Neural Audio Compression without Residual Vector Quantization

Anonymous submission

Abstract

In this work, we study high-fidelity stereo audio compression without using residual vector quantization (RVQ). We present preliminary findings of our neural audio codec approach, capable of compressing general (speech, music, and environment) stereo audio at 44.1 kHz to 13 kbps with minimal loss in audio fidelity. We achieve this compression by using scalar quantization (SQ) in combination with an autoregressive latent model (ARM), enabling efficient entropy modeling. This approach circumvents the pitfalls of widely-used RVQ approaches, and to the best of our knowledge is the first application of SQ with ARM to the general audio compression domain.

1 Introduction

Audio compression has a long line of research, enabling ever greater storage and transmission efficiency. Traditionally, codecs have relied on handcrafted transforms and quantization rules that are tuned to the human hearing system and psychoacoustics. Recent advances in general audio codecs have shown that learned compression can replace these manually designed pipelines with neural networks trained end-to-end, achieving significantly higher reconstruction quality at considerably lower bitrates. Instead of optimizing each component separately, these approaches train autoencoders to reconstruct the signal while quantizing the latent representation. The reconstruction quality is measured by various losses, based either on signal processing and psychoacoustics or through discriminators inspired by generative adversarial networks (GANs). The overall objective of these networks is to compress audio into the fewest possible bits while reconstructing the audio as faithfully as possible. This objective can be expressed as a rate-distortion (RD) optimization problem:

$$L = R + \lambda D, \tag{1}$$

where λ is a Lagrange multiplier that controls the trade-off between bitrate (R) and reconstruction distortion (D). The rate is measured by the negative log-likelihood of the quantized latent codes, while distortion is often evaluated using perceptually motivated metrics or GAN-losses. This framework has enabled learned neural codecs to surpass conventional codecs in rate-distortion performance in many domains (e.g., image, video, and audio compression).

While entropy models can compress the latent space efficiently, they require quantized latents. Common quantization methods are vector quantization (VQ) and scalar quantization (SQ). VQ maps encoder outputs to discrete codewords from a learned codebook, capturing correlations between latent dimensions. SQ rounds each latent element to the nearest integer independently. Quantization introduces non-differentiability, making end-to-end rate-distortion optimization challenging.

Recent neural audio codecs (Zeghidour et al. 2022; Défossez et al. 2022; Kumar et al. 2023; Siuzdak et al. 2024) rely on residual vector quantization (RVQ), a variant of VQ. While VQ is appealing for generative audio modeling as it allows downstream tasks to use tokenized representations of audio (Agostinelli et al. 2023; Copet et al. 2023; Streich et al. 2025), it is difficult to train and scale. Large codebooks can lead to codebook collapse, while small codebooks limit expressiveness. Moreover, the computational cost of encoding grows exponentially with the dimensionality of the latent space, making VQ infeasible for high-dimensional representations. To mitigate these issues, RVQ is used along with various strategies to avoid codebook collapse. While effective, these approaches add complexity, instability, and latency to the training and inference pipelines.

Scalar quantization offers several practical advantages over VQ in learned compression. While VQ is theoretically optimal for minimizing rate-distortion in finite-dimensional spaces, its computational and memory requirements grow exponentially with latent dimensionality. RVQ mitigates many of these issues but requires multiple codebooks per timestep, increasing complexity. In contrast, SQ treats each latent dimension independently, avoiding the combinatorial explosion of codebook search. Because each scalar latent is quantized independently, its marginal distribution can be modeled using flexible yet tractable probability models, such as factorized or autoregressive priors, without estimating the joint distribution of high-dimensional codewords. Combined with a learned nonlinear transform, uniform scalar quantization is expressive enough to emulate complex, non-uniform quantization behavior while remaining efficient to optimize and deploy (Ballé et al. 2020).

Additionally, the non-differentiability of SQ is easier to overcome. During training, hard quantization can be replaced with additive uniform noise, allowing the quantized latent

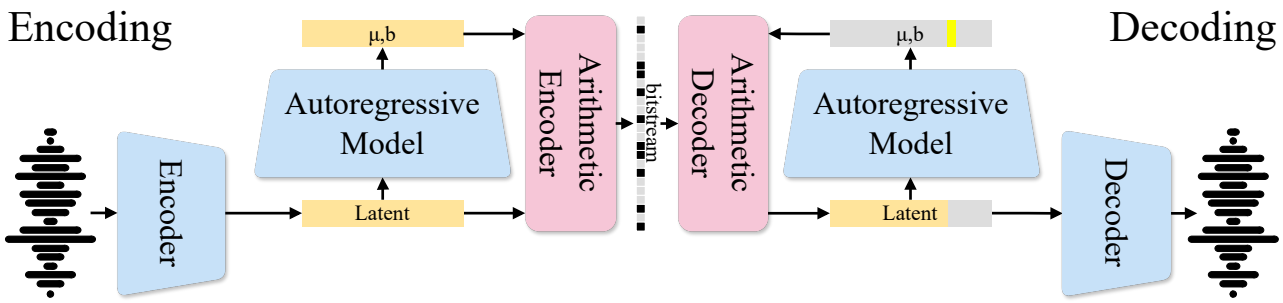

Figure 1: Overview of the proposed autoregressive audio compression model. The encoder transforms raw audio into a latent representation. Latents are quantized using scalar quantization and processed by an autoregressive model that predicts per-latent distribution parameters (mean μ and scale b) conditioned on previously decoded elements. These parameters guide arithmetic encoding to produce a compact bitstream. During decoding, latents are reconstructed sequentially: for each latent, the autoregressive model predicts its distribution parameters, which are then used by the arithmetic decoder to recover the latent. The reconstructed latents are finally passed through the decoder to synthesize the waveform.

distributions to remain differentiable. This stochastic relaxation enables direct optimization of the rate-distortion trade-off with continuous entropy models, resulting in stable and efficient training. In contrast, VQ relies on discrete codebook lookup that break differentiability, requiring additional heuristics such as straight-through estimators or EMA updates, which can lead to instability or codebook collapse. Because of the simplicity and robustness of SQ, in combination with findings that it obtains near-optimal rate-distortion performance (Ballé et al. 2020), SQ has become the standard approach in many learned image compression models (Ballé, Laparra, and Simoncelli 2016; Ballé et al. 2018; Minnen, Ballé, and Toderici 2018; He et al. 2022), and represents an appealing direction for learned audio codecs.

In image compression, ARMs are typically implemented using causal context or checkerboard dependencies (Minnen, Ballé, and Toderici 2018; Minnen and Singh 2020; He et al. 2021). In audio compression, entropy modeling has seen limited exploration. Some works extend hyperprior-based models from image to audio compression (Ballé et al. 2018; Yao et al. 2023; Byun et al. 2023), where Zhen et al. (2021) use an ARM model on a vector quantized latent space to encode speech. Finite Scalar Quantization (FSQ) has recently been proposed as an alternative to vector quantization (VQ) (Mentzer et al. 2023), in which the latent space is discretized onto a fixed set of scalar levels rather than a learned codebook. Several works have explored FSQ for speech audio compression (Langman et al. 2024; Parker et al. 2024; Ye et al. 2025). However, to the best of our knowledge, we are the first to explore scalar quantization with entropy modeling in the general audio domain at 44.1 kHz sampling rate.

Building on these ideas, we introduce a high-fidelity neural audio codec that integrates scalar quantization with an autoregressive latent model and arithmetic coding. Our model is built on the variational autoencoder component of Stable Audio Open (Evans et al. 2024), which reconstructs stereo audio at high-fidelity 44.1 kHz. Our approach efficiently captures temporal dependencies while maintaining training stability, low bitrates, and high perceptual fidelity.

Our contributions can be summarized as follows:

- We show preliminary research on a novel neural audio codec capable of compressing stereo audio at 44.1 kHz without the widely-used RVQ approach. Our model achieves results comparable to those of a state-of-the-art stereophonic neural audio codec using RVQ without the need to mitigate codebook collapse and using considerably less training steps and data.
- To the best of our knowledge, we are the first to apply scalar quantization and autoregressive latent modeling to general audio compression. This enables efficient entropy estimation and temporal dependency modeling, demonstrating that VQ and RVQ are not required to achieve high-fidelity general audio compression.

2 Method

Figure 1 illustrates the proposed compression pipeline. Audio is first encoded into a compact latent representation via a learned encoder network. Latents are quantized using scalar quantization: each element is rounded to the nearest integer at inference, while additive uniform noise is applied during training to preserve differentiability and enable stable rate-distortion optimization. The quantized latents are modeled with an autoregressive entropy model, which predicts the parameters of a per-element Laplace distribution conditioned on previously decoded latents. These distributions are then leveraged by arithmetic coding to produce a compact bitstream of the latent representation.

During decoding, the bitstream is decoded sequentially. At each step, the autoregressive model outputs distribution parameters conditioned on the previously decoded latents, which the arithmetic decoder uses to recover the current latent. The complete latent sequence is then passed through the decoder to reconstruct the waveform. This design effectively captures temporal dependencies while avoiding the complexity and potential instability of vector or residual vector quantization. It further allows stable end-to-end differentiable training, leading to high compression efficiency.

2.1 Autoencoder

We adopt the variational autoencoder from Stable Audio Open (Evans et al. 2024), which operates on raw waveforms. The encoder downsamples the input using strided convolutional blocks, each preceded by residual dilated 1D convolutions with Snake activations (Ziyin, Hartwig, and Ueda 2020). The decoder mirrors the encoder using upsampling blocks followed by residual dilated convolutions with Snake.

Given a waveform x of length T at sampling rate f_s, the encoder with downsampling ratio M produces a latent sequence $z \in \mathbb{R}^{T' \times D}$, where $T' = T/M$ and D is the latent dimension. We index latents by time $t = 1, \ldots, T'$ and channel $d = 1, \ldots, D$.

2.2 Scalar Quantization

We employ scalar quantization (Sullivan 2002), approximated during training with additive uniform noise:

$$\tilde{z}_{t,d} = z_{t,d} + u_{t,d}, \quad u_{t,d} \sim \mathcal{U}\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right), \tag{2}$$

and at inference:

$$\hat{z}_{t,d} = \Delta \cdot \text{round}\left(\frac{z_{t,d}}{\Delta}\right), \tag{3}$$

with $\Delta = 1$. This approach preserves differentiability and enables direct optimization of the rate-distortion objective.

2.3 Autoregressive Latent Model

Efficient entropy modeling is essential to achieve low bitrates. Simple factorized priors, which treat each latent independently, fail to capture the strong temporal dependencies present in audio signals. Autoregressive latent models (ARMs) address this limitation by modeling the probability of each quantized latent element conditioned on all previously decoded time steps. Formally, the joint probability of a latent sequence $z \in \mathbb{R}^{T \times D}$ can be expressed as

$$P(z) = \prod_{t=1}^{T} \prod_{d=1}^{D} P(z_{t,d} \mid z_{<t}), \tag{4}$$

where $z_{<t}$ denotes all latents at previous time steps. While ARMs are autoregressive over time, the latent dimensions at each time step are typically modeled as conditionally independent for computational efficiency. This conditional formulation enables the model to capture temporal correlations accurately, leading to lower bitrates.

A continuous transformer is used as the autoregressive latent model. A start token $s \in \mathbb{R}^{1 \times D}$ is prepended to the latent sequence. At each time step t, the model outputs per-dimension Laplace parameters $\mu_{t,d}$ and $b_{t,d}$, conditioned on all previously decoded latents:

$$p(z_{t,d} \mid \tilde{z}_{<t}) = \text{Laplace}(z_{t,d}; \mu_{t,d}, b_{t,d}). \tag{5}$$

To estimate the number of bits required, we compute the probability mass of the quantization bin centered at $\tilde{z}_{t,d}$:

$$\begin{aligned} r_{t,d} = -\log_2 \Big(& \text{CDF}_{\text{Lap}}\big(\tilde{z}_{t,d} + \tfrac{\Delta}{2}; \mu_{t,d}, b_{t,d}\big) \\ & - \text{CDF}_{\text{Lap}}\big(\tilde{z}_{t,d} - \tfrac{\Delta}{2}; \mu_{t,d}, b_{t,d}\big) \Big). \end{aligned} \tag{6}$$

Let $S = f_s/M$ denote the latent frame rate. The bitrate in bits per second is then

$$\text{bps} = S \cdot \mathbb{E}_{t,d}[r_{t,d}] = \frac{f_s}{TD} \sum_{t=1}^{T'} \sum_{d=1}^{D} r_{t,d}. \tag{7}$$

The resulting bitrate is dynamic and content-dependent: if a latent element is easy to predict (e.g., during silent or highly regular audio segments), the model assigns high probability to the correct quantized value, resulting in near-zero bits. Conversely, for unpredictable or noisy latents, the model assigns a lower probability, leading to a higher number of bits to encode the information accurately.

2.4 Training Objectives

We adopt the Stable Audio Open (Evans et al. 2024) configuration for spectral and adversarial losses. The spectral loss $\mathcal{L}_{\text{SPEC}}$ consists of multi-resolution STFT losses (Yamamoto, Song, and Kim 2020) applied to stereo audio (sum-difference and left/right channels). The perceptual loss $\mathcal{L}_{\text{GAN}}$ combines an adversarial loss and a feature-matching term. Additionally, we include the ARM loss:

$$\mathcal{L}_{\text{LM}} = \text{mean}_{t,d}[r_{t,d}]. \tag{8}$$

During training, an identity mapping layer is inserted before the ARM to control the gradient flow between ARM and the autoencoder. This layer passes the forward signal unchanged, but scales the backward gradients by a factor of λ. This ensures that the ARM is trained at full strength, with an effective learning rate independent of λ, while the autoencoder receives only a fraction λ of the ARM's gradient.

3 Experiments

3.1 Setup

The training dataset consists of 1k hours, aggregated randomly from various sources used by Kumar et al. (2023). For speech, we use DAPS (Gautham 2014), Clean Speech from DNS Challenge 4 (Dubey et al. 2022), Common Voice (Ardila et al. 2019), and VCTK (Veaux, Yamagishi, and Macdonald 2017). For music, we use MUSDB (Rafii, Liutkus, and Stöter 2017) and MTG-Jamendo (Bogdanov et al. 2019). For environmental sounds, we use training split from Audioset (Daniel et al. 2017). All audio is resampled to 44.1 kHz and randomly cropped into 0.37 s segments. For evaluation, we extract 3000 ten-second segments from the test split of Audioset, held-out speakers from DAPS (F1 and M1), and the test split of MUSDB. We follow the balanced data sampling strategy of Kumar et al. (2023), ensuring each batch contains the same number of samples from each domain and at least one full-band audio segment.

We adopt the same training setup as Evans et al. (2024), but reduce the segment length and incorporate an ARM in the latent space. Both the autoencoder and the ARM are trained with AdamW ($\beta_1 = 0.8$, $\beta_2 = 0.99$, weight decay = 1e-3). The base learning rates are 1.5e-4 for the autoencoder and 4e-6 for the ARM. We apply an EMA of model weights ($\beta = 0.9999$). The ARM loss is weighted by λ, and we train two variants with $\lambda = 0.5$ (which operates at 13 kbps) and $\lambda = 0.3$ (which operates at 20.4 kbps). Our models are trained on 4×H100 GPUs with a batch size of 6 for 500k steps.

Model	Bitrate↓	Mel↓	STFT↓	SI-SDR↑	ViSQOL↑
Opus	12.0	0.95	2.20	7.78	2.89
Encodec	12.0	0.69	0.97	8.59	4.05
Ours	13.0	0.63	1.00	10.81	3.83
Opus	20.0	0.58	0.96	7.79	3.99
Ours	20.4	0.52	0.95	13.84	3.93
Encodec	24.0	0.60	0.95	10.80	4.15
Stable Audio	44.1	0.73	0.79	8.38	4.05

Table 1: Objective reconstruction quality of various codecs and our models. Some codecs are evaluated at multiple bitrates. Our model achieves the best Mel distance and SI-SDR, while matching EnCodec in STFT distance.

3.2 Evaluation

We train two models, one version operates at 13 kbps bitrate and the other at 20.4 kbps bitrate. We compare the reconstruction performance of our models with the stereo version of Encodec (Défossez et al. 2022), a learned neural audio codec based on RVQ, and Opus (Valin, Vos, and Terriberry 2012), a widely-used multi-channel audio codec. For each codec, we use two different bitrate settings. We also evaluate against the original variational autoencoder from Stable Audio Open (Evans et al. 2024), which uses FP32 precision, giving a bitrate of $S \cdot D \cdot b = 21.53 \cdot 64 \cdot 32 = 44.1$ kbps. We run each of these models on the evaluation dataset, and compare their reconstruction result using both objective and subjective metrics.

For objective evaluation, we report the mel distance, Short-Time Fourier transform (STFT) distance (Yamamoto, Song, and Kim 2020), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) (Le Roux et al. 2019), and Virtual Speech Quality Objective Listener (ViSQOL) (Chinen et al. 2020). The Mel and STFT distances measure the reconstruction quality based on perceptually weighted spectral distortion and STFT magnitudes. SI-SDR quantifies signal fidelity in the time domain while being invariant to loudness scaling. ViSQOL estimates audio quality by approximating human perception. Lower Mel and STFT distances and higher SI-SDR and ViSQOL scores indicate better reconstruction quality.

For the subjective evaluation, we conduct a MUSHRA listening test.[1] We use ten unseen 10-second audio segments, covering a diverse range of music genres. A total of 20 participants were recruited to rate the audio reconstruction quality. During the test, participants first completed two practice trials, followed by eight test trials. In each trial, participants rated the reconstruction quality of the tested models along with a hidden reference and two anchors (3.5 kHz and 7 kHz), on a scale from 0 (Bad) to 100 (Excellent). Playback was unrestricted and the participants used headphones.

3.3 Results

The results for the objective evaluation are presented in Table 1. Both of our model variants outperform all other models in terms of Mel distance and SI-SDR. For STFT distance, our model achieves comparable performance to Encodec. While these objective metrics are informative, they do not fully align with perceptual quality. For instance, although the 44.1 kbps Stable Audio Open model underperforms compared to other models on the objective metrics, it achieves the best perceptual quality when rated by humans in the MUSHRA test.

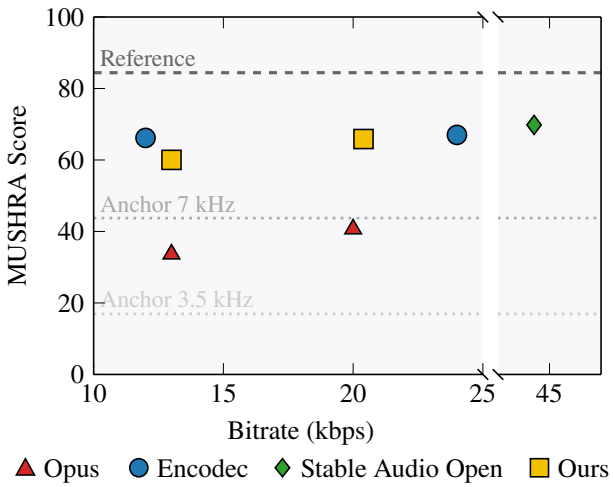

Figure 2: MUSHRA human evaluation results. All learned compression models outperform Opus, which falls below the 7 kHz anchor. Our model achieves performance comparable to EnCodec while using considerably fewer kbps than Stable Audio Open.

The results for MUSHRA listening test are reported in Figure 2. Overall, Stable Audio Open achieves the highest rating. Our 20.4 kbps model is roughly equivalent to the 24 kbps Encodec model, and the 12 kbps Encodec is slightly better compared to our 13 kbps model by roughly 6 points. Our model surpasses Opus across all conditions. Notably, we trained our model on a dataset of 1k hours compared to EnCodec which was trained on 80k hours. The fact that listeners consistently rated our reconstructions as comparable to Encodec suggests that SQ combined with ARM can achieve high perceptual quality at significantly reduced training cost. When compared to the original Stable Audio Open, our model showed a 6% to 14% decrease in MUSHRA scores, but achieved a 55% to 71% reduction in bitrate, demonstrating that our approach is highly promising.

4 Conclusion

We presented preliminary research into compressing stereo high-fidelity audio without leveraging vector quantization. Inspired by approaches used in image compression, we evaluate whether scalar quantization and a latent autoregressive model are feasible approaches for audio compression. Our findings indicate that while our approach cannot yet outperform existing RVQ-based approaches, the rate-distortion trade-off is comparable to the state-of-the-art stereo neural audio codec (Défossez et al. 2022). We believe SQ combined with ARM is an interesting alternative compared to RVQ for learned audio compression, meriting further investigation.

[1] https://www.mabyduck.com/

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common Voice: A massively-multilingual speech corpus. *arXiv [cs.CL]*.

Ballé, J.; Chou, P. A.; Minnen, D.; Singh, S.; Johnston, N.; Agustsson, E.; Hwang, S. J.; and Toderici, G. 2020. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2): 339–353.

Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2016. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*.

Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019. mtgjamendo dataset automatic music tagging. In *Machine Learning Music Discovery Workshop, International Conference Machine Learning (ICML 2019*. Long Beach, CA, United States.

Byun, J.; Shin, S.; Sung, J.; Beack, S.; and Park, Y. 2023. Perceptual improvement of Deep Neural Network (DNN) speech coder using parametric and non-parametric density models. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 859–863.

Chinen, M.; Lim, F. S. C.; Skoglund, J.; Gureev, N.; O'Gorman, F.; and Hines, A. 2020. ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36: 47704–47720.

Daniel, P. W.; Ellis, D.; Freedman, A.; Jansen, W.; Lawrence, C.; Moore, M.; and Plakal, M. 2017. Audio set: ontology human-labeled dataset audio events. In *IEEE international conference acoustics, speech signal processing (ICASSP)*, 776–780. IEEE.

Dubey, H.; Gopal, V.; Cutler, R.; Aazami, A.; Matusevych, S.; Braun, S.; Eskimez, S. E.; Thakker, M.; Yoshioka, T.; Gamper, H.; and Aichner, R. 2022. Icassp 2022 Deep Noise Suppression Challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2022. High Fidelity Neural Audio Compression. *arXiv [eess.AS]*.

Evans, Z.; Parker, J. D.; Carr, C. J.; Zukowski, Z.; Taylor, J.; and Pons, J. 2024. Stable Audio Open. *arXiv [cs.SD]*.

Gautham, J. 2014. Can we automatically transform speech recorded common consumer devices real-world environments professional production quality speech?-a dataset, insights, challenges. *IEEE Signal Processing Letters*, 22(8): 1006–1010.

He, D.; Yang, Z.; Peng, W.; Ma, R.; Qin, H.; and Wang, Y. 2022. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5718–5727.

He, D.; Zheng, Y.; Sun, B.; Wang, Y.; and Qin, H. 2021. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14771–14780.

Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-fidelity audio compression with improved RVQGAN. *arXiv [cs.SD]*.

Langman, R.; Jukić, A.; Dhawan, K.; Koluguri, N. R.; and Ginsburg, B. 2024. Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis. *arXiv [eess.AS]*.

Le Roux, J.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630. IEEE.

Mentzer, F.; Minnen, D.; Agustsson, E.; and Tschannen, M. 2023. Finite Scalar Quantization: VQ-VAE Made Simple. *arXiv [cs.CV]*.

Minnen, D.; Ballé, J.; and Toderici, G. D. 2018. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31.

Minnen, D.; and Singh, S. 2020. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, 3339–3343. IEEE.

Parker, J. D.; Smirnov, A.; Pons, J.; Carr, C. J.; Zukowski, Z.; Evans, Z.; and Liu, X. 2024. Scaling transformers for low-bitrate high-quality speech coding. *arXiv [eess.AS]*.

Rafii, Z.; Liutkus, A.; and Stöter, F.-R. 2017. *Stylianos Ioannis Mimilakis, Rachel Bittner. musdb18 corpus music separation*.

Siuzdak, H.; et al. 2024. Snac: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*.

Streich, G.; et al. 2025. Generating Vocals from Lyrics and Musical Accompaniment. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Sullivan, G. J. 2002. Efficient scalar quantization of exponential and Laplacian random variables. *IEEE Transactions on information theory*, 42(5): 1365–1374.

Valin, J.-M.; Vos, K.; and Terriberry, T. 2012. *Definition opus audio codec*.

Veaux, C.; Yamagishi, J.; and Macdonald, K. 2017. *Cstr vctk corpus: English multi-speaker corpus cstr voice cloning*. University of Edinburgh. The Centre for Speech Technology Research (CSTR).

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram.

In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Yao, S.; Xiao, Z.; Wang, S.; Dai, J.; Niu, K.; and Zhang, P. 2023. Variational speech waveform compression to catalyze semantic communications. In *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–6. IEEE.

Ye, Z.; Zhu, X.; Chan, C.-M.; Wang, X.; Tan, X.; Lei, J.; Peng, Y.; Liu, H.; Jin, Y.; Dai, Z.; et al. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2022. SoundStream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507.

Zhen, K.; Sung, J.; Lee, M. S.; Beack, S.; and Kim, M. 2021. Scalable and efficient neural speech coding: A hybrid design. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 12–25.

Ziyin, L.; Hartwig, T.; and Ueda, M. 2020. Neural networks fail to learn periodic functions and how to fix it. *arXiv [cs.LG]*.