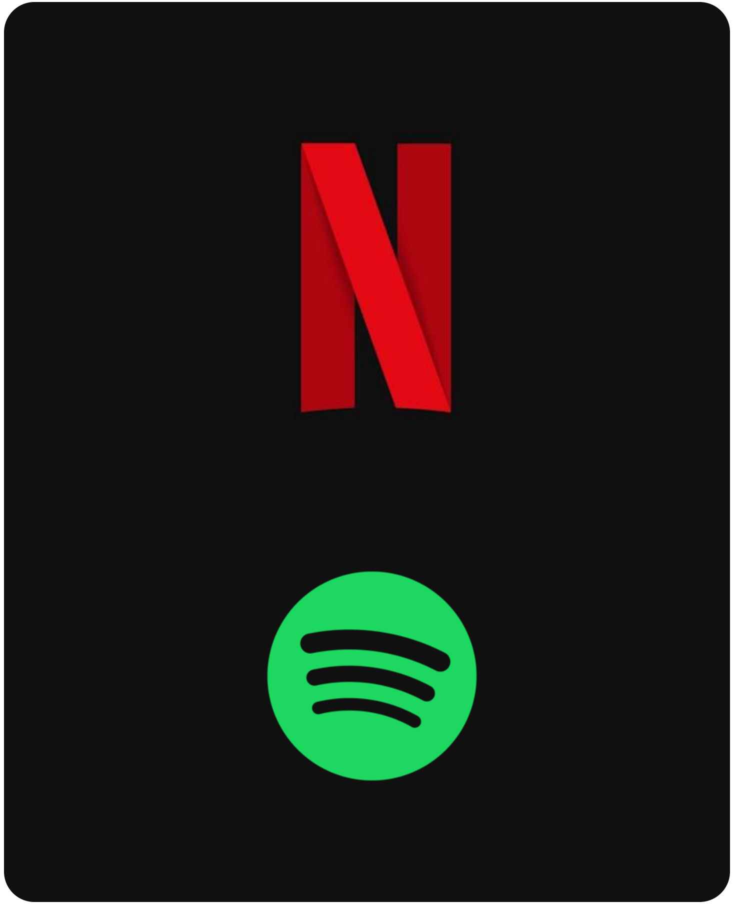
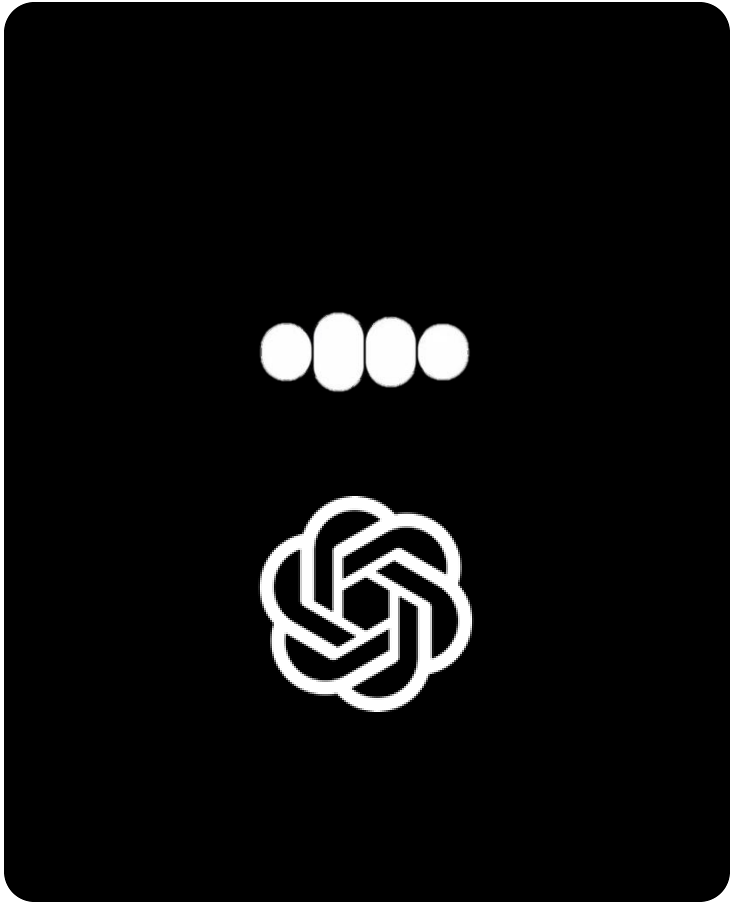
High-Fidelity Audio Compression with Improved RVQGAN

Presented by: Nandor Kofarago

ETH zürich

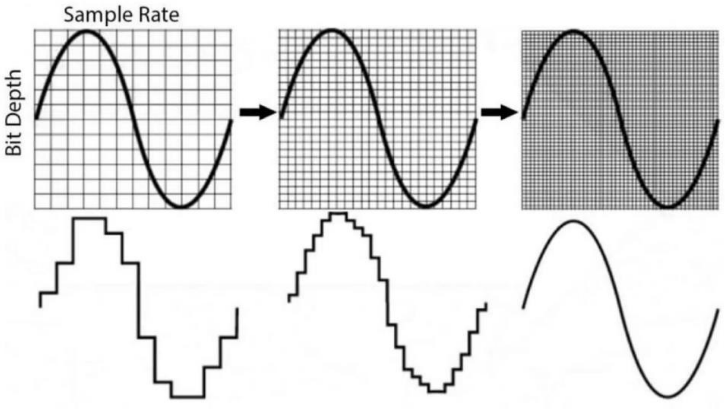




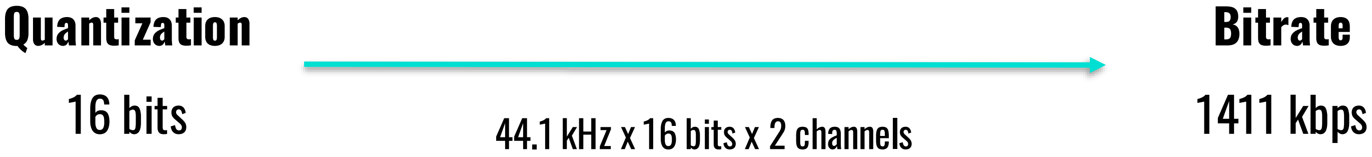
Audio sampling

Human hearing → **Sample rate** → **Eliminate aliasing**

20 Hz – 20 kHz → 40 kHz → 44.1 kHz



Digital quantization



Quantization

16 bits

44.1 kHz x 16 bits x 2 channels

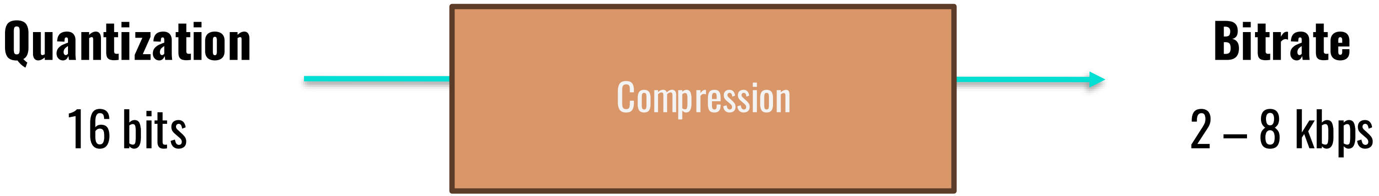
Bitrate

1411 kbps

Uncompressed audio: **630 MB / hour**

Tokenizing: **44100 tokens / s ?**

Digital quantization



Quantization

16 bits

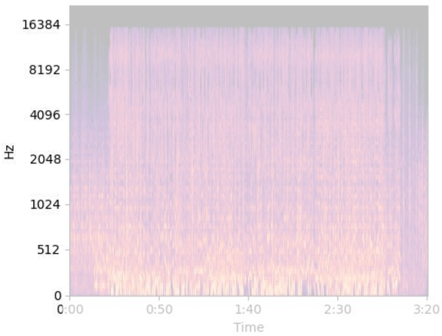
Compression

Bitrate

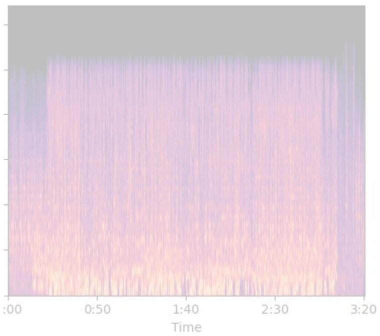
2 – 8 kbps

Compressed audio: ~ **2 MB / hour**

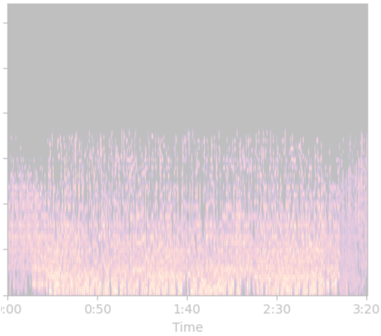
Demo



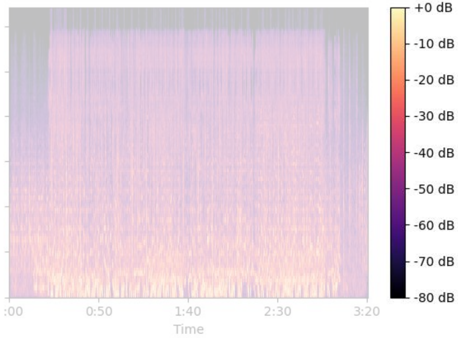
Original



Opus – 8kbps



MP3 – 8 kbps



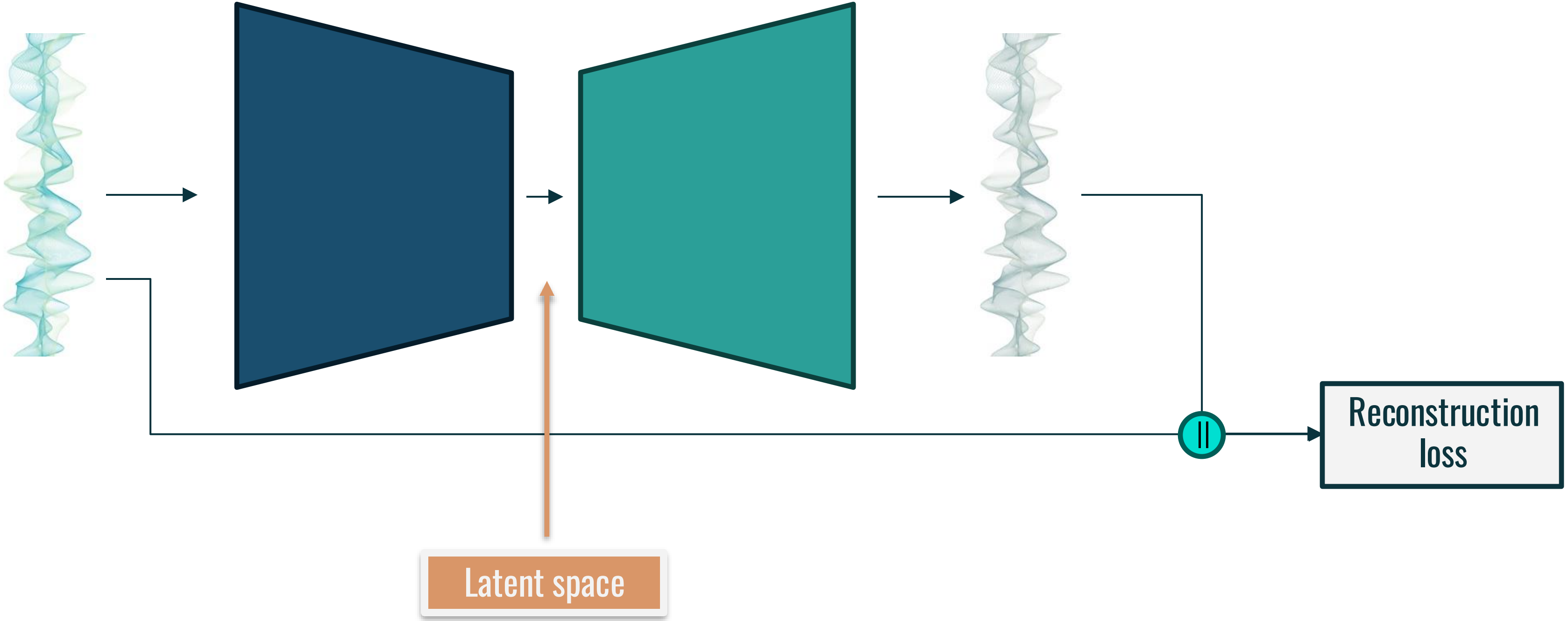
RVQGAN – 8 kbps

What do we need?

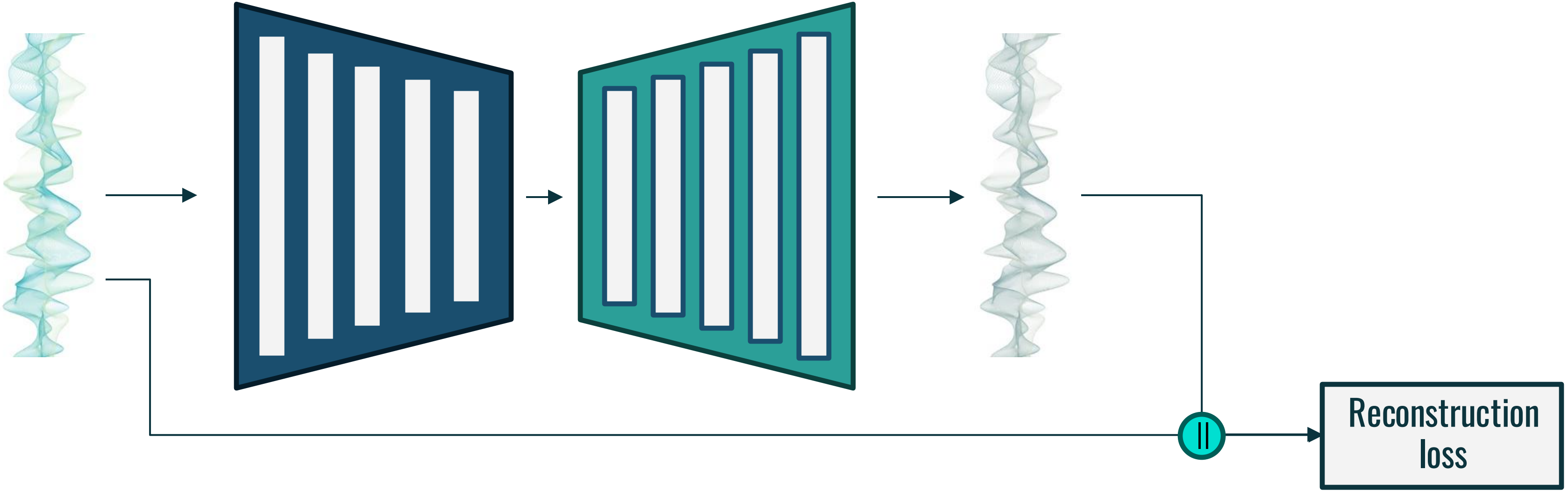
1. Efficient compression
2. Tokenizing audio
3. Generating audio

Model architecture

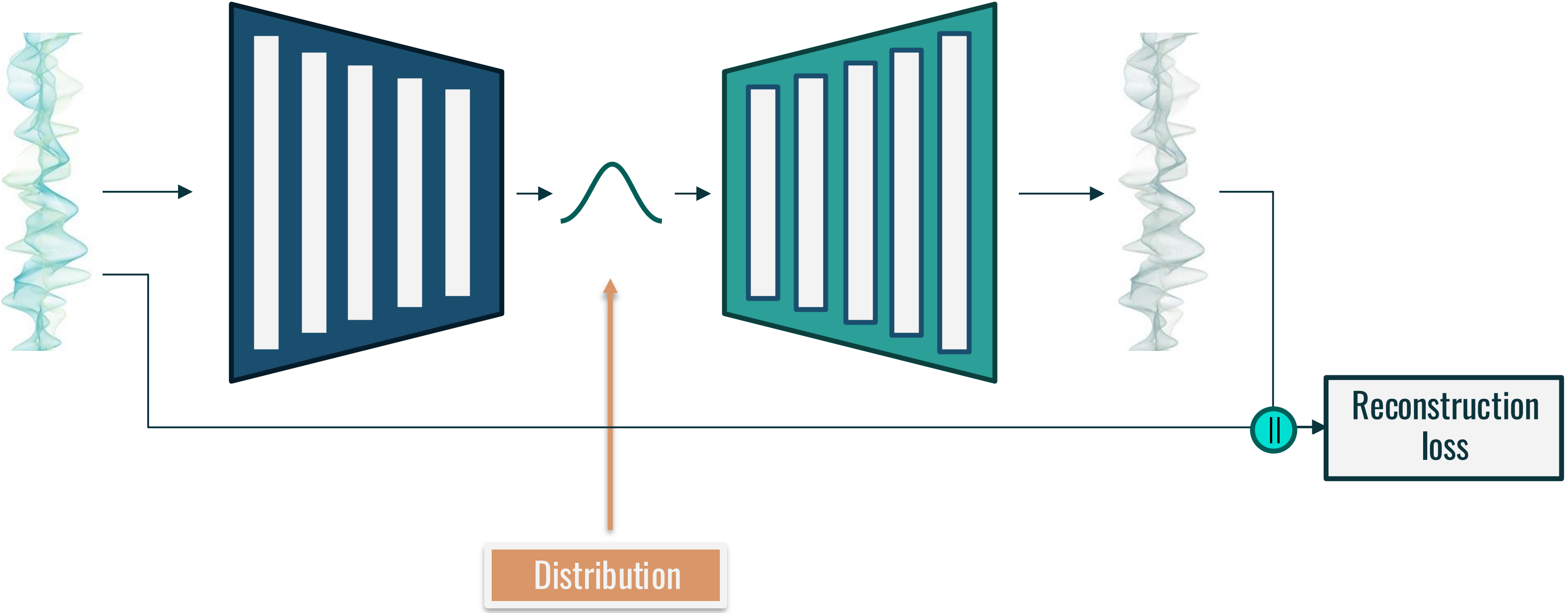
Autoencoder



Autoencoder



Variational autoencoder



Vector quantized VAE

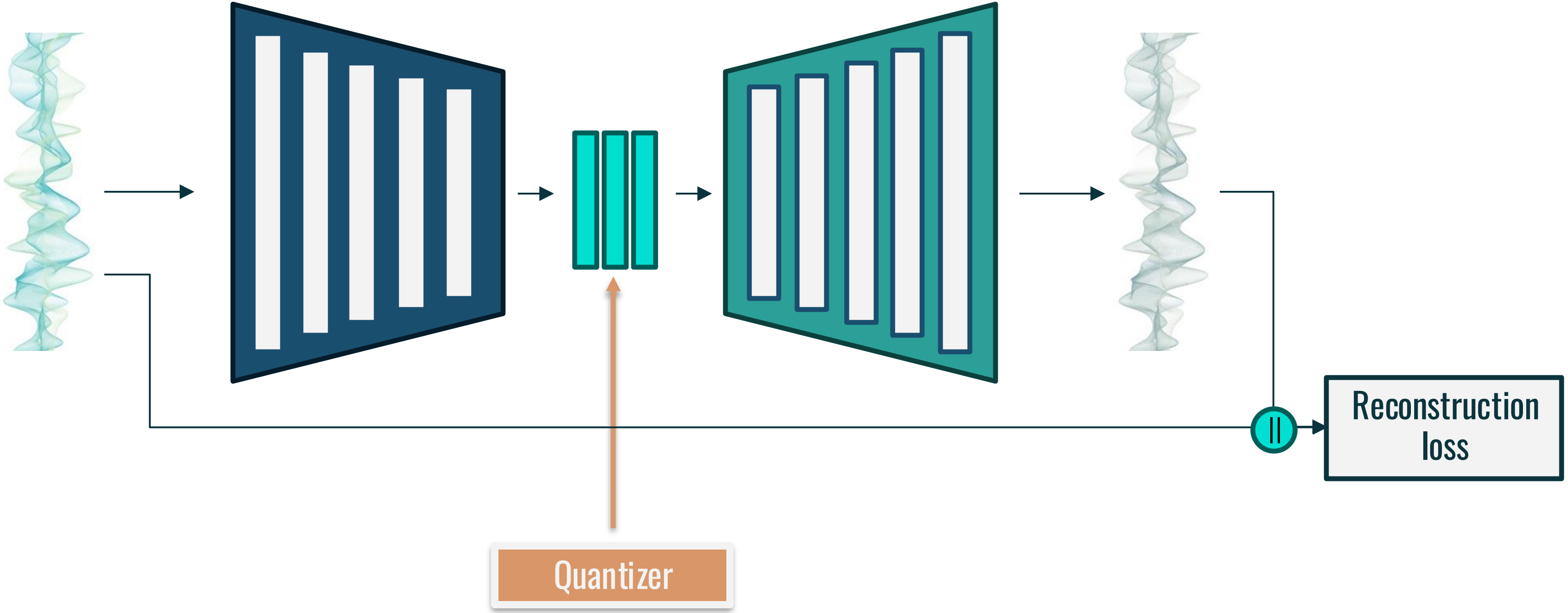


Figure based on: [38] Aaron Van Den Oord, et al. Neural discrete representation learning

Vector quantized VAE

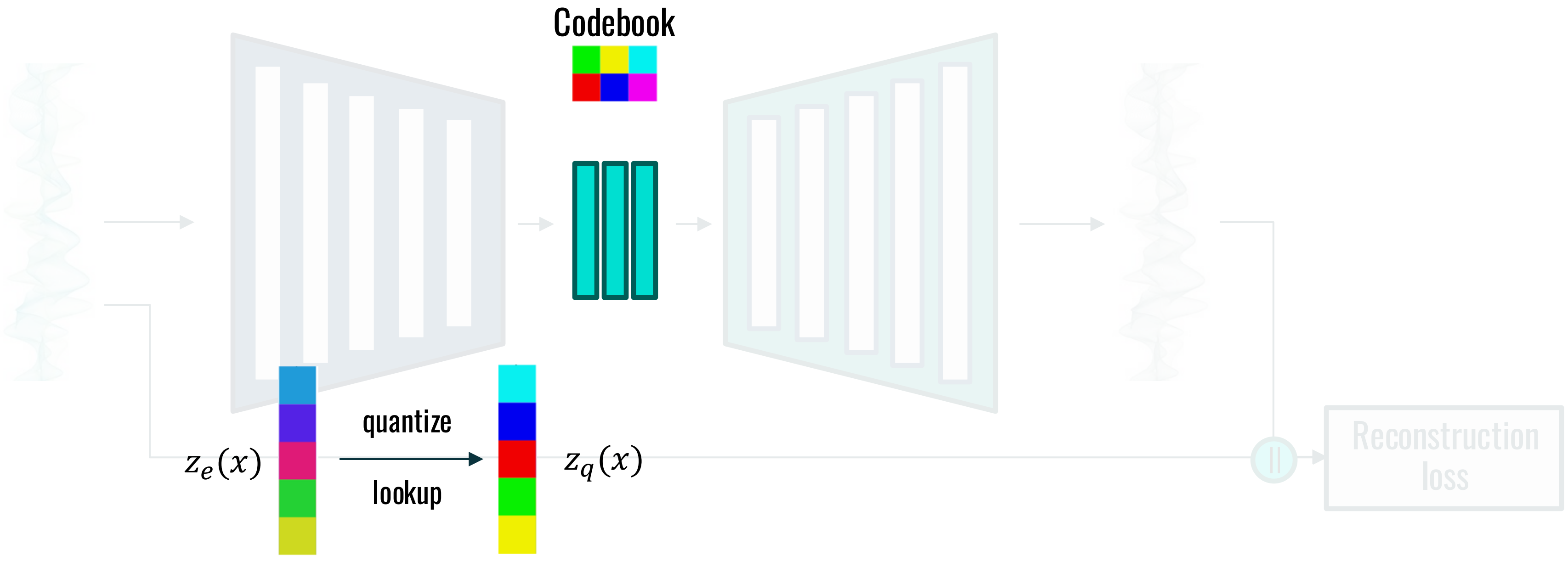


Figure based on: [9] Prafulla Dhariwal, et al. Jukebox: A generative model for music

Residual vector quantized VAE

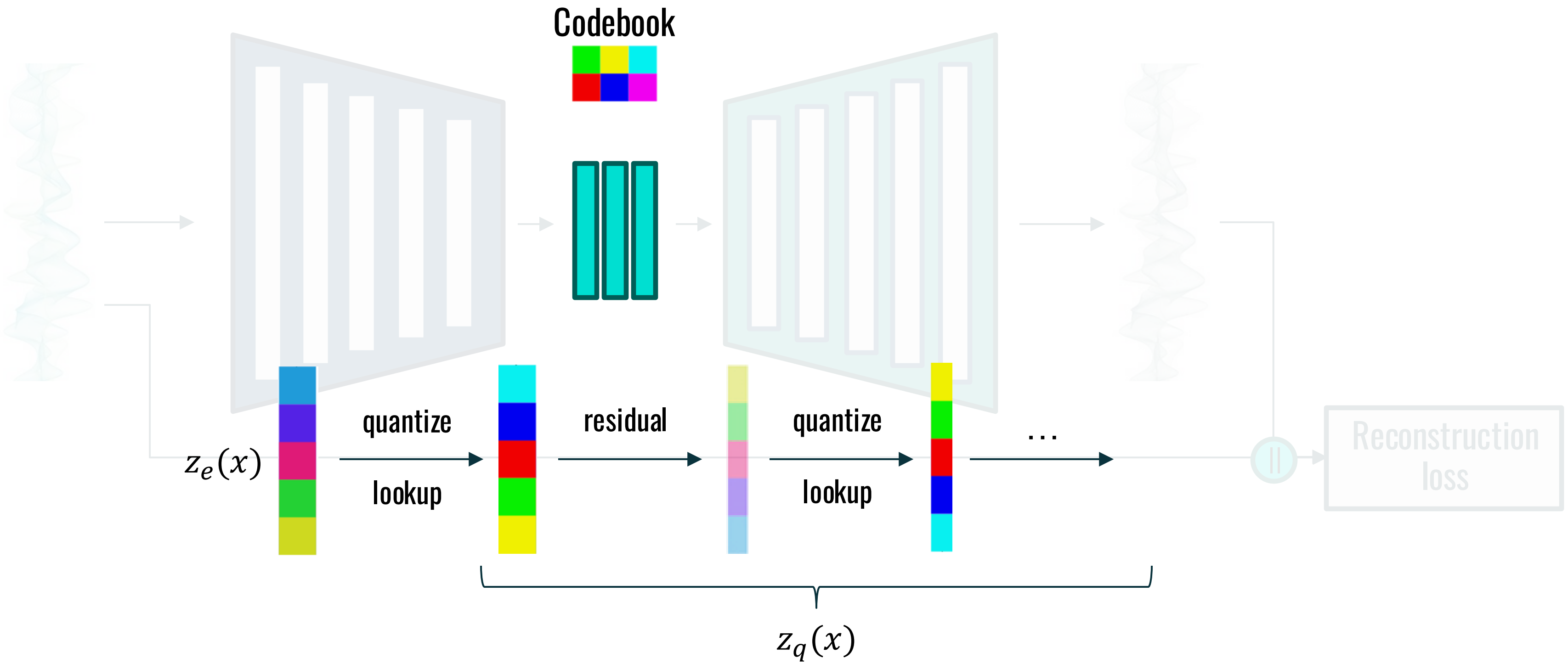


Figure based on: [9] Prafulla Dhariwal, et al. Jukebox: A generative model for music

Bitrate of RVQVAE

Codebook size: 1024 -> 10 bits
Number of quantizers: 1 – 9

→

~86 windows / s
90 bit / window

→

~ 8 kbps

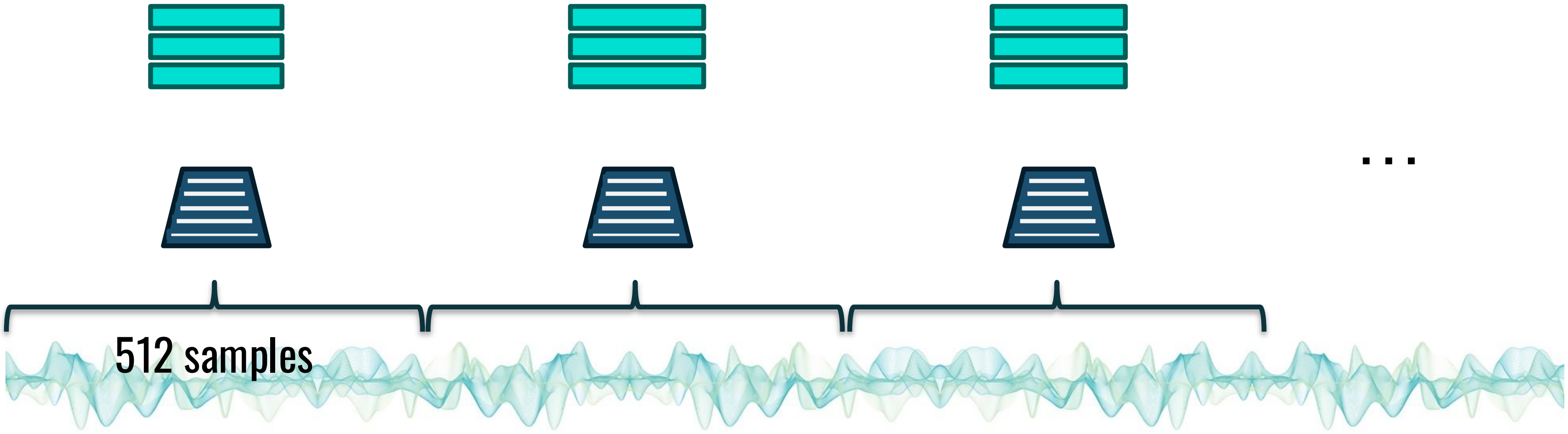
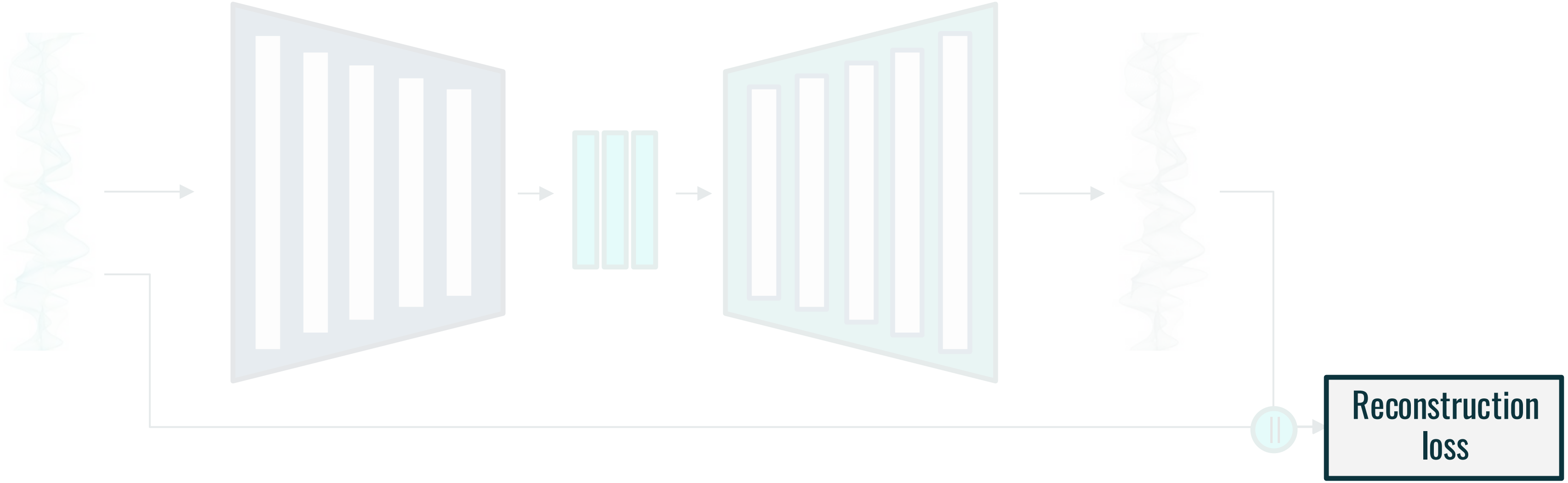


Figure based on: [8] Alexandre Défossez, et al. High fidelity neural audio compression

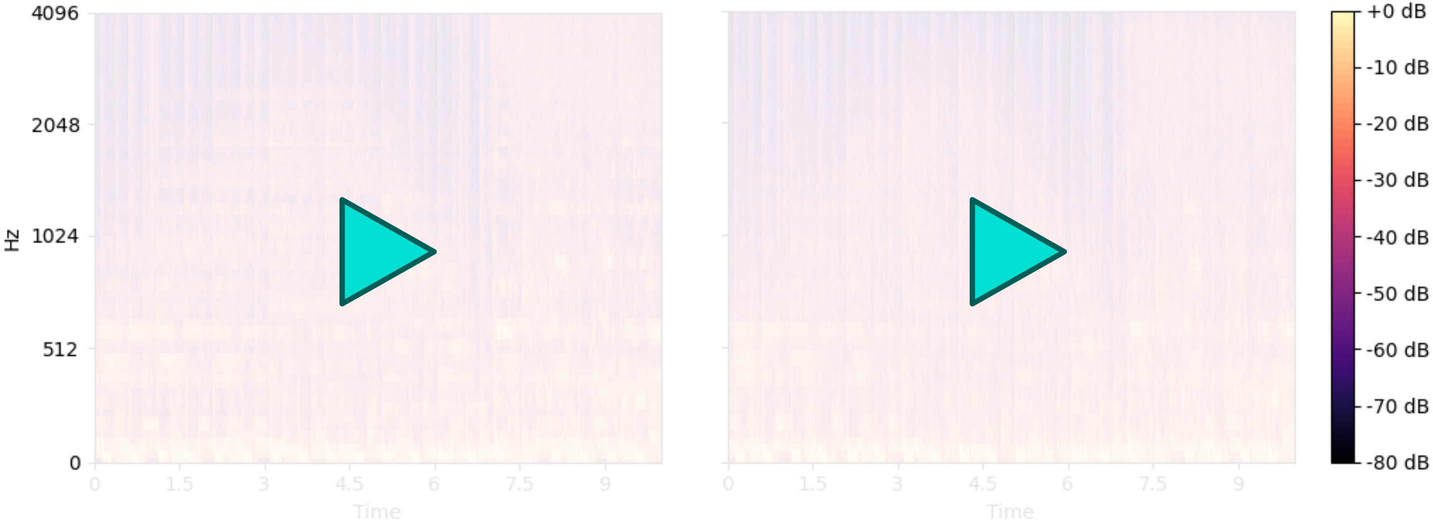
Residual vector quantized VAE



Reconstruction loss

Figure based on: [8] Alexandre Défossez, et al. High fidelity neural audio compression

Problem with simple loss



Residual vector quantized VAE

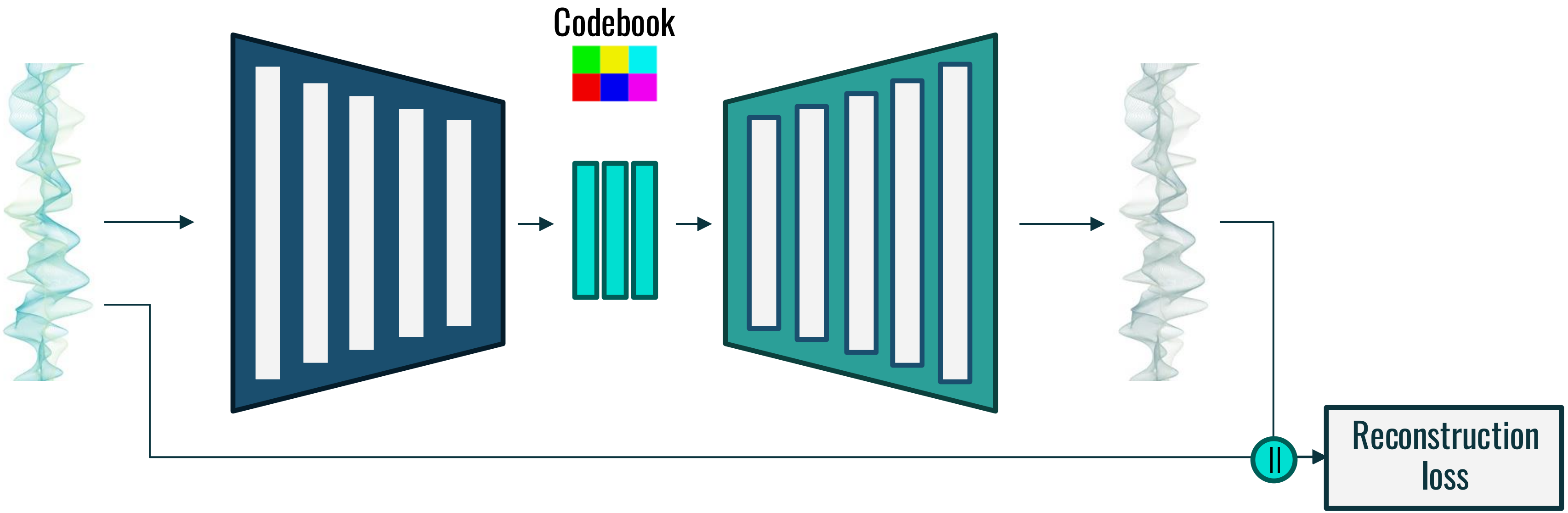


Figure based on: [38] Aaron Van Den Oord, et al. Neural discrete representation learning

Residual vector quantized VAE

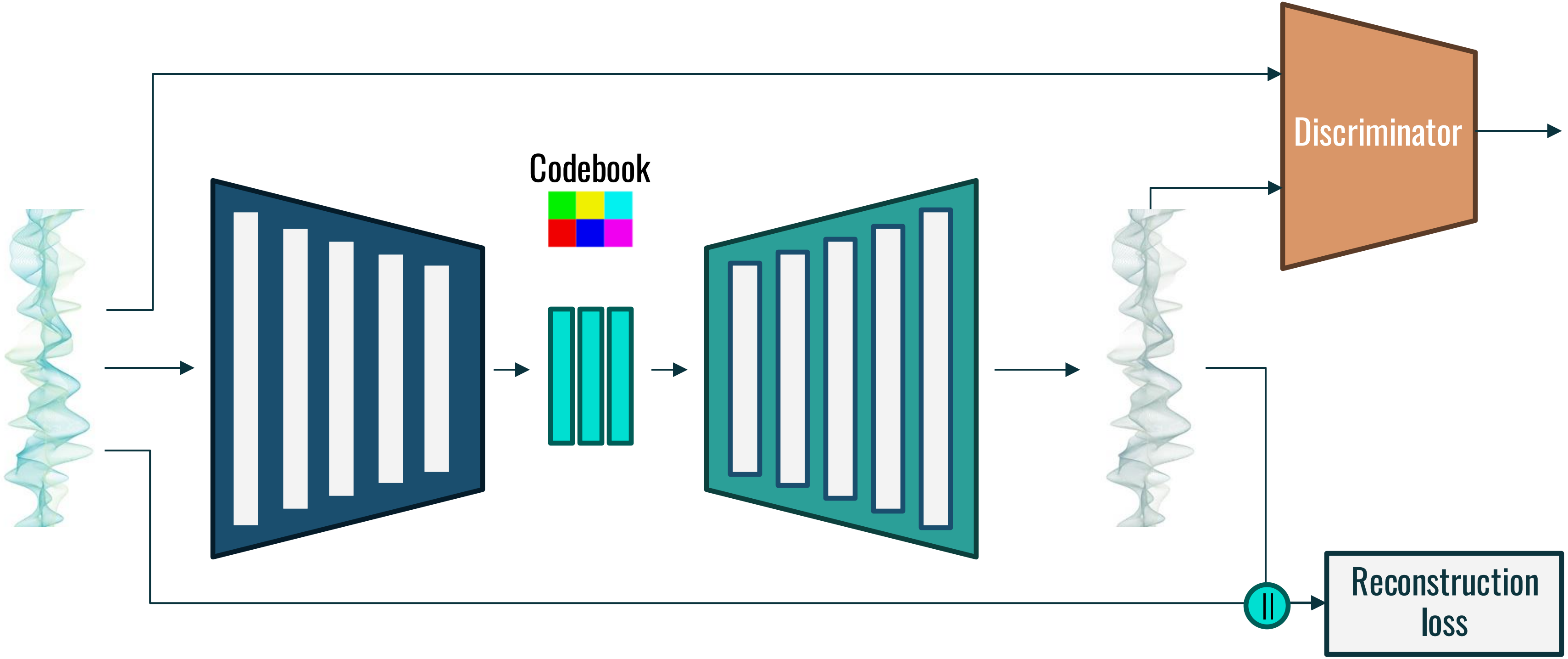
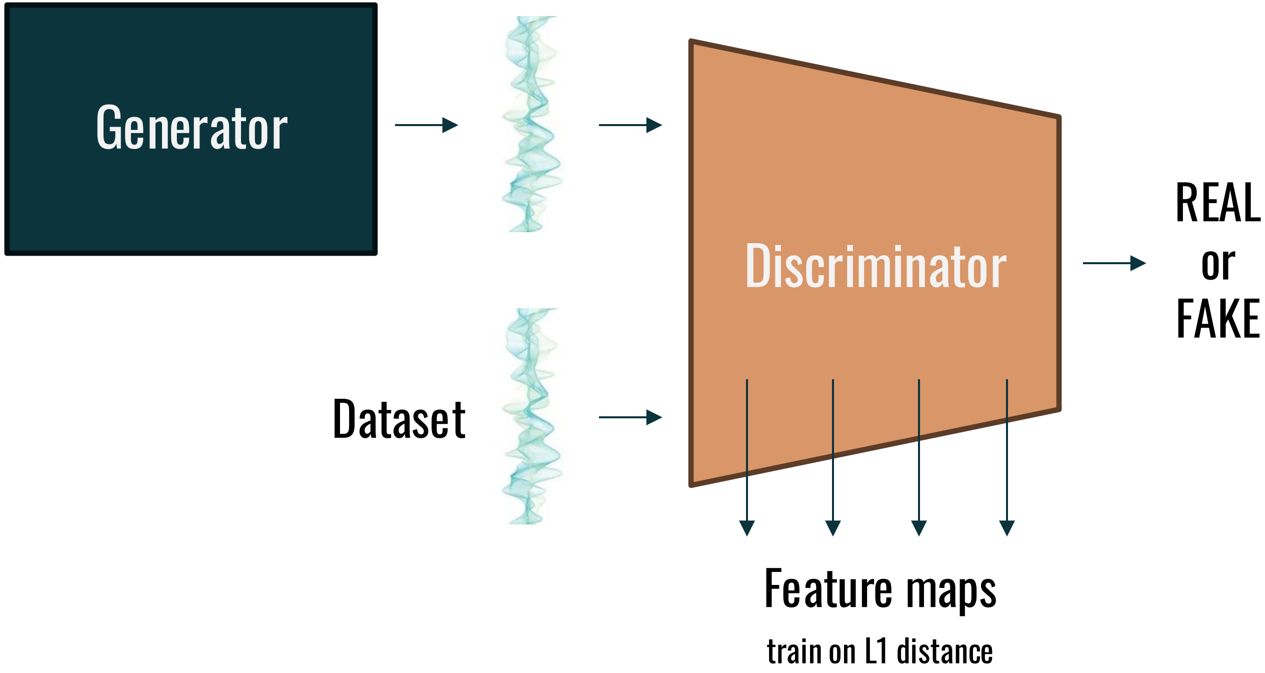


Figure based on: [8] Alexandre Défossez, et al. High fidelity neural audio compression

Generative adversarial networks



Generator

Dataset

Discriminator

REAL
or
FAKE

Feature maps

train on L1 distance

EnCodec

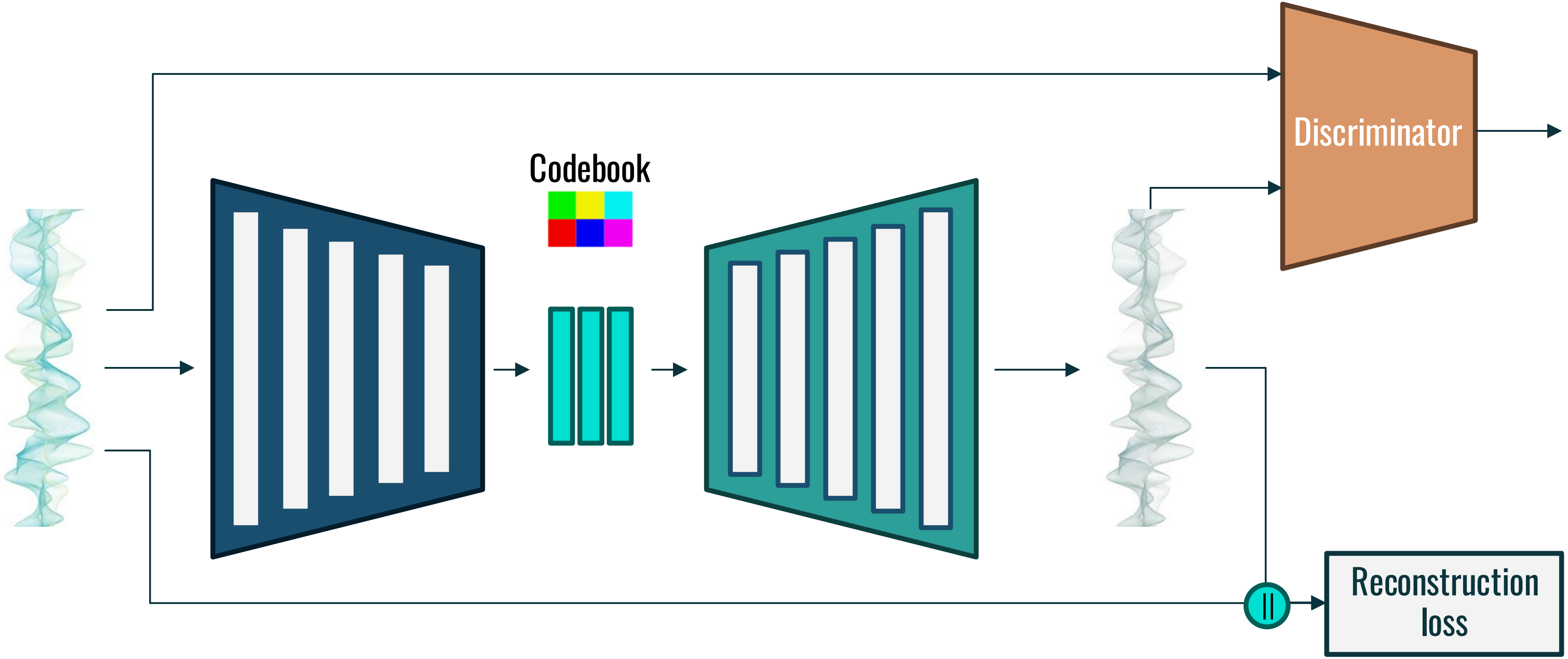


Figure based on: [8] Alexandre Défossez, et al. High fidelity neural audio compression

Improved RVQGAN

1. Periodic activation function

Snake activation function: $snake(x) = x + \frac{1}{\alpha}\sin^2(\alpha x)$

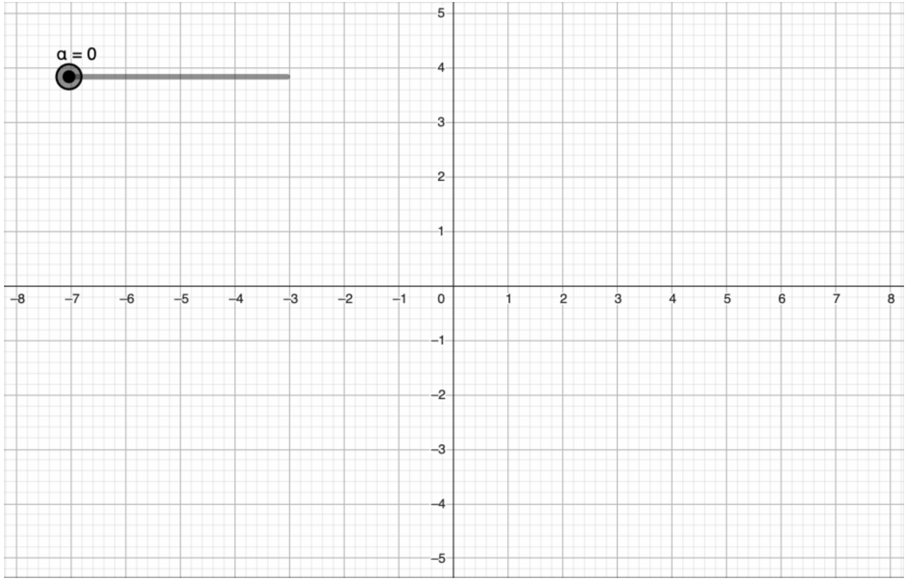
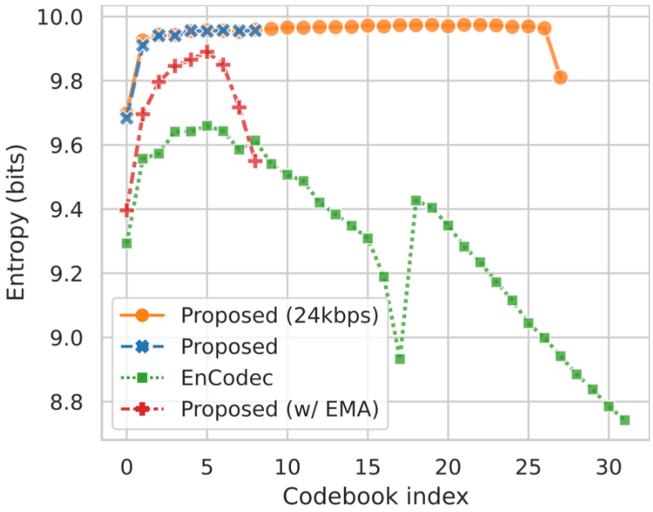


Figure based on: [47] Liu Ziyin, et al. Neural networks fail to learn periodic functions and how to fix it

2. Improved residual vector quantization

Low codebook utilization



Inefficient encoding

Lower quality reconstructions

Source: High-Fidelity Audio Compression with Improved RVQGAN

2. Improved residual vector quantization

K-means clustering to initialize codebook

❌ Randomized restart for underutilization

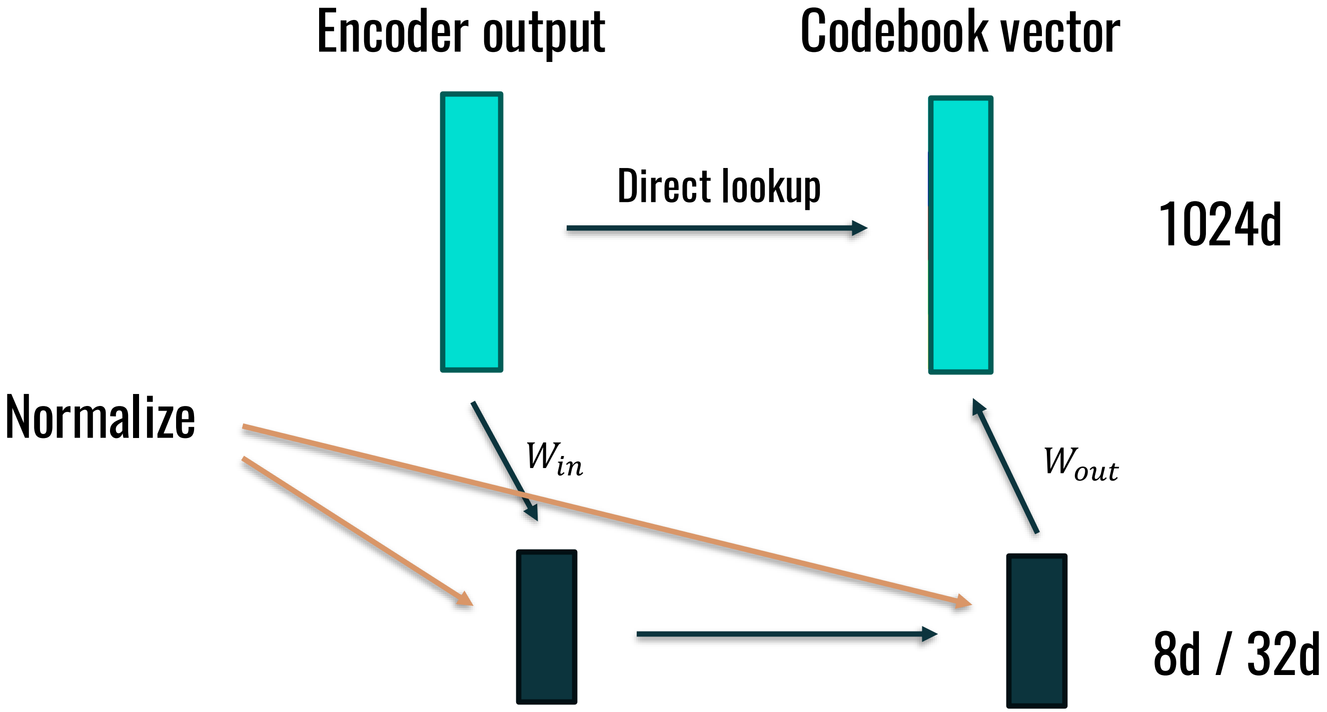
Exponential moving average (EMA)

↓

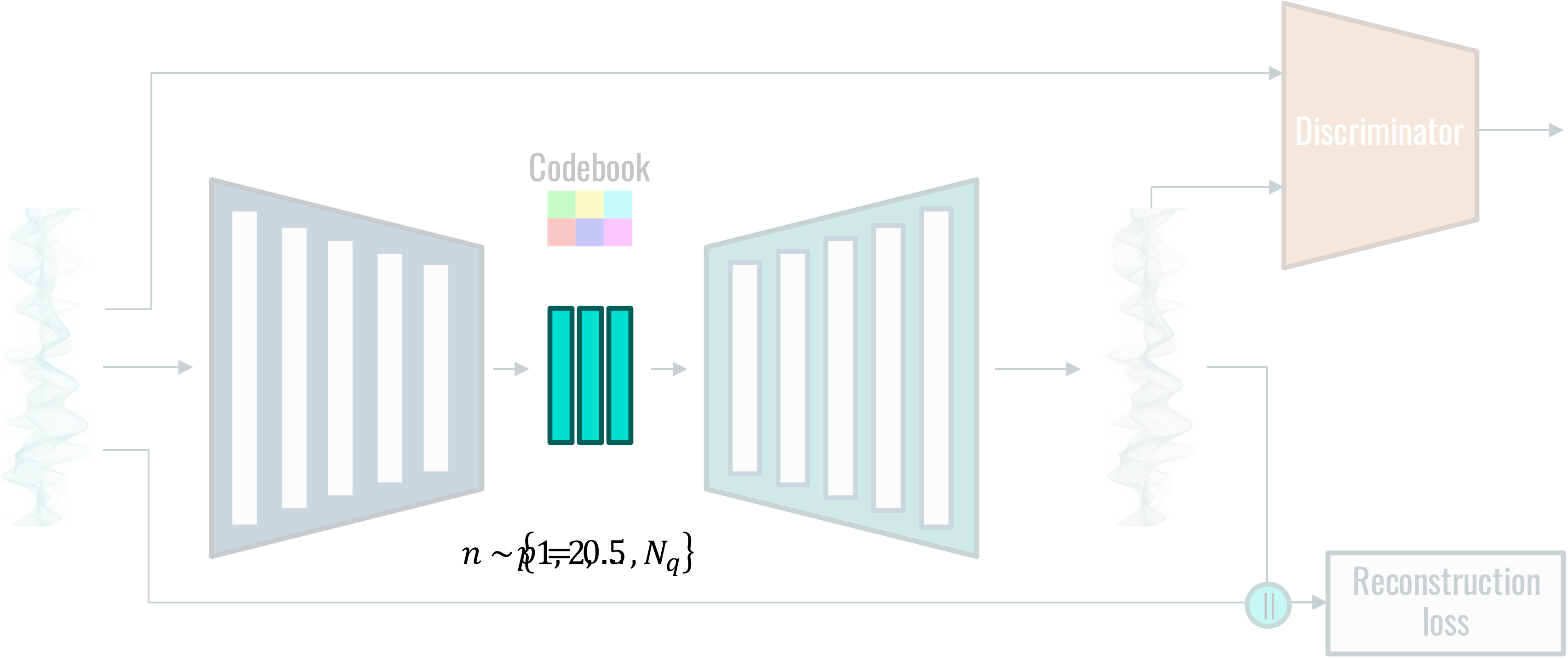
Factorized codes

L2-normalized codes

2. Improved residual vector quantization



3. Quantizer dropout rate



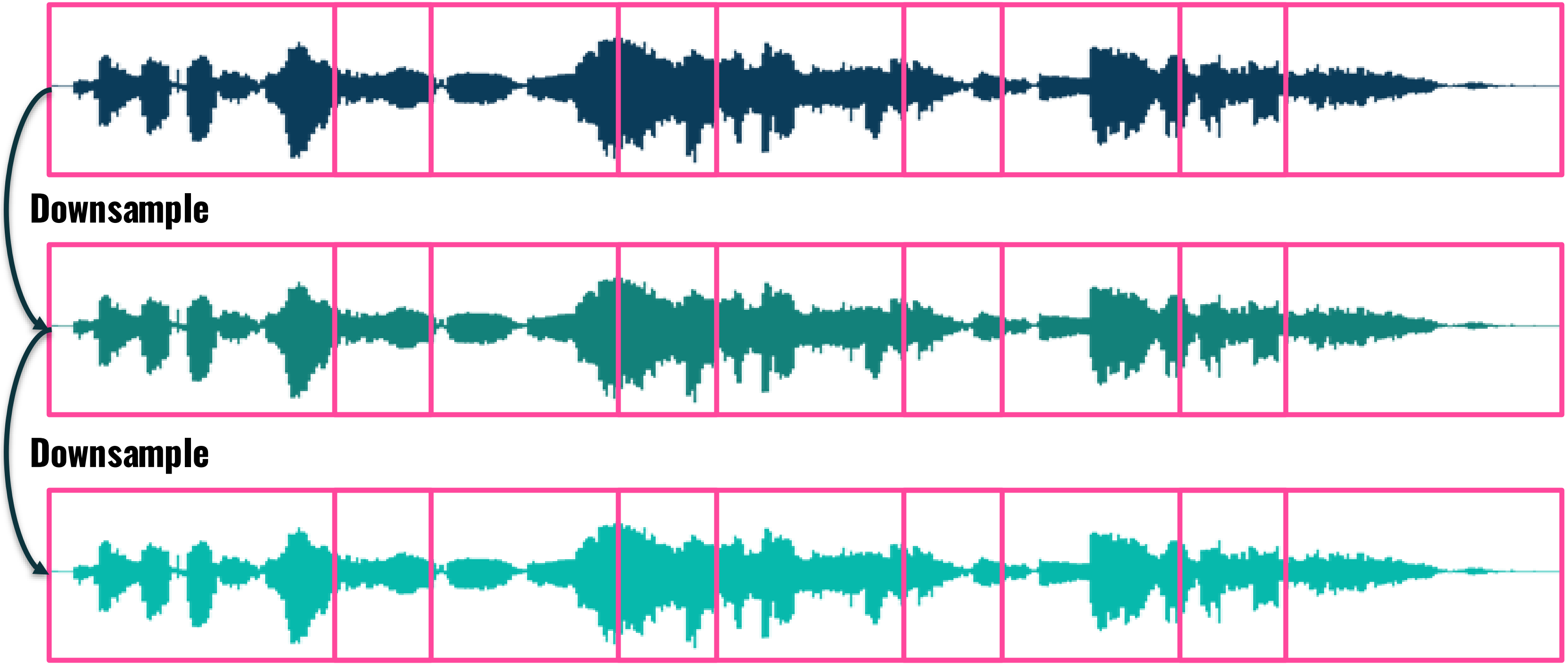
$n \sim \{1, 2, \ldots, N_q\}$

$p = 0.5$

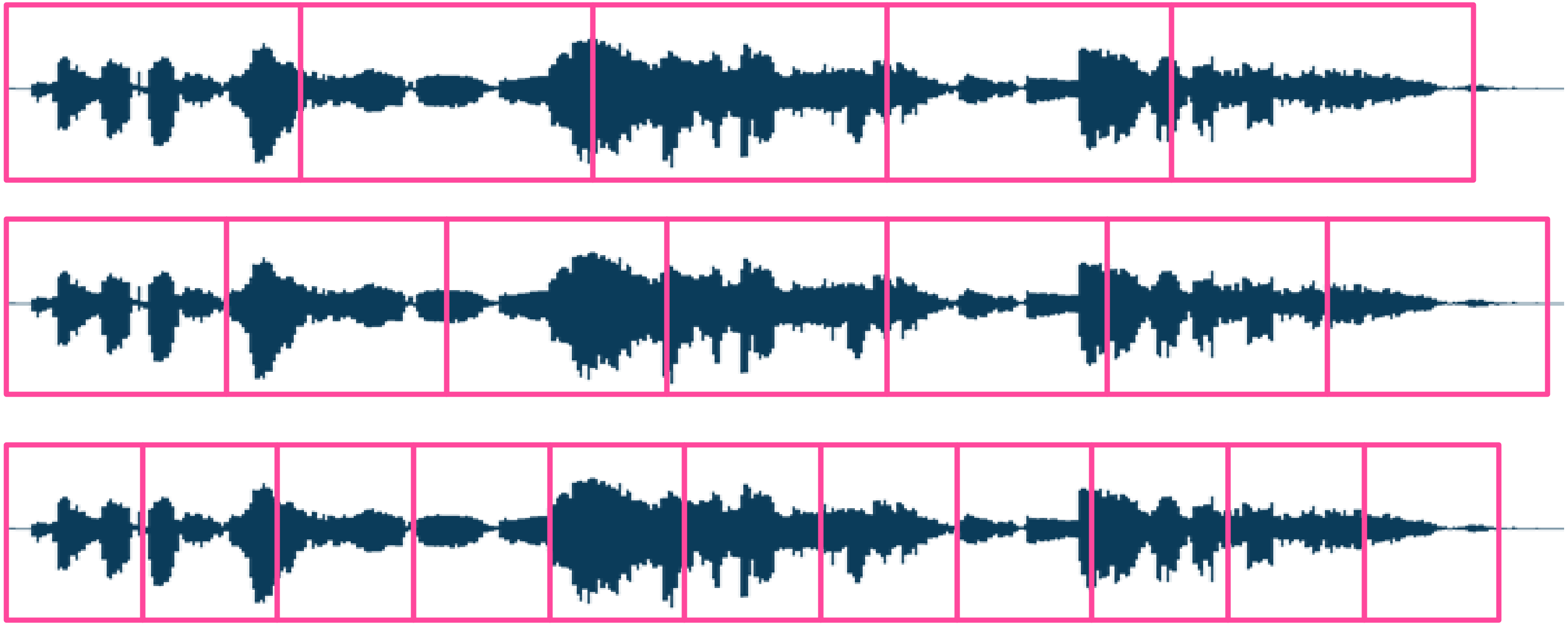
4. Discriminator design

1. Multi-scale discriminator (MSD) -> waveform
2. Multi-period discriminator (MPD) -> waveform
3. Complex short-time Fourier transform (STFT) discriminator at multiple time-scales -> frequency

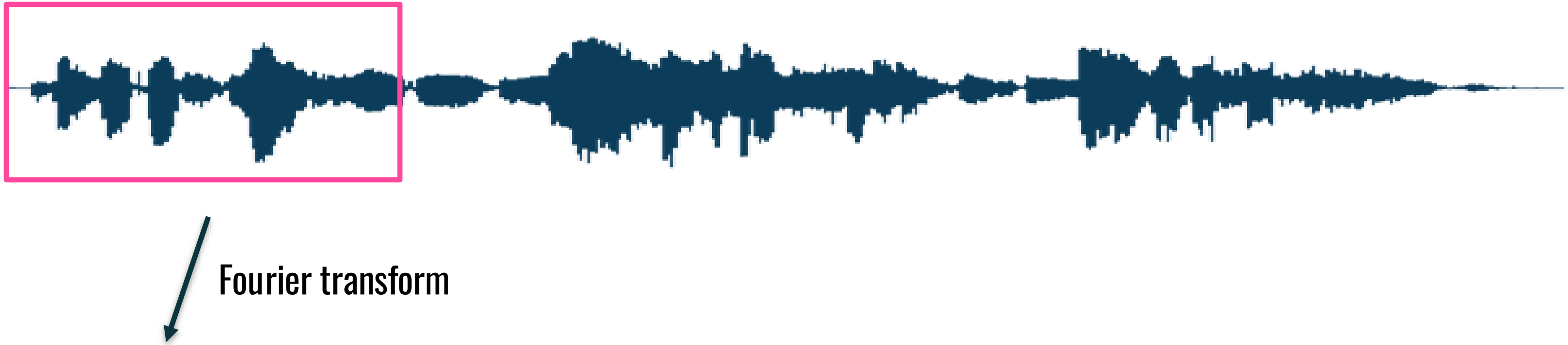
Multi-scale discriminator



Multi-period waveform discriminator



Complex STFT at multiple time-scales



Real part: **frequency**

Imaginary part: **phase**

5. Loss functions

Reconstruction

- Mel-reconstruction loss
- Multi-scale spectral losses

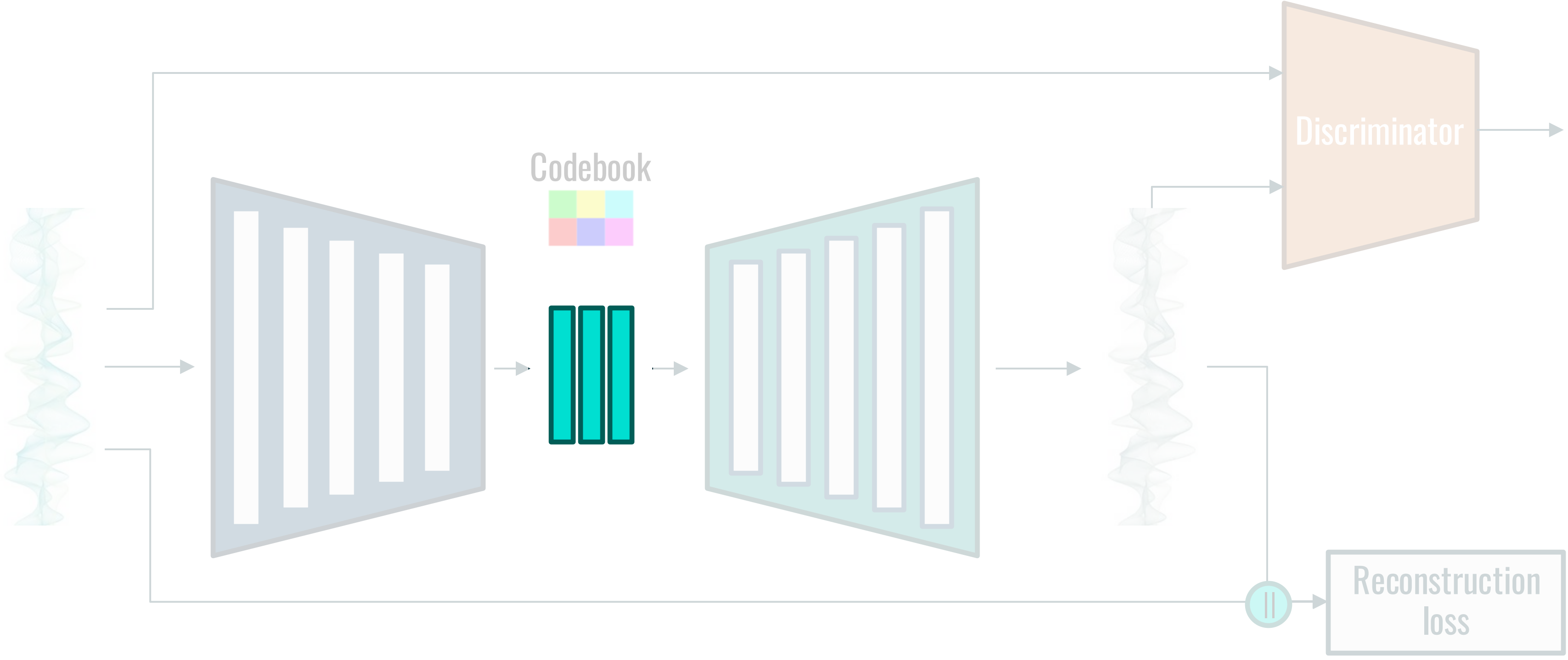
Codebook learning

- Codebook loss
- Commitment loss

Adversarial

- Multi-scale discriminator
- Multi-period discriminator
- Multi-band multi-scale STFT discriminator

RVQGAN



Training

Training data

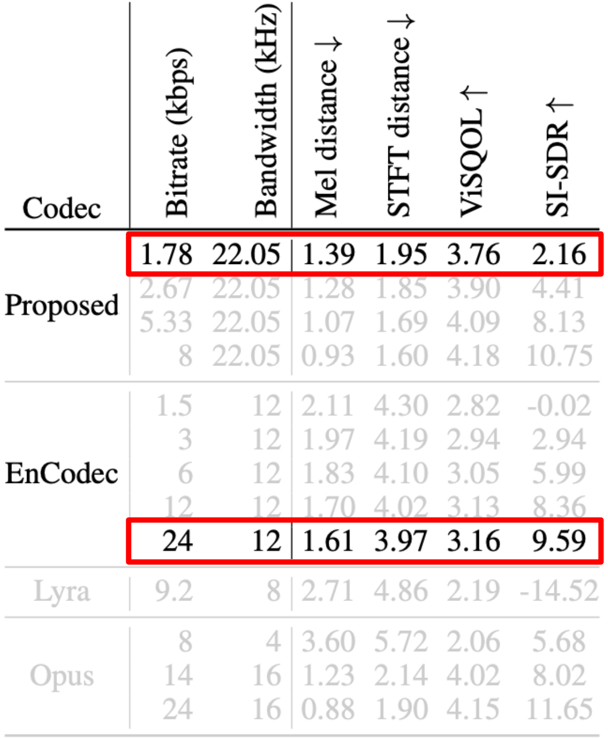
- Speech, music, and environmental sounds
- Balanced data sampling (full-band)

Experiments

Ablation study

- Discriminators
- Mel reconstruction loss
- Latent dimension of codebook
- Quantization setup
- Balanced data sampling

Objective metrics

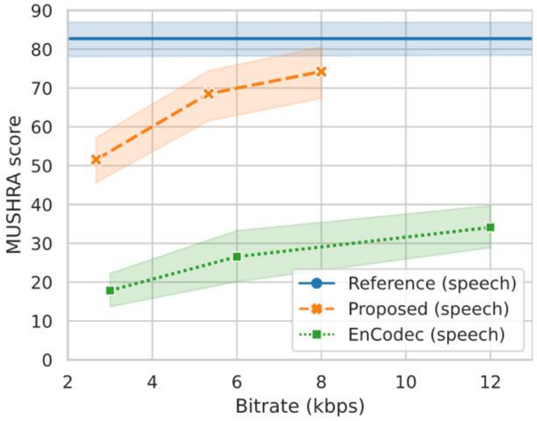


Codec	Bitrate (kbps)	Bandwidth (kHz)	Mel distance ↓	STFT distance ↓	ViSQOL ↑	SI-SDR ↑
Proposed	1.78	22.05	1.39	1.95	3.76	2.16
	2.67	22.05	1.28	1.85	3.90	4.41
	5.33	22.05	1.07	1.69	4.09	8.13
	8	22.05	0.93	1.60	4.18	10.75
EnCodec	1.5	12	2.11	4.30	2.82	-0.02
	3	12	1.97	4.19	2.94	2.94
	6	12	1.83	4.10	3.05	5.99
	12	12	1.70	4.02	3.13	8.36
	24	12	1.61	3.97	3.16	9.59
Lyra	9.2	8	2.71	4.86	2.19	-14.52
Opus	8	4	3.60	5.72	2.06	5.68
	14	16	1.23	2.14	4.02	8.02
	24	16	0.88	1.90	4.15	11.65

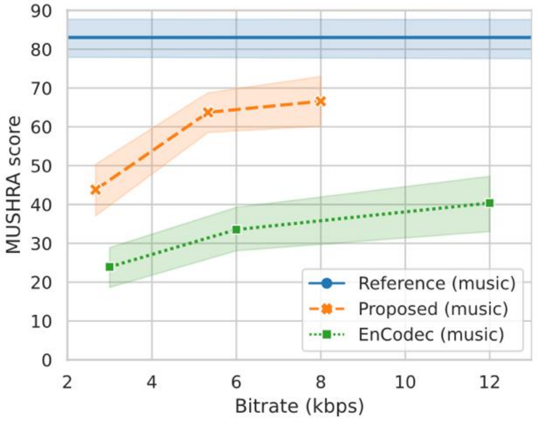
Source: High-Fidelity Audio Compression with Improved RVQGAN

Subjective metrics

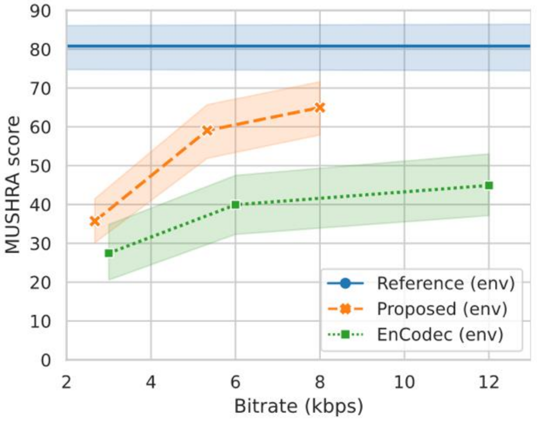
Speech



Music



Environmental s.



MUSHRA = MUltiple Stimuli with Hidden Reference and Anchor

Source: High-Fidelity Audio Compression with Improved RVQGAN

Opinion

Opinion

+ Impressive results, very clearly presented

+ Clean codebase, 1-line usage from command line

+ Focus on new applications (encoding and audio generation)

– Lacking speed test (is it real-time?)

– Streamability

– Hard to compare sampling rates

– EnCodec uses entropy coding -> low codebook utilization is OK

– Reviewers criticize novelty (OpenReview)

Discussion time

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Evaluation metrics

1. **Mel Distance**

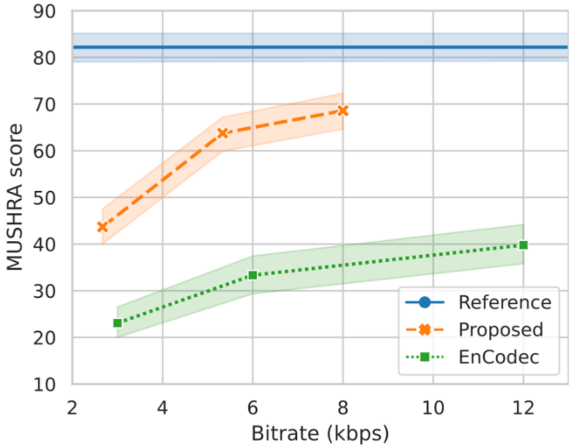
2. **STFT Distance**

3. **ViSQOL (Virtual Speech Quality Objective Listener)**
   -> deep learning model trained on human hearing data to predict Mean Opinion Score

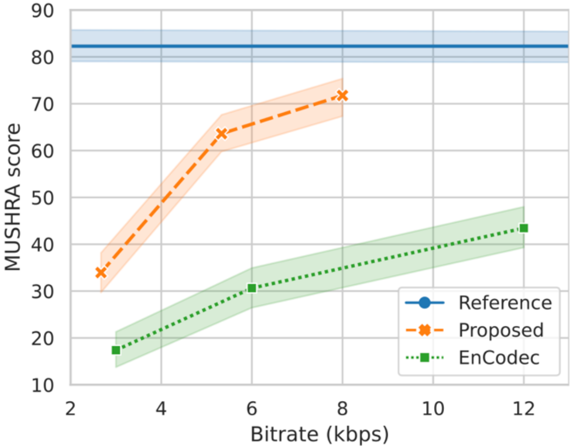
4. **SI-SDR (Scale-Invariant Signal-to-Distortion Ratio)**
   -> similar to signal-to-noise ratio, with modifications so that it is invariant to scale differences, indicates the quality of the phase reconstruction of the audio

Sample rate comparisons

44 kHz



24 kHz



Loss function

$$\mathcal{L}_{\text{VQ}} = ||\text{sg}[\ell_2(z_{\text{proj}}(x))] - \ell_2(e_k)||_2^2 + \beta||\ell_2(z_{\text{proj}}(x)) - \text{sg}[\ell_2(e_k)]||_2^2$$

Reconstruction loss (multi-scale mel, multi-scale spectral): 15.0

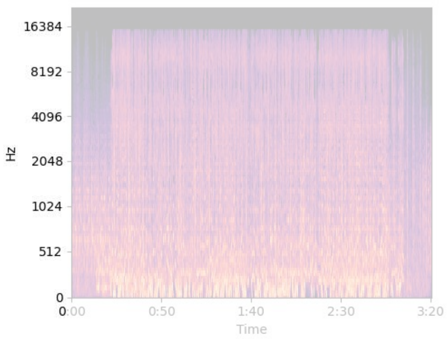
Feature matching loss: 2.0

Adversarial loss: 1.0

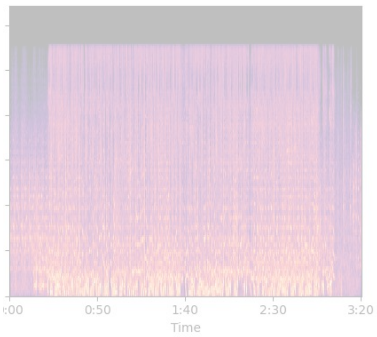
Codebook loss: 1.0

Commitment losses: 0.25

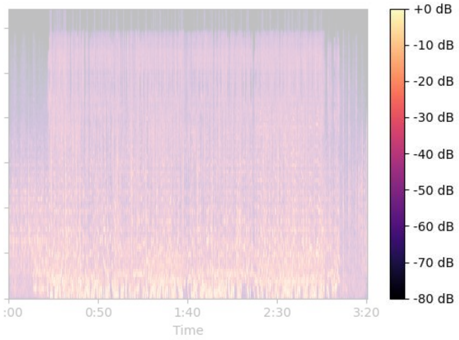
EnCodec demo



Original



EnCodec – 12 kbps



RVQGAN – 8 kbps