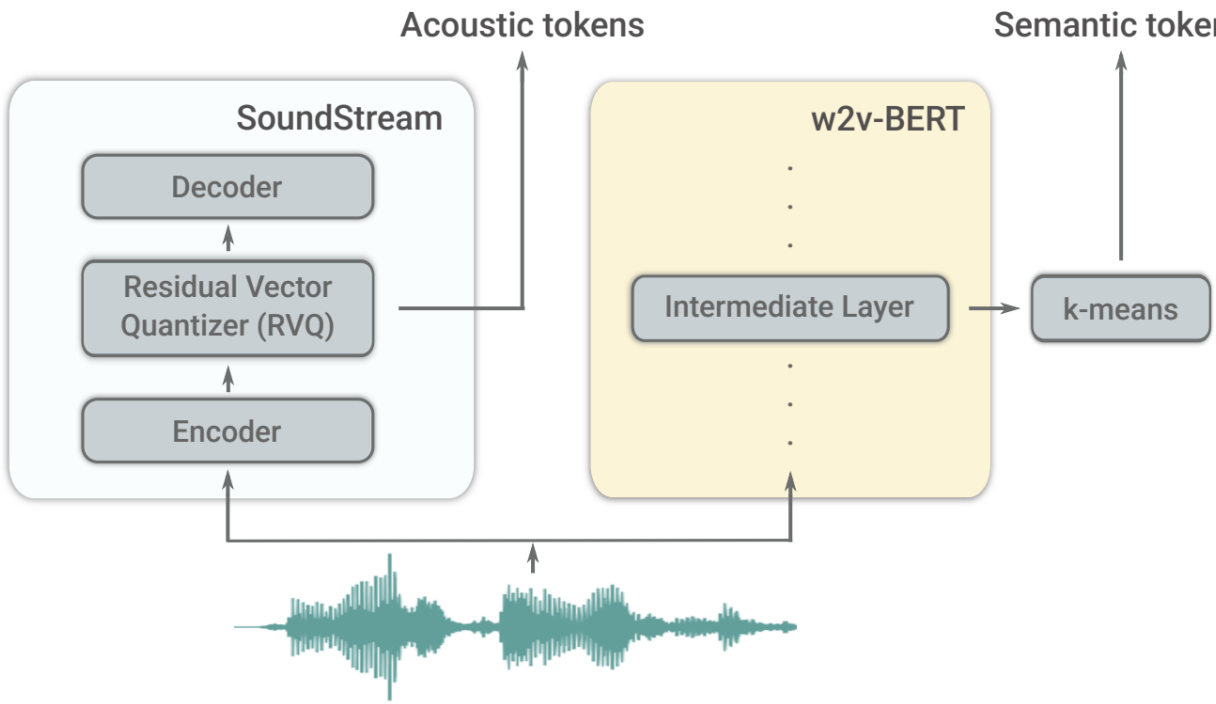


Figure: Latent space interpolation between two seeds of the same prompt

- 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

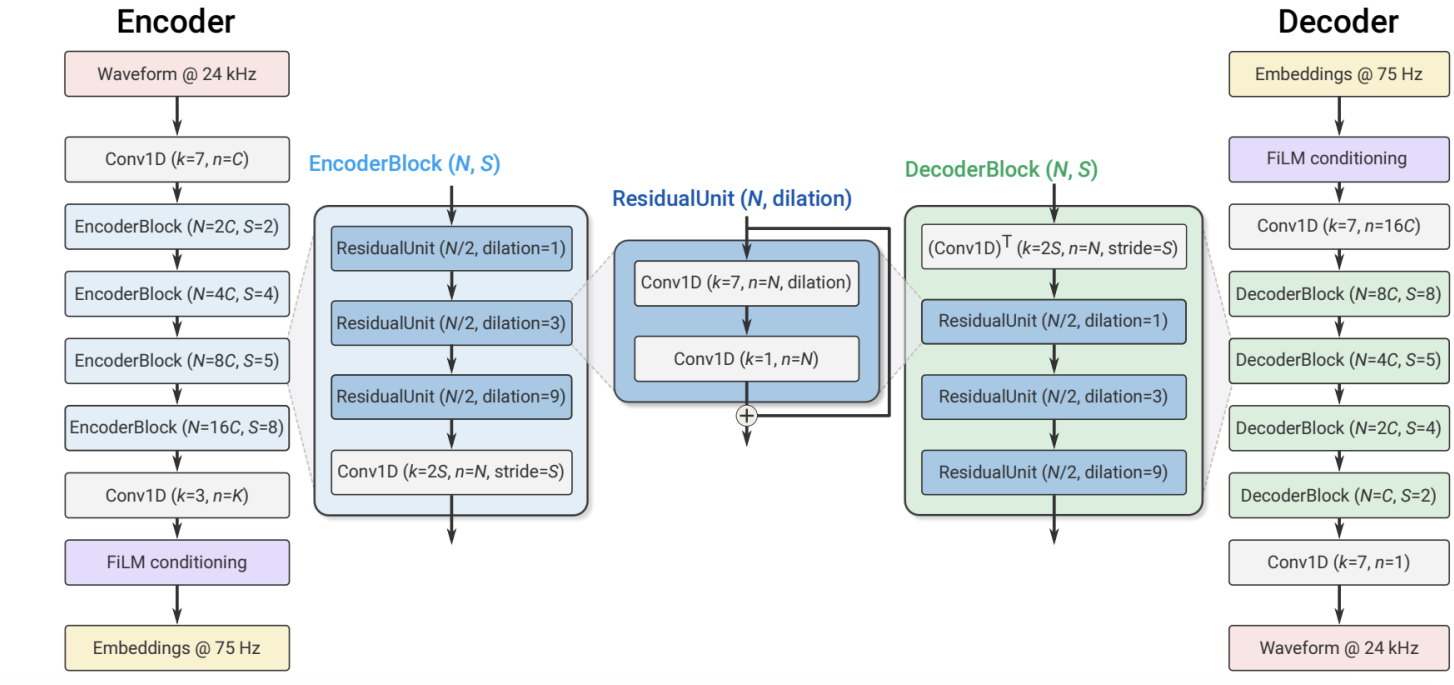
AudioLM (Google Research, 2022)



- 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

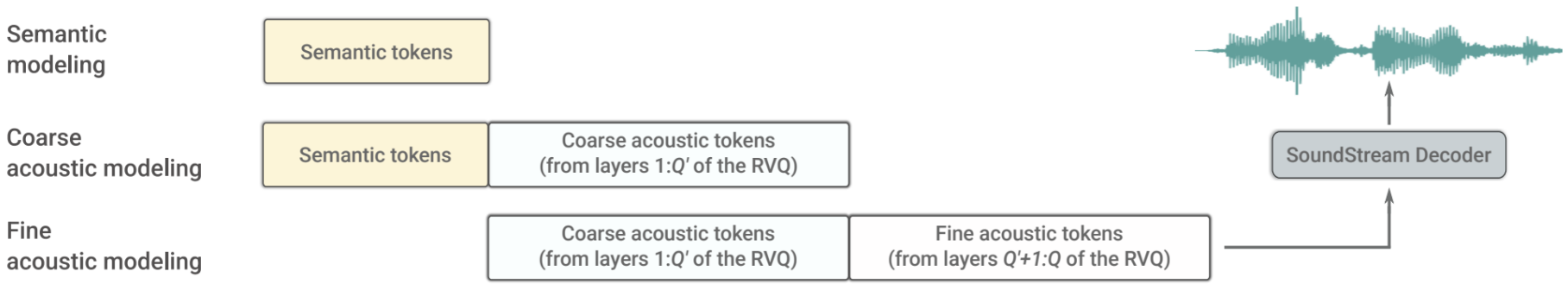
Soundstream codec



- 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

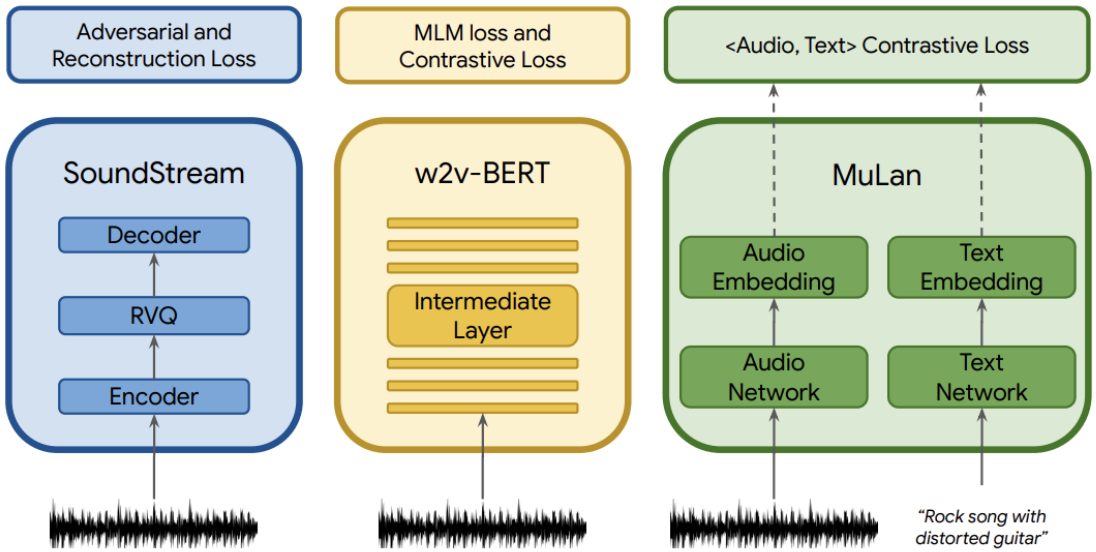
Zeghidour et al., SoundStream: An End-to-End Neural Audio Codec, IEEE/ACM, 2021

AudioLM: decoding



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

MusicLM (Google Research, 2023)



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

Kim, Jaehyeon et al., MusicLM: Generating Music From Text, Arxiv preprint, 2023

MusicLM: train & generating

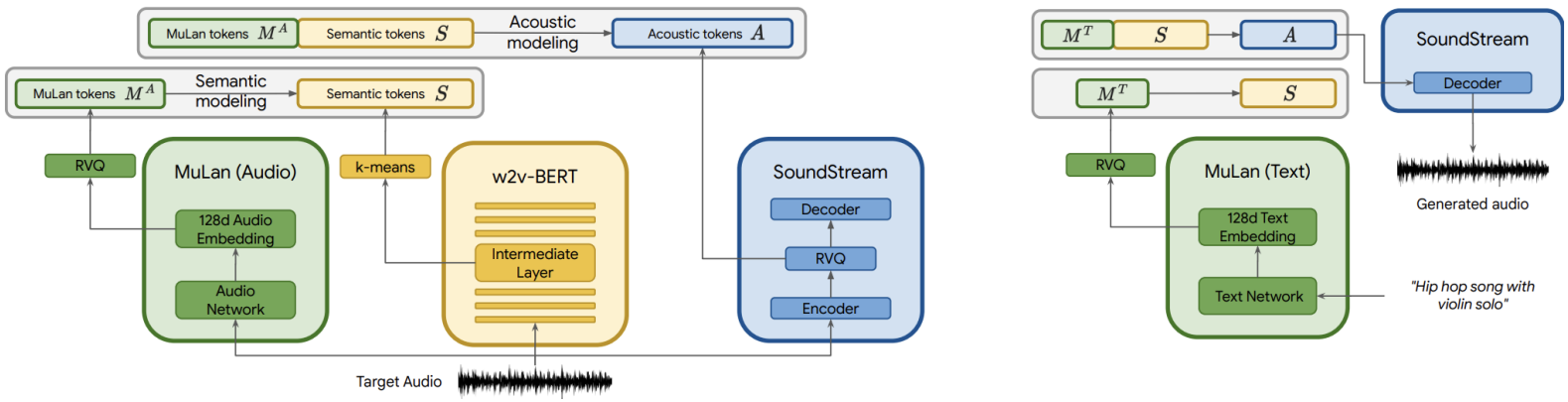
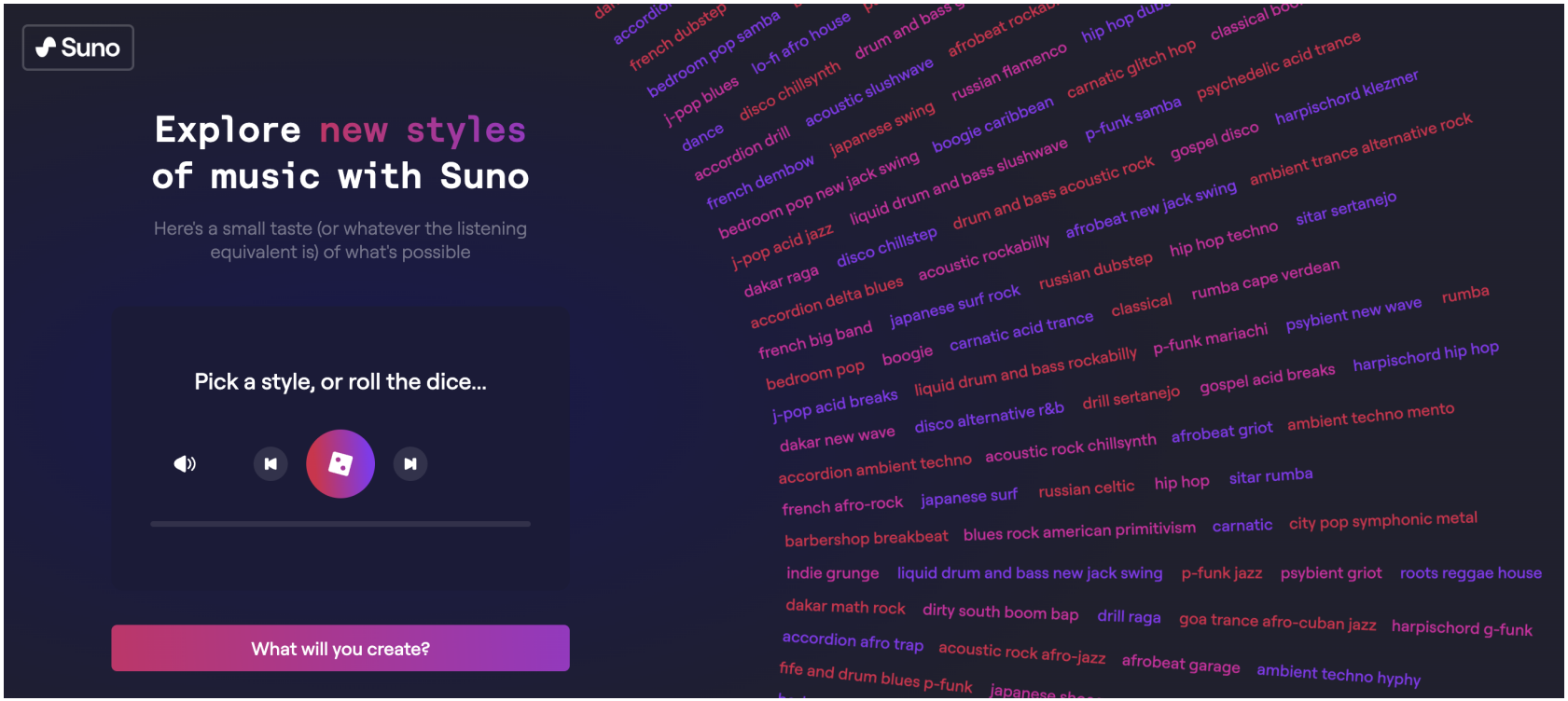


Figure: Training on the left, generating on the right

Kim, Jaehyeon et al., MusicLM: Generating Music From Text, Arxiv preprint, 2023

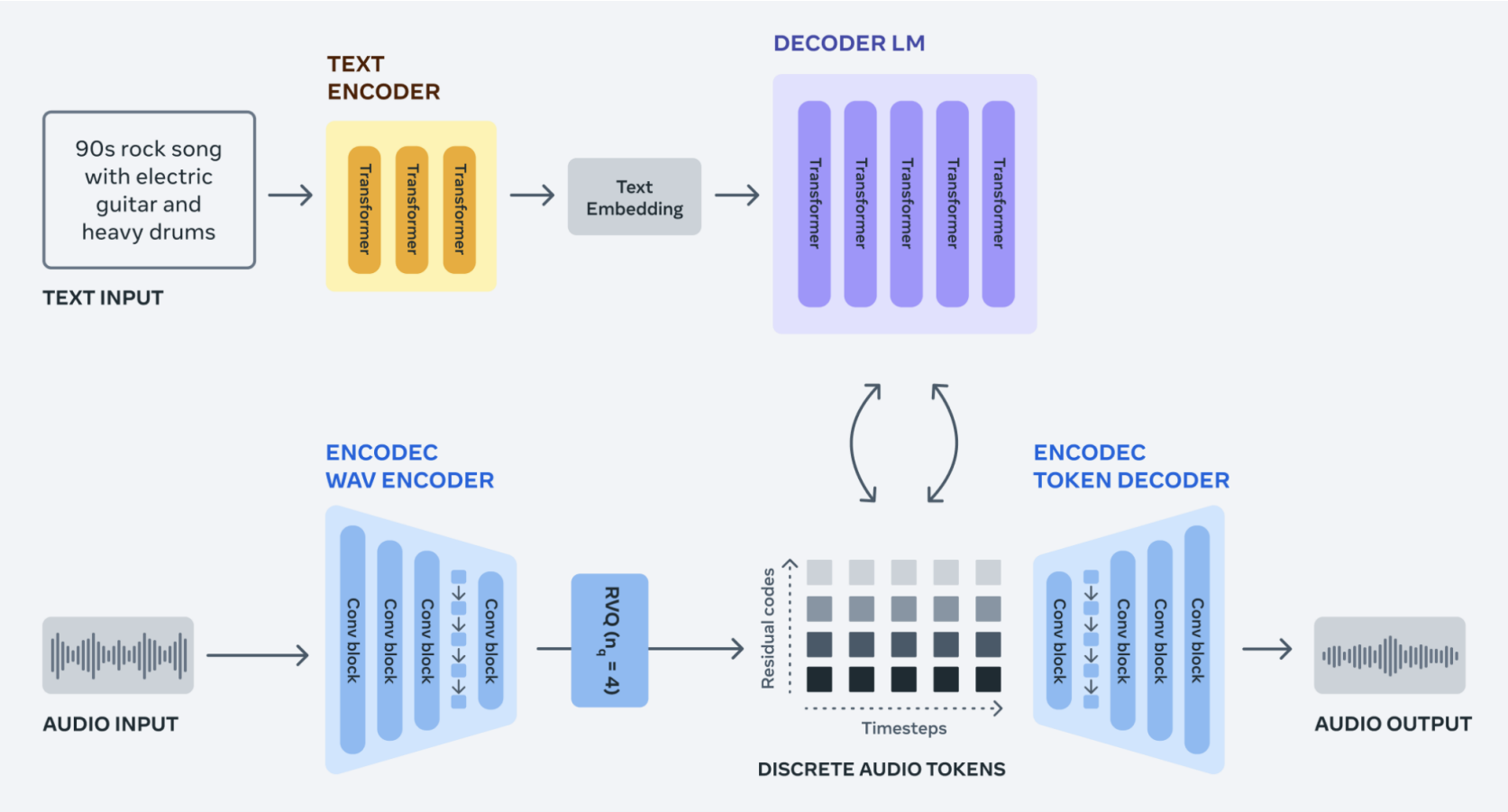
Suno: Bark



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

Outline

AudioCraft (Meta, 2023)



- AudioGen & MusicGen