Encodec

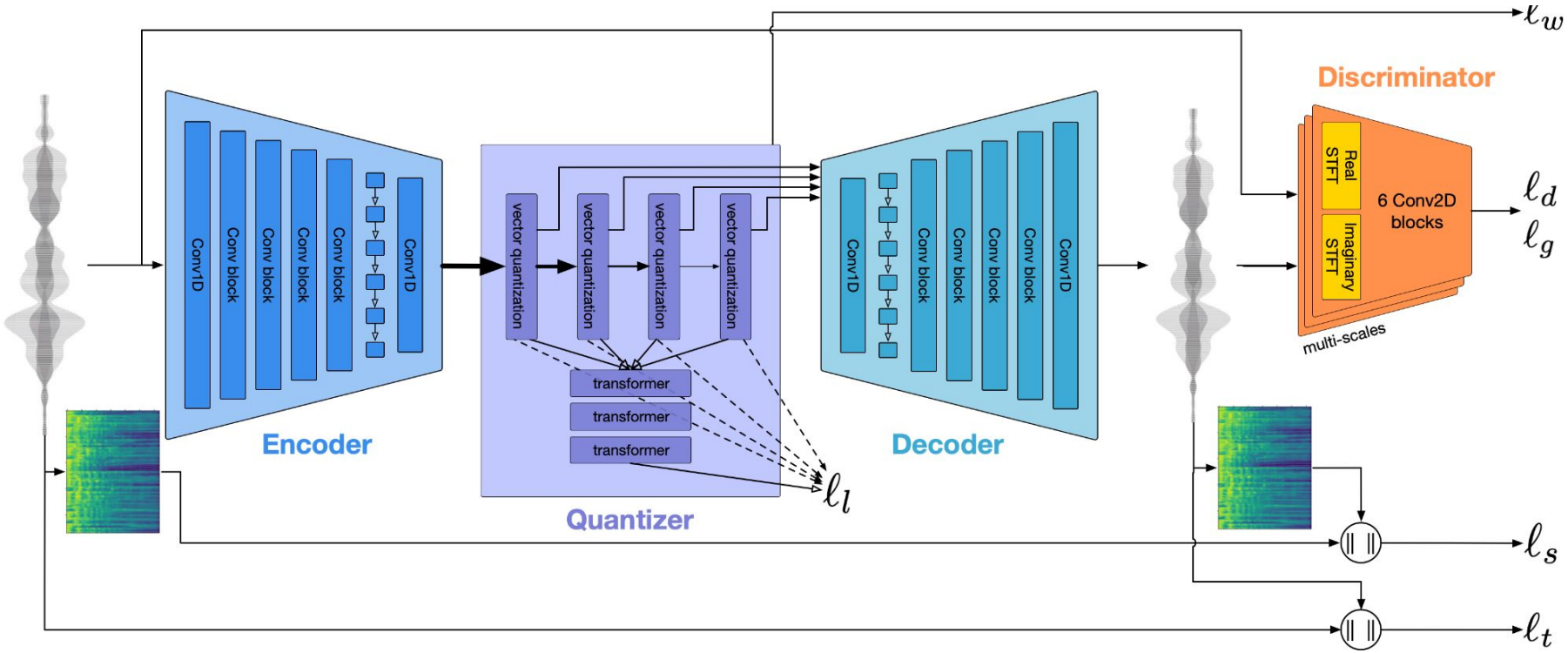
High Fidelity Neural Audio Compression

Кодеки

MP3. Это формат сжатия с потерями, который значительно уменьшает размер файла за счет удаления звуковой информации, которую человеческое ухо не воспринимает.

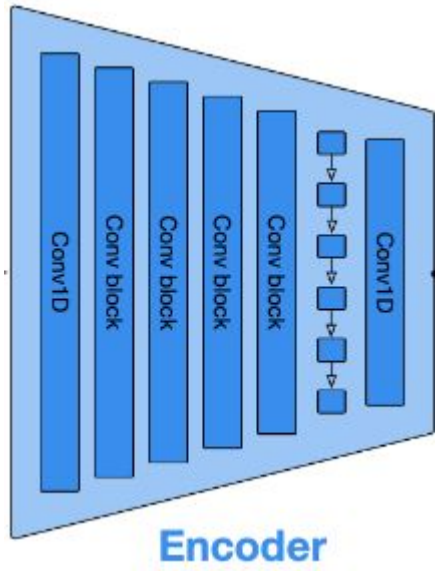
FLAC. Это формат без потерь, который сжимает файлы без утраты качества.

Encodec


ℓ_w
Discriminator
Real STFT
Imaginary STFT
6 Conv2D blocks
multi-scales
ℓ_d
ℓ_g
Conv1D
Conv block
Conv1D
Encoder
vector quantization
transformer
Quantizer
ℓ_l
Decoder
ℓ_s
ℓ_t

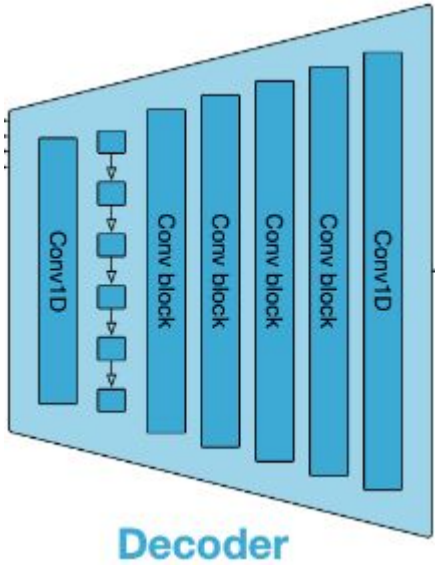
Encoder

- Состоит из 1D сверток со skip-connection и stride.
- Каждый блок сжимает данные за счет stride, но при этом увеличивает число каналов.
- Предпоследний слой двухслойный LSTM



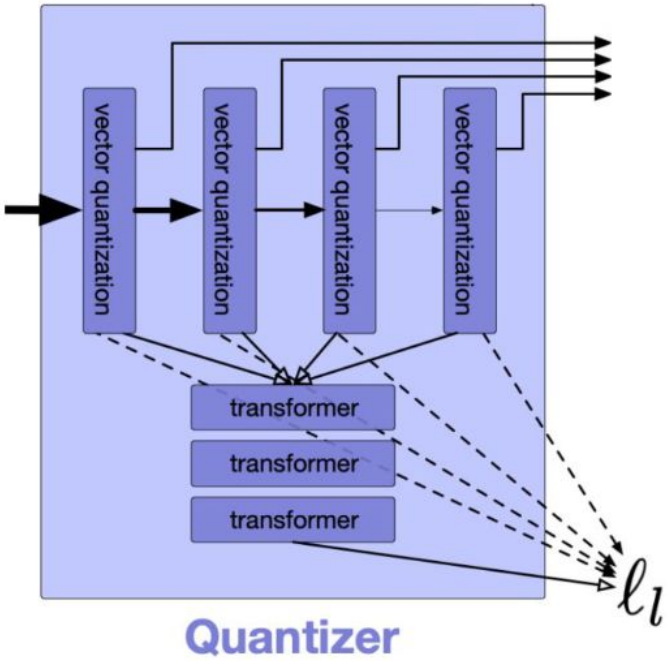
Decoder

- Зеркальное отображение Encoder, за исключением того что используется transposed 1D свертки.



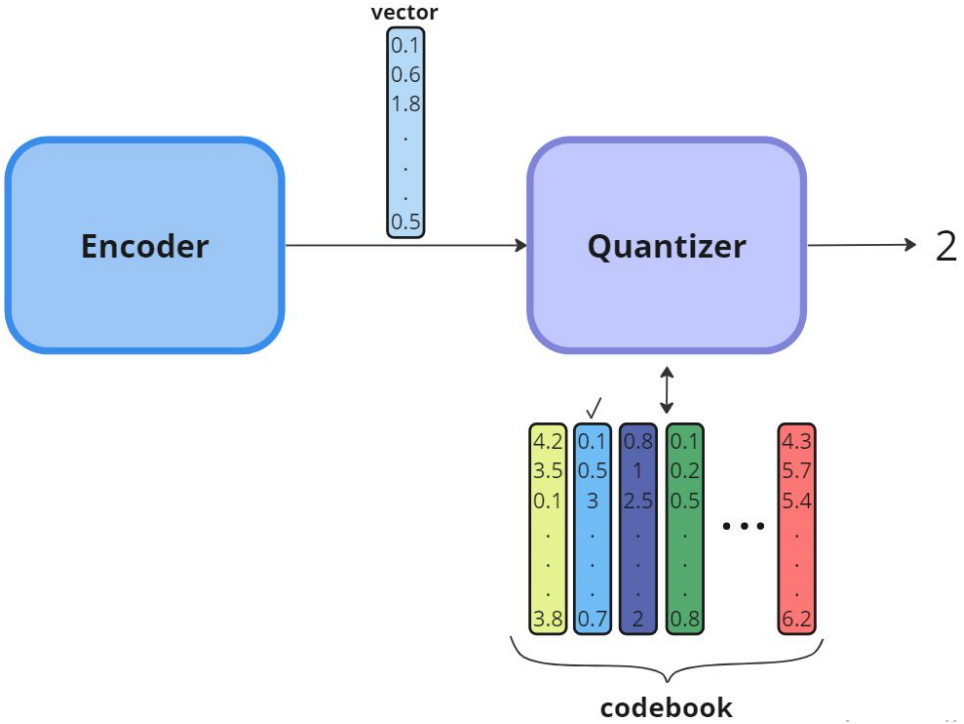
Quantizer

- Еще больше сжимает информацию, с помощью остаточной векторной квантизации, и арифметического кодирования.

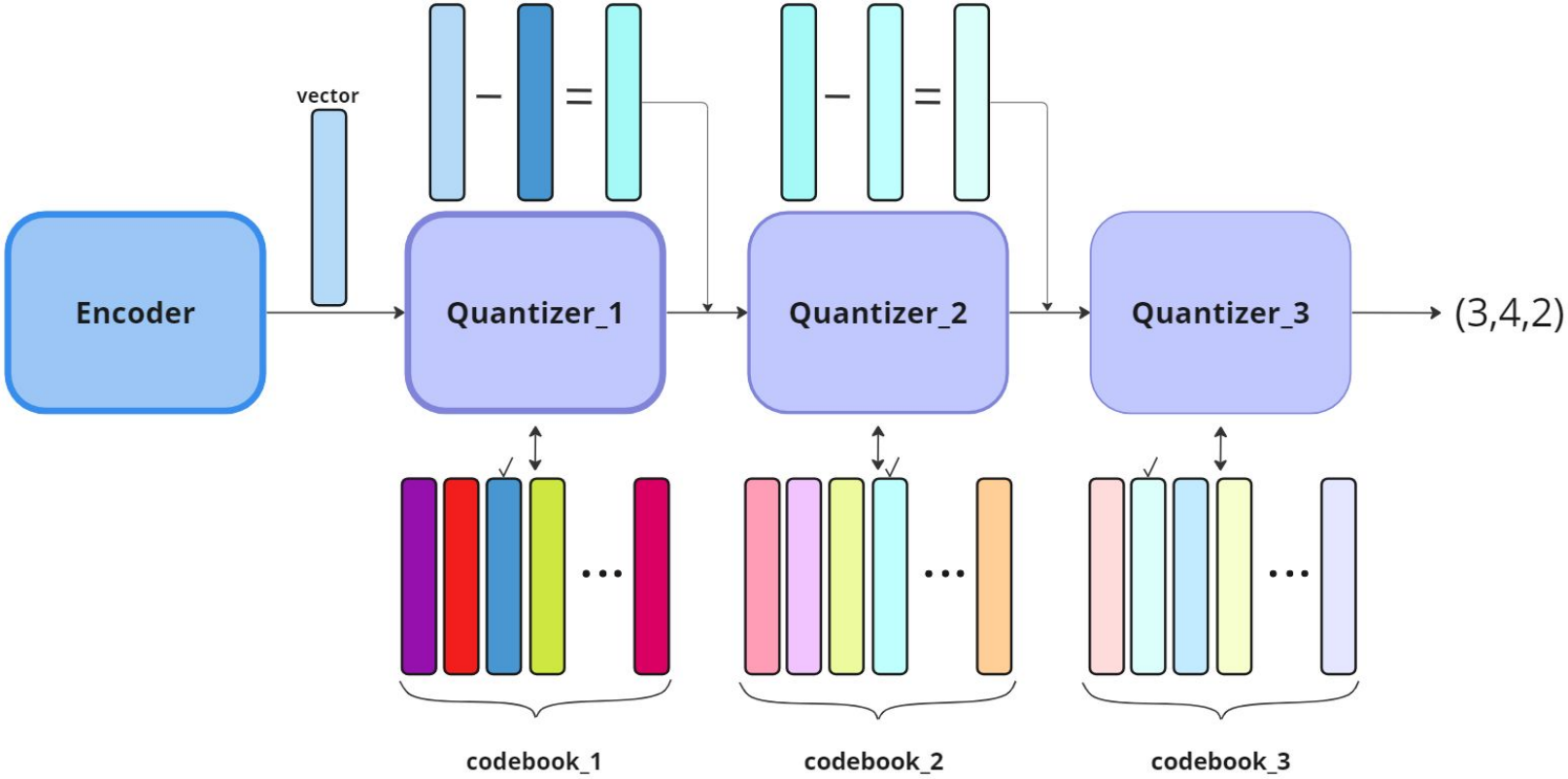


Vector Quantization

- Сопоставляет вектору пришедшему на вход, ближайший вектор по модулю разности из codebook, и возвращает его номер.
- Не работает ввиду большой размерности vector.



Residual Vector Quantization



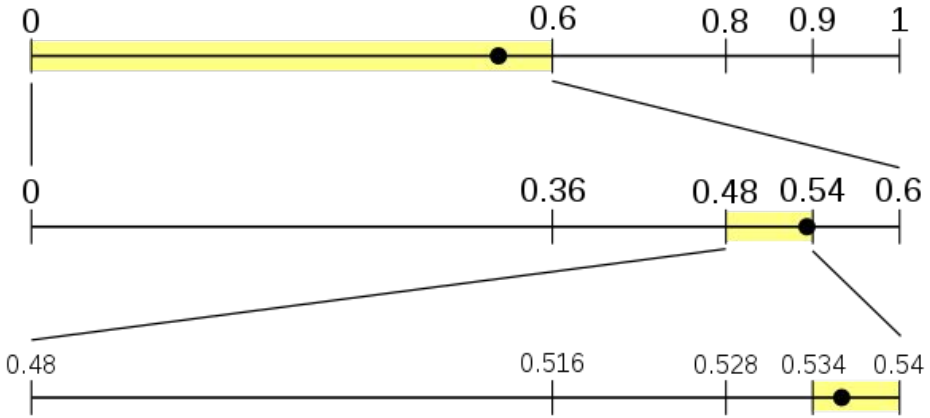
Arithmetic coder

Пусть у нас есть алфавит {A,B,C,D},и мы знаем их распределение.

P(A) = 0.6 , P(B) = 0.2

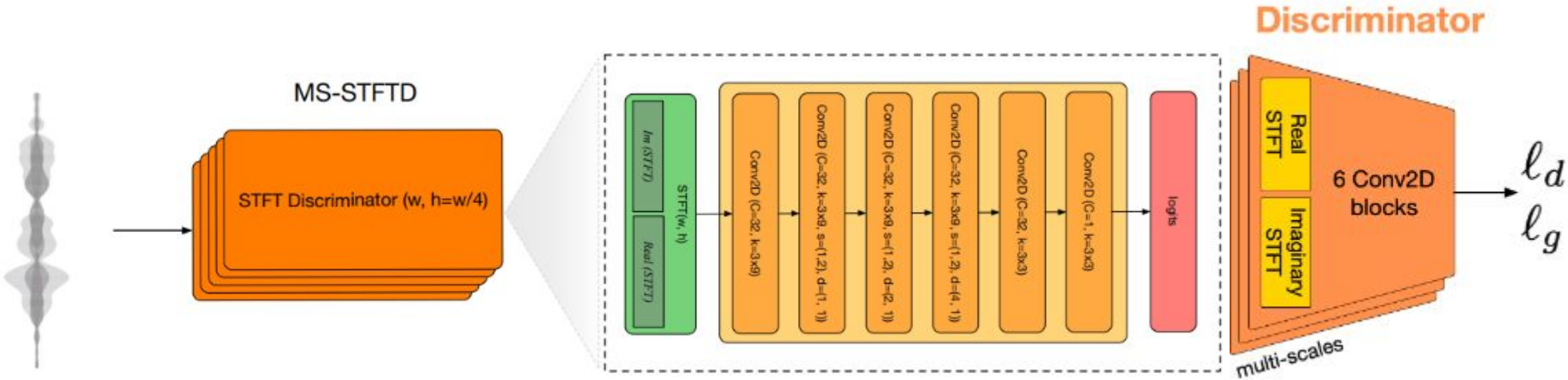
P(C) = P(D) = 0.1

На примере закодирована последовательность ACD, числом 0.538



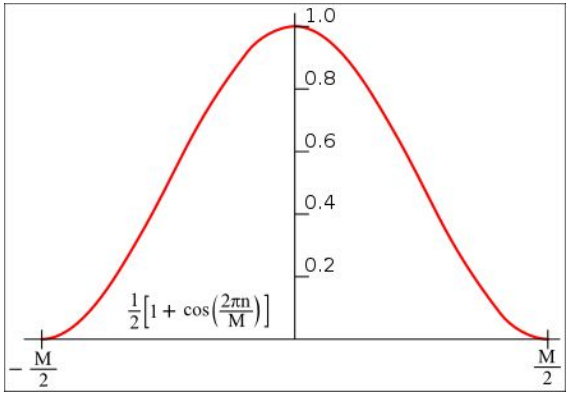
Discriminator

- Предсказывает является аудио, оригинальным или сгенерированным.

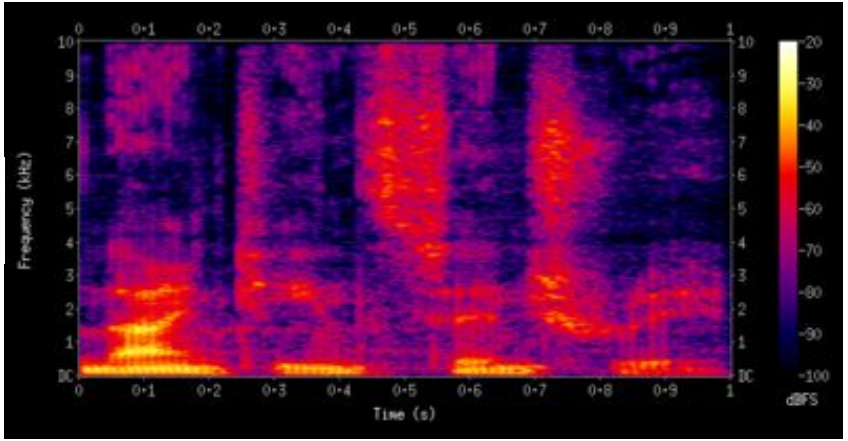


STFT

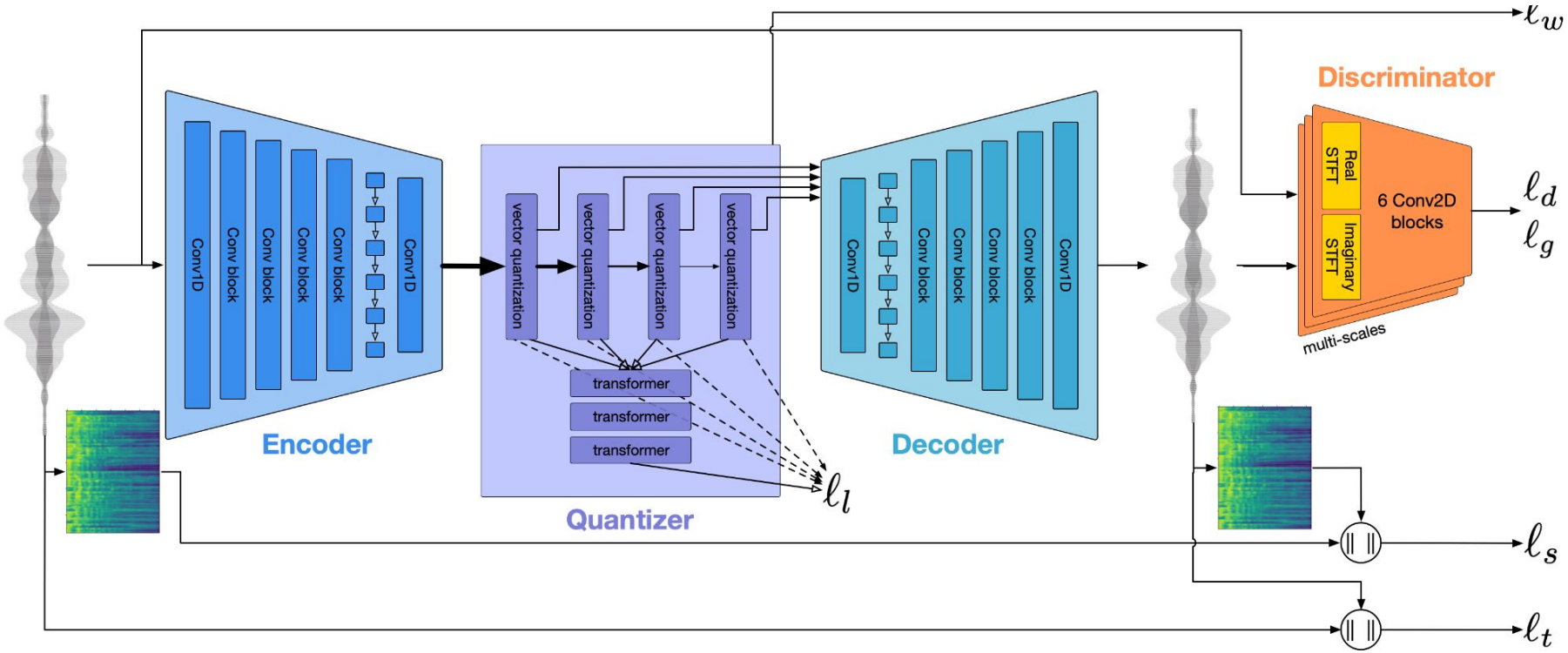
$$\mathbf{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i\omega n}$$



$$\text{spectrogram}\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2$$



Encodec


ℓ_w
Discriminator
Real STFT
Imaginary STFT
6 Conv2D blocks
multi-scales
ℓ_d
ℓ_g
Conv1D
Conv block
Conv1D
Encoder
vector quantization
transformer
Quantizer
ℓ_l
Decoder
ℓ_s
ℓ_t

Reconstruction Loss

Лосс по дискретизации

$$\ell_t(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \left\| \boldsymbol{x} - \hat{\boldsymbol{x}} \right\|_1$$

Лосс по спектрограммам

$$\ell_f(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{|\alpha| \cdot |s|} \sum_{\alpha_i \in \alpha} \sum_{i \in e} \|\mathcal{S}_i(\boldsymbol{x}) - \mathcal{S}_i(\hat{\boldsymbol{x}})\|_1 + \alpha_i \|\mathcal{S}_i(\boldsymbol{x}) - \mathcal{S}_i(\hat{\boldsymbol{x}})\|_2$$

Vector Quantization Loss

Лосс, по которому обучается квантизатор

$$l_w = \sum_{c=1}^{C} \|\boldsymbol{z}_c - q_c(\boldsymbol{z}_c)\|_2^2.$$

Discriminative Loss

Лосс по которому Encoder-Decoder обманывает дискриминатор

$$\ell_g(\hat{\boldsymbol{x}}) = \frac{1}{K}\sum_k \max(0, 1 - D_k(\hat{\boldsymbol{x}}))$$

Лосс по которому дискриминатор не обманывается

$$L_d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{K}\sum_{k=1}^{K} \max(0, 1 - D_k(\boldsymbol{x})) + \max(0, 1 + D_k(\hat{\boldsymbol{x}}))$$

Общий лосс дискриминатора

$$\ell_{feat}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\|D_k^l(\boldsymbol{x}) - D_k^l(\hat{\boldsymbol{x}})\|_1}{\text{mean}\left(\|D_k^l(\boldsymbol{x})\|_1\right)}$$

Final Loss

$$L_G = \lambda_t \cdot \ell_t(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_f \cdot \ell_f(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_g \cdot \ell_g(\hat{\boldsymbol{x}}) + \lambda_{feat} \cdot \ell_{feat}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_w \cdot \ell_w(w)$$

where λ_t, λ_f, λ_g, λ_{feat}, and λ_w the scalar coefficients to balance between the terms.

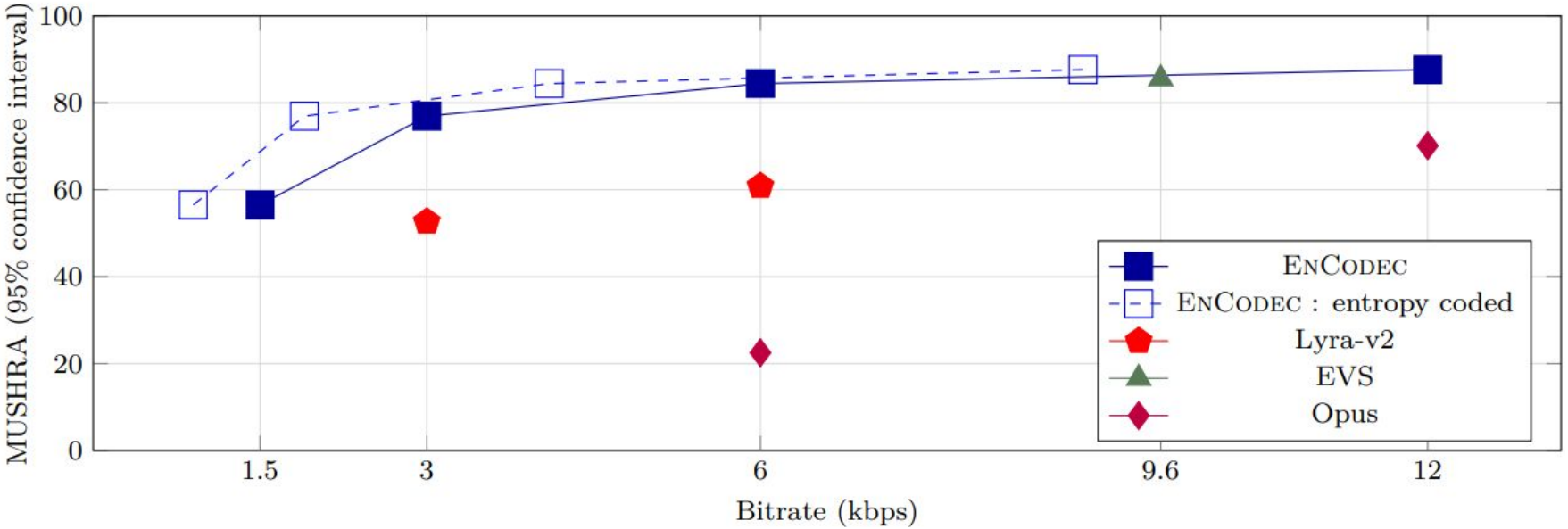
Балансировщик лоссов

$$\tilde{g}_i = R\frac{\lambda_i}{\sum_j \lambda_j} \cdot \frac{g_i}{\langle \|g_i\|_2 \rangle_\beta}$$

$g_i = \frac{\partial \ell_i}{\partial \hat{x}}$ - градиент i-ого лосса

$\langle \|g_i\|_2 \rangle_\beta$ - экспоненциальное скользящее среднее

Эксперименты



MUSHRA

Model	Bandwidth	Entropy Coded	Clean Speech	Noisy Speech	Music Set-1	Music Set-2
Reference	-	-	95.5±1.6	93.9±1.8	93.2±2.5	97.1±1.3
Opus	6.0 kbps	-	30.1±2.8	19.1±5.9	20.6±5.8	17.9±5.3
Opus	12.0 kbps	-	76.5±2.3	61.9±2.1	77.8±3.2	65.4±2.7
EVS	9.6 kbps	-	84.4±2.5	80.0±2.4	89.9±2.3	87.7±2.3
Lyra-v2	3.0 kbps	-	53.1±1.9	52.0±4.7	69.3±3.3	42.3±3.5
Lyra-v2	6.0 kbps	-	66.2±2.9	59.9±3.3	75.7±2.6	48.6±2.1
EnCodec	1.5 kbps	0.9 kbps	49.2±2.4	41.3±3.6	68.2±2.2	66.5±2.3
EnCodec	3.0 kbps	1.9 kbps	67.0±1.5	62.5±2.3	89.6±3.1	87.8±2.9
EnCodec	6.0 kbps	4.1 kbps	83.1±2.7	69.4±2.3	92.9±1.8	91.3±2.1
EnCodec	12.0 kbps	8.9 kbps	90.6±2.6	80.1±2.5	91.8±2.5	92.9±1.2

Результат

Model	Streamable	SI-SNR	ViSQOL
Opus	✓	2.45	2.60
EVS	✓	1.89	2.74
EnCodec	✓	6.67	4.35
EnCodec	✗	7.46	4.39

Источники:

https://github.com/facebookresearch/encodec/tree/main/encodec

https://en.wikipedia.org/wiki/Short-time_Fourier_transform

https://en.wikipedia.org/wiki/Arithmetic_coding