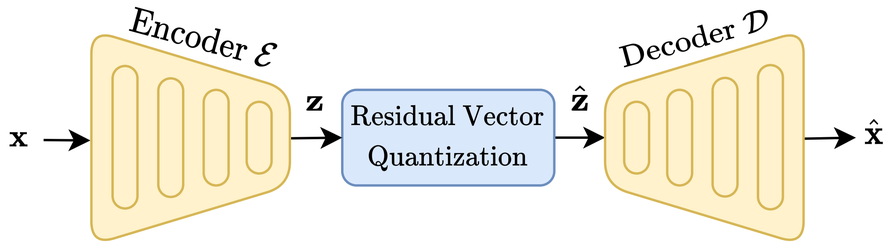
Post-Training Latent Dimension Reduction in Neural Audio Coding

Thomas Muller[1, 2], Stéphane Ragot[1], Pierrick Philippe[1] and Pascal Scalart[2]

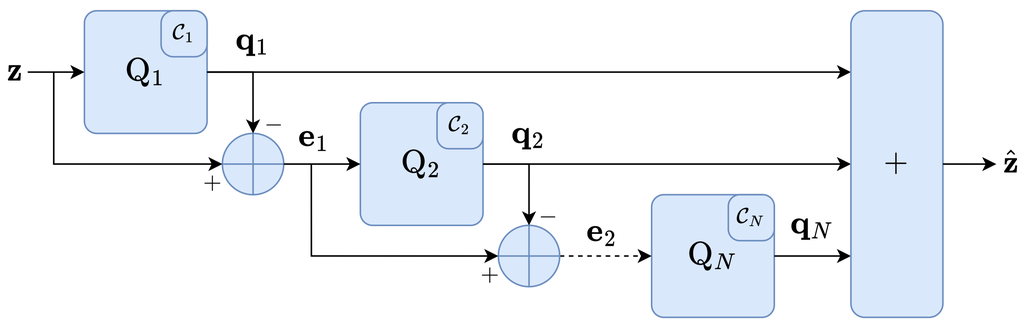
[1]Orange Innovation, France [2]IRISA – University of Rennes, France

1. Neural Audio Coding

Context and motivations: A new generation of audio codecs has emerged using deep learning. Neural audio codecs such as SoundStream or EnCodec demonstrate promising audio quality at low bitrates at the cost of higher computational complexity compared to traditional audio codecs.



Most neural audio codecs use Residual Vector Quantization (RVQ) whose codebooks are learned during end-to-end model training. The learned codebooks need to be stored, and the nearest neighbor search complexity increases with the number of VQs.

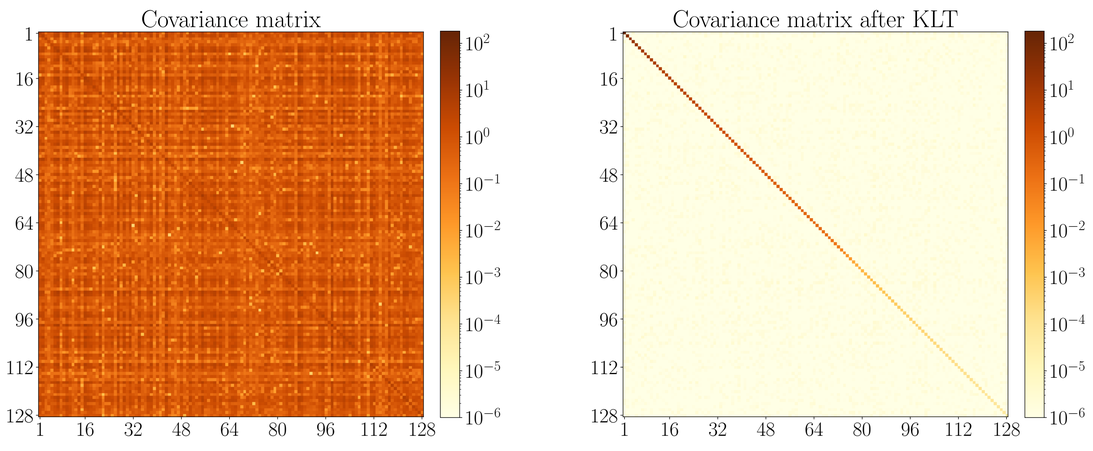


Goal: Optimize Residual Vector Quantization to reduce storage and computation complexity.

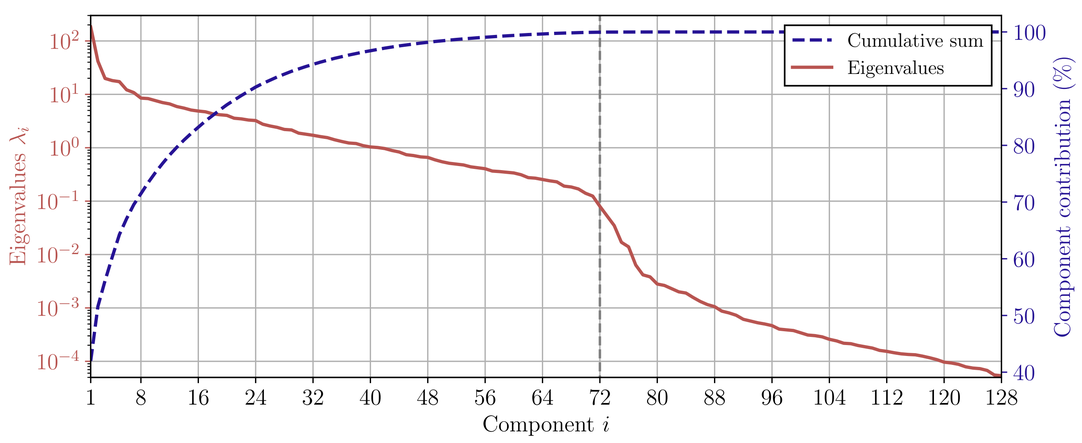
2. Proposed Method – Example of EnCodec

Case study: The neural audio codec EnCodec compresses mono audio sampled at 24 kHz. Each frame $\mathbf{x}$ of 320 samples is converted into a 128-dimension latent vector $\mathbf{z}$ quantized by RVQ. The bitrate goes from 1.5 to 24 kbps (from 2 to $N_{max} = 32$ quantization stages using 10-bit codebooks).

Latent space analysis: Elements of latent vectors $\mathbf{z}$ are highly correlated. The proposed method decorrelates latent space before quantization.



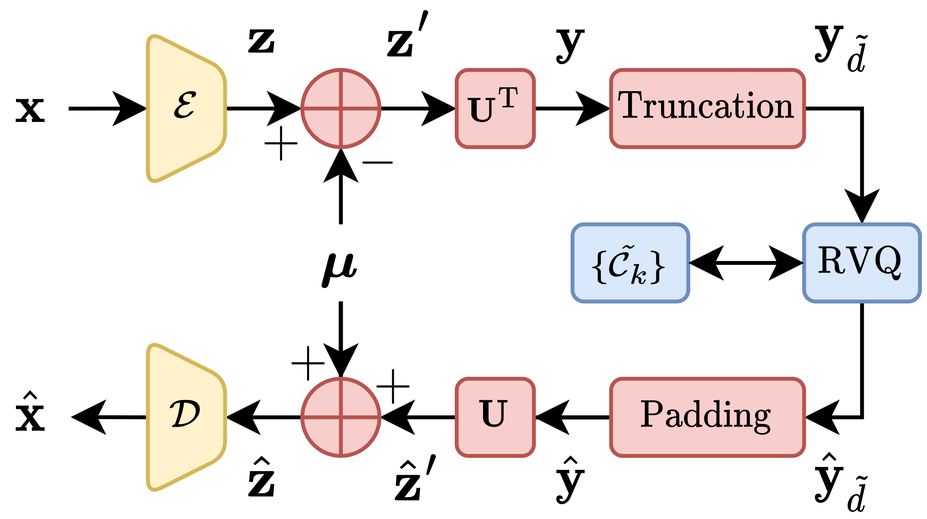
Decorrelation: A Karhunen-Loève Transform (KLT) is performed on latent space, by multiplying the (mean-removed) latent vectors with the eigenvector matrix $\mathbf{U}$ from the eigen decomposition of the covariance matrix $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$. One can observe a drop of eigenvalues for last elements.



Covariance estimation: The covariance matrix $\mathbf{R}$ is estimated using the first N_{cov} codebooks that capture most of the information of the latent distribution.

Proposed RVQ optimization: The method consists in truncating last dimensions in the transformed space (corresponding to smallest eigenvalues). Quantization codebooks are modified as follows:

$$\begin{cases} \tilde{\mathcal{C}}_1 = \left[\mathbf{U}^{\mathrm{T}}(\mathcal{C}_1 - \boldsymbol{\mu})\right]_{1:\tilde{d}} \\ \tilde{\mathcal{C}}_k = \left[\mathbf{U}^{\mathrm{T}}\mathcal{C}_k\right]_{1:\tilde{d}} \quad k = 2, \ldots, N_{max} \end{cases}$$

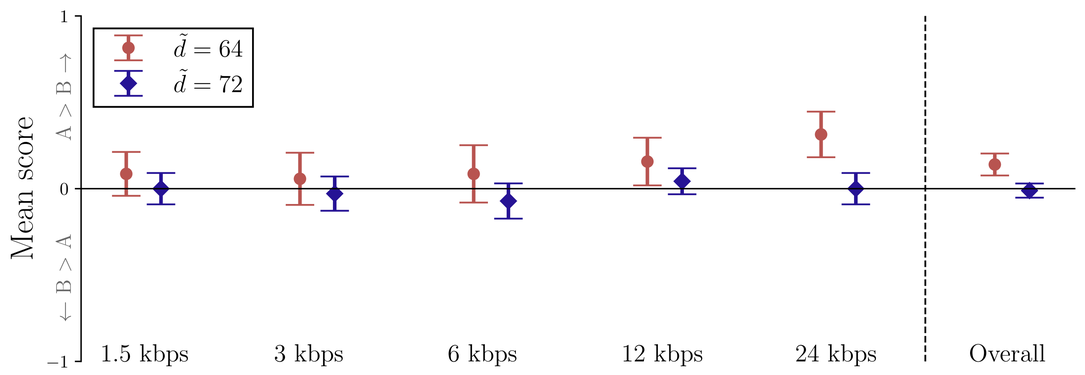


Practical implementation: For EnCodec, the covariance matrix $\mathbf{R}$ and mean $\boldsymbol{\mu}$ are computed from the first $N_{cov} = 2$ learned codebooks. The truncation is performed by keeping only the first $\tilde{d} = 72$ dimensions out of 128.

3. Experimental Results for EnCodec

Truncation of the transformed latent space: The truncation is performed by keeping only $\tilde{d} = 72$ dimensions out of 128. The truncation to 72 dimensions instead of 128 brings about 43.4% gain in codebook storage and computation complexity.

Subjective testing: Tests were conducted to verify that the proposed method does not degrade audio quality. AB tests were performed at each bitrate (1.5, 3, 6, 12 and 24 kbps) and for two truncations ($\tilde{d} = 64$ and 72). Results with 7 expert listeners show that while truncating at 64 dimensions can be perceived, keeping the first 72 dimensions is sufficient and not noticeable in the decoded audio.



4. Application to Other Codecs

- Latent space analysis and decorrelation using eigenvalues decomposition can be generalized. The proposed method to estimate the covariance matrix and mean vector of the discretized latent space based on pre-trained codebooks was applied to other neural audio codecs, e.g., Lyra V2, AudioDec, Descript Audio Codec.
- The proposed method requires only slight modifications to RVQ. The gain in storage and computation complexity is codec dependent.