SNAC: Multi-Scale Neural Audio Codec

Hubert Siuzdak
Papla Media
hubert@papla.media

Florian Grötschla
ETH Zurich
fgroetschla@ethz.ch

Luca A. Lanzendörfer
ETH Zurich
lanzendoerfer@ethz.ch

Abstract

Neural audio codecs have recently gained popularity because they can represent audio signals with high fidelity at very low bitrates, making it feasible to use language modeling approaches for audio generation and understanding. Residual Vector Quantization (RVQ) has become the standard technique for neural audio compression using a cascade of VQ codebooks. This paper proposes the Multi-Scale Neural Audio Codec, a simple extension of RVQ where the quantizers can operate at different temporal resolutions. By applying a hierarchy of quantizers at variable frame rates, the codec adapts to the audio structure across multiple timescales. This leads to more efficient compression, as demonstrated by extensive objective and subjective evaluations. The code and model weights are open-sourced at `https://github.com/hubertsiuzdak/snac`.

1 Introduction

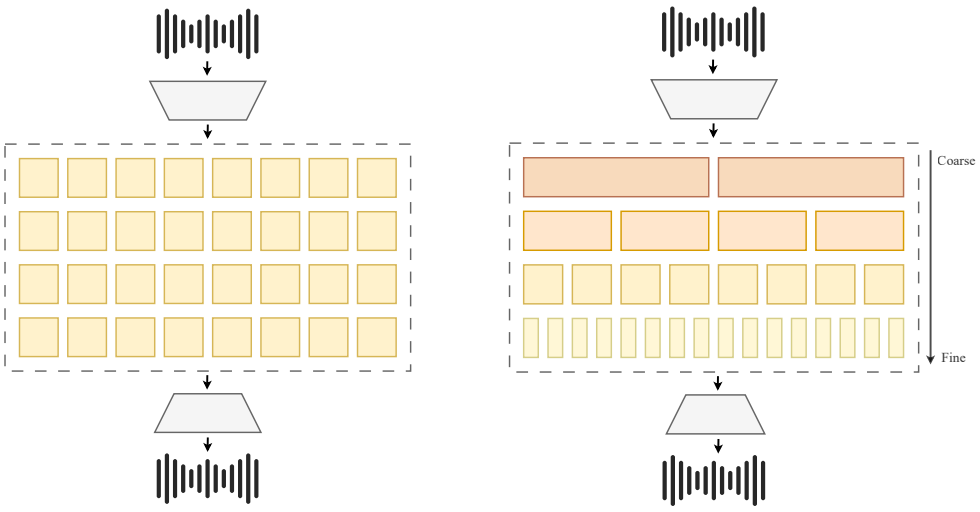
Neural audio compression methods have recently achieved lower bitrates than traditional codecs while maintaining competitive quality [1, 2]. By representing audio using discrete latent variables, these methods have proven useful beyond compression, particularly in generative models. This approach has led to successful generation of natural and coherent continuations of speech and music [3].

However, existing audio tokenizers, while capable of high reconstruction quality, suffer from high token granularity that limits their ability to capture long-term structures. Specifically, in transformer architectures [4], processing extended attention context windows becomes impractical. Ideally, discrete audio tokens should also represent higher-level aspects of sounds [5]. *Semantic* tokens are commonly used in audio generation systems due to their higher information density, which enables autoregressive models to recognize patterns more reliably, much like in natural language processing. In contrast, high-frequency acoustic tokens often contain noise and less patterned data.

Audio signals inherently involve multiple levels of abstraction. Speech can be analyzed at local acoustic or phonetic levels, as well as in terms of prosody or semantics. Music also exhibits hierarchical structures over longer time scales. Research on the human auditory cortex also indicates that acoustic signals are processed hierarchically [6, 7].

In this paper, we introduce **SNAC** (Multi-**S**cale **N**eural **A**udio **C**odec), a simple extension to the current audio residual quantization approach by introducing quantization at different temporal resolutions to form a multi-scale discrete representation of audio. Our experiments – including both objective metrics and subjective evaluations – demonstrate that the proposed method achieves more efficient compression. Additionally, we introduce minor enhancements to the RVQGAN framework by incorporating noise blocks, depthwise convolutions, and local windowed attention.

Neural Information Processing Systems (NeurIPS) 2024 Workshop on AI-Driven Speech, Music, and Sound Generation.



(a) Residual Vector Quantization

(b) Multi-scale Residual Vector Quantization

Figure 1: Comparison between traditional Residual Vector Quantization (RVQ) and our proposed Multi-Scale Residual Vector Quantization. The figures depict the discrete tokens produced by both methods. In the traditional RVQ approach, tokens are generated at a fixed temporal resolution, whereas SNAC utilizes a hierarchy of quantizers operating at multiple temporal resolutions, allowing the codec to capture coarse and fine details more efficiently.

2 Related Work

Neural Audio Codecs Recent advancements in audio codecs have leveraged deep learning to learn efficient audio representations, moving beyond traditional signal processing methods [8]. Vector quantization (VQ) [9], which maps high-dimensional data onto a discrete set of code vectors, has been a foundational tool in compression. It was later applied in deep learning, leading to powerful discrete latent representations in VQ-VAE model [10]. Successful applications of VQ-VAE in audio coding have been demonstrated [11]. However, VQ becomes inefficient at higher bitrates due to the rapid growth of the codebook size. This issue was addressed early on by structuring the quantizer into multiple stages [12]. This approach has now been effectively applied in modern end-to-end models, introduced as Residual Vector Quantization (RVQ) [1], enabling high-quality audio compression with scalable bitrates. Further advancements in RVQ have improved both the efficiency and quality of neural audio compression [13, 2].

Multi-Scale Models Capturing long-term structure in generative models has proven challenging, and various approaches have been proposed to address this issue. A hierarchical decoder was introduced to manage the diffuculties of modeling long-term musical structures [14]. Subsequently, multi-level models for discrete sequences were explored [15]. The multi-scale hierarchical organization of vector quantized codes was proposed to model large images effectively [16]. Additionally, separate VQ-VAE models with different temporal resolutions were used to represent longer dependencies in musical compositions [17].

3 Method

Our model builds upon RVQGAN [2], an encoder-decoder network with Residual Vector Quantization (RVQ) in the bottleneck. It uses a cascade of N_q vector quantization layers, where each layer maps the residual $\mathbf{x} \in \mathbb{R}^{T \times C}$ to a sequence of one-hot vectors of shape $T \times D$, where T denotes the number of frames, C is the encoder dimension, and D is the codeword dimension.

Multi-Scale Residual Vector Quantization Our work extends RVQGAN by introducing Multi-scale Residual Vector Quantization (illustrated in Figure 1). At each iteration i, we downsample the

residuals by a factor of W_i, perform codebook lookup, and then upsample by the same factor W_i to match the original temporal resolution T of $\mathbf{x}$. In practice, we use average pooling for downsampling and nearest-neighbor interpolation for upsampling.

Noise Block To introduce stochasticity and enhance the decoder's expressiveness, we add a *Noise Block* after each upsampling layer. This block adds noise to the activations by updating the input: $\mathbf{x} \leftarrow \mathbf{x} + \text{Linear}(\mathbf{x}) \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ is Gaussian noise, and $\odot$ denotes element-wise multiplication. This mechanism allows the model to inject input-dependent noise. We find that the Noise Block improves reconstruction quality and leads to better utilization of the codebook.

Depthwise Convolution Depthwise separable convolutions were introduced to create lighter models for vision applications [18]. By applying a single filter to each input channel, this method significantly reduces computation and model size. We propose using depthwise convolutions in the generator to not only decrease the number of parameters but also to stabilize training. GAN-based vocoders are notoriously unstable, often experiencing divergent gradients early in training which can lead to instabilities and even model collapse [19].

Local Windowed Attention In our model, we include a single layer of local windowed attention [20] at the lowest temporal resolution in both the encoder and decoder. This is motivated by the attention mechanism's ability to adaptively focus on relevant features across different inputs. Also it can complement the subsequent average pooling by helping to capture contextual representations. Similarly, [13] integrates LSTM layers to model temporal dependencies more effectively.

4 Experiments

We carried out two studies: one where we trained the SNAC codec for general audio – covering music, sound effects, and environmental sounds – and another focused on speech. The speech-specific study aimed to maintain quality with lower bitrates and reduced compute. This approach was inspired by traditional audio codecs that offer specialized algorithms for speech coding [21].

4.1 Model architecture

General Audio Our model follows the encoder-decoder architecture described in [2], with additional noise blocks after each transposed convolution in the decoder. Both the encoder and decoder incorporate local windowed attention layers at the lowest temporal resolution. We replace most convolutions with depthwise convolutions, except in the embedding, output projection, and upsample layers. The encoder uses a cascade of downsampling layers with rates of [2, 3, 8, 8], with corresponding upsampling layers in the decoder at rates [8, 8, 3, 2]. In the RVQ, we use the downsample factors (strides) of [8, 4, 2, 1], effectively compressing a 44.1 kHz input signal into four sequences of tokens at rates of 14, 29, 57, and 115 Hz. Each codebook holds 4096 entries (12-bit), leading to a total bitrate of 2.6 kbps. The model consists of 16 M parameters in the encoder and 38.3 M in the decoder, totaling 54.5 M parameters. We apply the same architecture to train on 32 kHz audio, resulting in token rates of 10, 21, 42, and 83 Hz, with a total bitrate of 1.9 kbps.

Speech For the speech codec, we modify the architecture by adjusting the downsampling factors in the encoder (and correspondingly, the decoder) to [2, 4, 8, 8]. In the residual vector quantization, we use strides of [4, 2, 1]. This model is trained on 24 kHz audio, resulting in token rates of 12, 23, and 47 Hz, with an effective bitrate of 984 bits per second. Additionally, we reduce the number of convolution channels, resulting in 6.7 million parameters in the encoder and 13.0 million in the decoder, for a total of 19.8 million parameters. We omit local windowed attention layers in the speech codec, making the architecture purely convolutional.

4.2 Training

Our training largely follows the GAN setup described in [2] with a multi-period discriminator [22] and a complex multi-scale STFT discriminator [23]. Notably, thanks to depthwise convolutions, we are able to use a higher initial learning rate for AdamW at 6e-4, decaying at a rate of $\lambda = 0.999994$ per iteration. We do not use gradient clipping, as the training remains stable.

Table 1: Ablation study results comparing various model variants. *Single-scale RVQ* uses a single temporal resolution for all VQ codebooks. *w/o Noise Block* is the model without the noise blocks in the decoder. *w/o LWA* indicates the model without local attention layers. *w/o DW Conv.* refers to the model without depthwise convolutions, which failed to produce stable results (metrics not available). The results highlight the effectiveness of our multi-scale quantization approach and the contributions of each component to overall performance.

Model Variant	**ViSQOL (↑)**	**SI-SDR (↑)**	**Mel Distance (↓)**	**STFT Distance (↓)**
SNAC (baseline)	**4.00**	**3.95**	**1.38**	**1.62**
Single-scale RVQ	3.89	3.73	1.44	1.66
w/o Noise Block	3.94	3.32	1.43	1.65
w/o LWA	3.99	3.90	1.39	1.63
w/o DW Conv.	N/A	N/A	N/A	N/A

The models are trained for 800k iterations, with each batch consisting of 16 audio clips of approximately 0.8 seconds each. This results in a total of 2730 hours of unique audio samples, sampled without replacement from our internal dataset. We train the SNAC 32 kHz and 44 kHz versions on a general audio dataset with sampling weights of 80% music, 10% sound effects, and 10% speech. The SNAC 24 kHz version is trained exclusively on a speech dataset. However, it is important to note that this dataset is diverse and may still expose the model to non-speech sounds, such as podcast intro music, jingles, or various background noises and effects.

4.3 Ablation Study

To evaluate the impact of the proposed components, we conducted an ablation study on a 44 kHz version of our model, trained on a general audio dataset. We used a smaller batch size of 8 samples and trained each variant for 250k iterations on the same subset of audio files. The results, summarized in Table 1, demonstrate the effectiveness of each component.

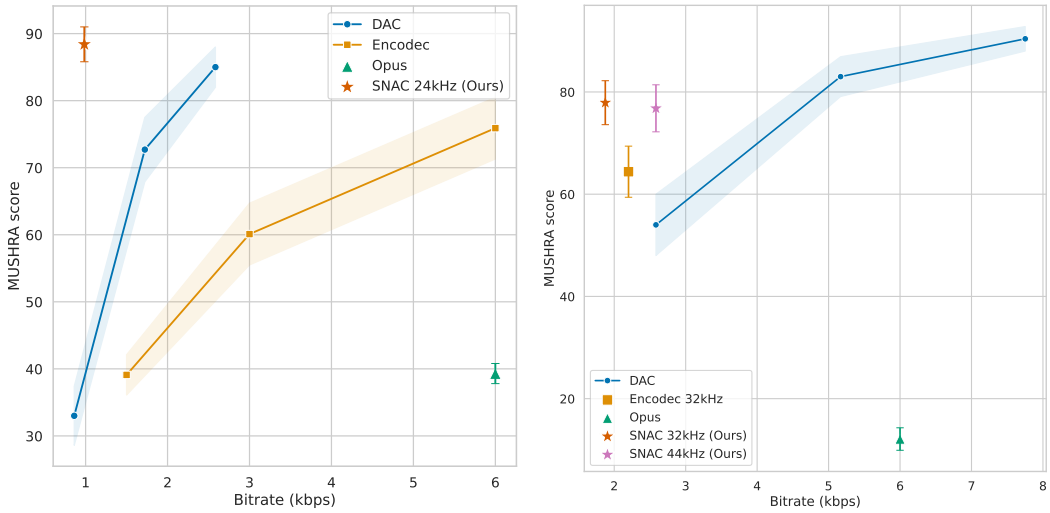
The *Single-scale* variant represents a typical RVQ setup, where a single temporal resolution is used for all VQ codebooks. All other modules, including noise blocks, local attention, depthwise convolutions, and training hyperparameters, remain consistent with the baseline. In this setup, the encoder downsample rates are set to [2, 4, 8, 8], and the model uses 3 levels of RVQ, each with 4096 codebook entries, resulting in a total bitrate of 3.1 kbps. Despite the larger bitrate compared to the SNAC baseline (2.6 kbps), this variant performs worse across all objective metrics, highlighting the importance of our multi-scale quantization approach for audio.

The Noise Block leads to a notable improvement in reconstruction quality, especially in terms of the SI-SDR score. In contrast, the local attention layers provide only marginal gains. We hypothesize that deeper attention networks with larger context windows might yield more meaningful contextual representations. However, these also increase computational costs and latency. For our lightweight speech codec, we chose to omit the local attention layers. Additionally, when training with standard (non-depthwise) convolutions using the same hyperparameters as the baseline, we observed unstable training and eventual model collapse.

4.4 Evaluation

We evaluate SNAC using a combination of objective metrics and a MUSHRA-like (Multiple Stimuli with Hidden Reference and Anchor) [24] listening study to assess the perceived quality of speech and music samples generated by different codecs. For the speech evaluation, we used the DAPS dataset [25], selecting 10 samples of varying lengths (5–9 seconds), with five female voices and five male voices. For the music evaluation, we used 10 samples trimmed to 5 seconds from the MUSDB18-HQ dataset [26]. In addition to the listening study, we evaluated the codecs on commonly used objective metrics to quantitatively assess signal quality and speech intelligibility.

Participants in the MUSHRA test were presented with a hidden reference (the original, unprocessed audio), an anchor (compressed using the Opus codec at 6 kbps, with the corresponding audio and speech preset), and samples that had been encoded and decoded by various neural audio codecs. All samples were loudness normalized to -12dB LUFS. Ratings were collected using the webMUSHRA



(a) Speech evaluation

(b) Music evaluation

Figure 2: Results of the MUSHRA listening study with 95% confidence intervals. We visualize the performance of SNAC compared to previous state-of-the-art approaches. We find that SNAC outperforms existing speech codecs while using a significantly lower bitrate and performs comparably to DAC in music reconstruction quality at a considerably lower bitrate.

framework [27] from a group of audio experts. Results of the human evaluation are presented in Figure 2, and the full results, including objective metrics, are presented in Table 2 and Table 3.

Music We compare the two SNAC variants for general audio introduced in Section 4.1 with the 32 kHz checkpoint of Encodec [13] from MusicGen [28] and the official DAC [2] checkpoint using 3, 6, or 9 codebooks. We observe that SNAC significantly outperforms other codecs, such as Encodec (32 kHz) and DAC (with 3 codebooks), which operate at comparable bitrates. Notably, SNAC even competes with codecs operating at more than twice its bitrate. Furthermore, the difference in perceived audio quality between the SNAC models at 32 kHz and 44 kHz is marginal, suggesting that the 32 kHz model is adequate for most tasks, offering the added benefit of a lower bitrate.

Speech For speech, we compare the SNAC speech model with EnCodec (24 kHz checkpoint) and DAC using varying numbers of codebooks. In our evaluation, SNAC consistently outperforms all other codecs. Notably, even at bitrates below 1 kbit/s, SNAC maintains audio quality that closely approaches the reference signal. This efficiency makes it particularly advantageous for bandwidth-constrained applications, where preserving intelligibility and clarity of voices is crucial.

5 Conclusion

We introduced the Multi-Scale Neural Audio Codec (SNAC), an extension of Residual Vector Quantization that uses quantizers operating at multiple temporal resolutions. This multi-scale approach adapts to the inherent structure of audio signals, leading to more efficient compression. Ablation studies confirmed the significance of our design choices. SNAC outperformed existing state-of-the-art codecs in both music and speech domains, delivering higher audio quality at lower bitrates, as demonstrated by extensive objective and subjective evaluation. By open-sourcing our code and models, we aim to contribute to the advancement of neural audio compression research.

References

[1] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[2] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Emili Balaguer-Ballester, Nicholas R Clark, Martin Coath, Katrin Krumbholz, and Susan L Denham. Understanding pitch perception as a hierarchical process with top-down modulation. *PLoS computational biology*, 5(3):e1000301, 2009.

[7] Jonathan E Peelle, Ingrid Johnsrude, and Matthew H Davis. Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in human neuroscience*, 4:1735, 2010.

[8] Srihari Kankanahalli. End-to-end optimized speech coding with deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2521–2525. IEEE, 2018.

[9] Robert Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.

[10] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[11] Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 735–739. IEEE, 2019.

[12] Biing-Hwang Juang and A Gray. Multiple stage vector quantization for speech coding. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 597–600. IEEE, 1982.

[13] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[14] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, pages 4364–4373. PMLR, 2018.

[15] Sander Dieleman, Aaron Van Den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. *Advances in neural information processing systems*, 31, 2018.

[16] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[17] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[18] Andrew Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[19] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023.

[20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[21] Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry. Definition of the Opus Audio Codec. RFC 6716, 2012.

[22] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[23] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *Interspeech 2021*, pages 2207–2211, 2021.

[24] B Series. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2014.

[25] Gautham J. Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010, 2015.

[26] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18-hq - an uncompressed version of musdb18, 2019.

[27] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stoter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jurgen Herre. webmushra-a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(7), 2018.

[28] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

A Full results

Table 2: Comparison of SNAC with previous neural audio codecs on speech samples. The bitrate is measured in kbps, Mel and STFT measure the L1 distance between the reconstruction and the ground-truth. MUSHRA scores are reported with a 95% confidence interval, and the reference signal in MUSHRA was rated at 99.5 ± 0.3.

Model	**Bitrate**	**ViSQOL** (↑)	**SI-SDR** (↑)	**Mel** (↓)	**STFT** (↓)	**MUSHRA** (↑)
Opus	6	3.84	2.28	4.99	4.01	39.3 ± 1.5
EnCodec	6	4.36	6.83	1.60	1.81	75.9 ± 4.6
	3	4.12	3.72	1.80	1.94	60.1 ± 4.6
	1.5	3.75	0.85	2.05	2.07	39.1 ± 3.0
DAC	2.5	4.28	6.43	1.37	1.65	85.0 ± 3.0
	1.7	4.03	4.22	1.56	1.74	72.7 ± 4.8
	0.8	3.49	-0.16	2.03	1.96	33.0 ± 4.4
SNAC (ours)	0.98	4.14	0.82	1.50	1.78	88.4 ± 2.6

Table 3: Comparison of SNAC with previous neural audio codecs on music samples. The bitrate is measured in kbps, Mel and STFT measure the L1 distance between the reconstruction and the ground-truth. MUSHRA scores are reported with a 95% confidence interval, and the reference signal in MUSHRA was rated at 99.8 ± 0.3.

Model	**Bitrate**	**ViSQOL** (↑)	**SI-SDR** (↑)	**Mel** (↓)	**STFT** (↓)	**MUSHRA** (↑)
Opus	6	1.54	2.24	7.33	5.71	12.1 ± 2.2
EnCodec	2.2	3.66	5.41	1.91	1.97	64.4 ± 5.0
DAC	7.74	4.19	10.60	1.15	1.50	90.4 ± 2.4
	5.16	4.09	8.31	1.30	1.56	83.0 ± 3.9
	2.5	3.91	5.04	1.54	1.69	54.0 ± 6.0
SNAC (ours)	1.9	3.79	4.01	1.75	1.89	77.9 ± 4.3
	2.6	4.04	5.17	1.42	1.59	76.8 ± 4.6