

Neural Audio Codec-Based Multi-Style Singing Voice Generation and Music Reconstruction

Anil Yhopra

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

anil1991@unr.edu

Casper Rao

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.

casperrao00@ku.edu

Abstract

The convergence of neural audio compression and generative sequence modeling has produced a new class of systems capable of reconstructing and synthesizing singing voices across multiple stylistic conditions. This paper examines the system-level foundations of neural audio codec-based multi-style singing voice generation and music reconstruction. It argues that the choice of discrete audio representation, the architecture of the generative backbone, and the design of conditioning mechanisms jointly determine trade-offs in fidelity, style control, data efficiency, and computational cost. The discussion moves beyond model architecture to address the broader socio-technical infrastructure required for responsible deployment. The paper analyzes the structural tensions between high-resolution acoustic reconstruction and controllable stylistic generation, the role of large-scale singing corpora with realistic music scores, and the implications of dataset provenance for voice identity and cultural representation. It further considers deployment challenges including latency, streaming architectures, model compression, and edge inference. Robustness and fairness concerns are examined through the lens of voice cloning, accent and language coverage, and the risk of homogenizing vocal aesthetics. Sustainability and governance are treated as integral system properties rather than external constraints. The paper contributes a cross-domain perspective that links audio codec design, generative music modeling, dataset governance, and policy development. It concludes that future progress depends not only on improved neural codecs or generative architectures but also on institutional mechanisms for evaluation, transparency, and participatory oversight.

Keywords

neural audio codec, singing voice generation, music reconstruction, multi-style synthesis, generative audio, socio-technical infrastructure.

1. Introduction

Neural audio codecs have emerged as a foundational substrate for generative audio systems by converting continuous waveforms into compact discrete representations that can be modeled with sequence-based methods. Early work on discrete representation learning demonstrated that high-dimensional signals such as audio could be quantized into finite codebooks while preserving enough perceptual information for reconstruction [1]. Subsequent codec designs improved the efficiency and perceptual quality of this process, enabling end-to-

end learned compression that operates at low bitrates while retaining musical and vocal detail [2, 3]. These codecs form the backbone of recent singing voice and music generation pipelines because they allow powerful autoregressive and diffusion models that were originally developed for text and symbolic sequences to be applied to audio generation tasks. The use of attention-based sequence modeling further enabled long-range dependencies in music and speech, supporting the emergence of hierarchical generative systems [4, 5].

This paper adopts a system-level perspective on neural audio codec-based singing voice generation and music reconstruction. Rather than focusing solely on quantitative benchmarks or specific architectural innovations, it examines the structural trade-offs that arise when discrete audio representations are coupled with generative models in multi-style singing contexts. Such trade-offs include the tension between acoustic fidelity and style controllability, the balance between model capacity and deployment efficiency, and the relationship between dataset scale and cultural representativeness. The paper also addresses governance and policy dimensions, including dataset documentation, model reporting, fairness, and environmental sustainability. By integrating technical and socio-technical analysis, the discussion aims to provide a publication-ready synthesis suitable for researchers, engineers, and policymakers working on next-generation music and voice systems.

2. Neural Audio Codecs as Substrates for Singing Generation

The transition from raw waveform modeling to neural codec-based generation has reshaped the design space for singing voice synthesis. A neural audio codec typically consists of an encoder that maps audio into a latent sequence, a vector quantizer that converts the latent into discrete indices, and a decoder that reconstructs the waveform. This discrete representation allows generative models to operate over a finite vocabulary of audio tokens rather than continuous samples. The conceptual advantage is that singing voice generation can be reframed as a sequence modeling problem in which the generative network predicts audio tokens conditioned on lyrics, pitch, duration, and style information. The codec thus performs a critical abstraction function: it compresses the acoustic surface of singing into a manipulable symbolic-like space while retaining enough acoustic texture for natural reconstruction [1, 2].

Several system-level implications follow from this design. First, the perceptual quality of generated singing depends heavily on the codec used to create the token space. Codecs that preserve expressive attributes such as vibrato, breathiness, formant movement, and dynamic variation provide a richer generative target. Second, the codec bitrate and codebook size determine the sequence length and vocabulary complexity that downstream generative models must handle. A larger codebook may reduce quantization error but increases the modeling burden, while a smaller codebook may simplify generation at the cost of blurring timbral and stylistic distinctions [3]. Third, codec design affects the feasibility of streaming and low-latency systems. High-fidelity codecs with long temporal receptive fields can introduce algorithmic delay, which matters for live performance or interactive music reconstruction. Therefore, codec selection is not merely an upstream preprocessing choice but a central architectural decision with ramifications for fidelity, efficiency, and control.

3. Architectural Trade-offs in Multi-Style Singing Synthesis

Multi-style singing voice generation poses a distinctive challenge because the system must simultaneously model linguistic content, musical pitch, timing, and stylistic performance attributes such as genre, ornamentation, vocal register, and emotional delivery. Architectures built on attention mechanisms provide a flexible backbone for this task by allowing the model

to attend to different conditioning signals across time [4]. Hierarchical autoregressive approaches represented an early attempt to decompose music generation into multiple temporal scales, with a high-level model producing coarse musical structure and a lower-level model generating detailed audio tokens [5]. This decomposition reduces the effective sequence length for each model but introduces coordination overhead and can produce inconsistencies between levels.

Codec-based language models simplify this hierarchy by operating directly on discrete audio tokens and conditioning on text or musical features. These models inherit the strengths of large language model training, including scalable pretraining and transfer across tasks [8, 11]. However, singing voice generation differs from text-to-speech or instrumental music generation because singing requires precise alignment between lyrics and melody while maintaining natural prosody and expressive timing. Diffusion-based singing synthesis systems have addressed some of these challenges by iteratively refining mel-spectrogram or latent representations, providing more stable pitch control and smoother stylistic transitions [6]. The trade-off is that iterative refinement increases inference cost relative to autoregressive decoding.

A central structural tension is between controllability and acoustic realism. Explicit conditioning on pitch curves, phoneme durations, and style embeddings increases control but may constrain the generative model from discovering natural variations. Conversely, weaker conditioning can produce realistic performances but reduces the ability to specify a desired style. System designers therefore need modular conditioning architectures that allow style attributes to be injected at multiple levels, including global style tokens, phrase-level descriptors, and frame-level acoustic features. The availability of corpora that pair singing audio with realistic music scores and multi-technique annotations has improved the evaluation of such trade-offs, although large-scale standardized benchmarks remain underdeveloped [7].

4. Music Reconstruction and Continuity Across Codec Domains

Music reconstruction in this context refers to the generation or restoration of coherent musical signals from compressed, partial, or symbolic representations. Neural codec-based systems are particularly suited to this task because the discrete token representation can be shared across different generative objectives, including continuation, inpainting, style transfer, and source separation. The reconstruction process involves mapping a musical input into the codec space, manipulating or conditioning the token sequence, and decoding it back to a waveform. The continuity of musical structure depends on how well the generative model captures phrase-level repetition, harmonic progression, and expressive timing. Models trained on large musical corpora have demonstrated the ability to continue musical ideas in a coherent manner by learning these long-range dependencies [8, 9].

A key issue is the alignment between voice and accompaniment when reconstructing full music. Singing voice generation is often evaluated in isolation, but deployed systems must integrate vocal output with instrumental tracks without producing spectral collisions, timing drift, or unnatural balance. Codec-based models that jointly tokenize mixed audio can learn to generate vocal and instrumental components together, but this reduces the ability to control the vocal part independently. Alternatively, separate vocal and accompaniment codecs can be used with a coordination module, though this can introduce synchronization errors. The trade-off between joint modeling and modular reconstruction remains an open systems question with implications for remixing, karaoke generation, and interactive music production [10].

Cross-domain continuity also matters for voice conversion and style transfer. A singer may need to perform the same melody in different languages, genres, or vocal identities while preserving intelligibility and emotional intent. Neural codec language models trained on speech have shown strong zero-shot transfer abilities, suggesting that similar capabilities may be achievable for singing if sufficient paired data and conditioning strategies are available [11, 13]. The challenge is that singing style is not simply a timbral property but includes genre-specific ornamentation, rhythmic phrasing, and dynamic contours that are difficult to capture in a single style embedding. Multi-style systems therefore require representation spaces that disentangle voice identity, linguistic content, pitch, and performance style without sacrificing reconstruction quality.

5. Training Data, Governance, and Dataset Provenance

The performance of neural codec-based singing voice generation is closely tied to the scale, diversity, and annotation quality of training data. Singing datasets must provide aligned lyrics, musical scores, and often style labels to support controllable generation. Early datasets focused primarily on speech or instrumental music, which limited the transfer of techniques to singing tasks. More recent corpora have incorporated global multi-technique singing data with realistic music scores, enabling more systematic training and evaluation across languages, techniques, and expressive styles [7]. Such datasets are valuable because they reduce the reliance on weakly aligned or synthetic labels and support the study of style transfer and reconstruction in more ecologically valid musical settings.

Dataset governance becomes especially important when the audio contains human voices that may be linked to personal identity. Singing voice data can be used to clone or imitate vocalists without clear consent, raising legal and ethical concerns about publicity rights, performer control, and cultural appropriation. Documentation practices such as datasheets for datasets provide a mechanism for clarifying data collection methods, consent procedures, intended uses, and known limitations [22]. However, documentation alone does not resolve the deeper structural problem of who has the authority to decide how vocal data are used. Governance frameworks must therefore address not only transparency but also the distribution of decision-making power among singers, dataset creators, platform operators, and downstream users.

The provenance of musical styles further complicates data governance. Singing traditions are embedded in cultural contexts, and aggregating diverse styles into a single training corpus can flatten regional aesthetics or reproduce dominant commercial styles. A dataset that includes multiple techniques is an improvement over narrowly curated pop-oriented corpora, but representativeness cannot be measured solely by counting style labels. System designers need to work with music communities to understand how stylistic categories are defined, contested, and valued. Governance mechanisms should include participatory review processes, community-based licensing options, and regular audits of how generated outputs reflect or distort underrepresented singing traditions.

6. Infrastructure and Deployment Considerations

Deploying neural codec-based singing voice systems at scale requires careful attention to computational infrastructure. Training large generative audio models demands substantial accelerator resources, particularly when high-resolution codecs produce long token sequences. Preprocessing audio into discrete tokens also requires a trained codec that must be stable across diverse singing inputs. Inference infrastructure must support low-latency streaming for

interactive applications, batch generation for studio workflows, and model serving across different hardware platforms. The choice of codec bitrate and token sequence length directly affects memory usage, decoding speed, and network bandwidth, making it a key deployment parameter [2, 3].

Model compression and efficient decoding are particularly important for real-time voice generation. Quantization, distillation, and pruning can reduce inference cost, but aggressive compression may degrade the fine acoustic details that distinguish expressive singing from synthetic speech. Edge deployment introduces additional constraints on memory and power consumption, which may necessitate specialized architectures rather than simply compressing a large server-side model. Systems that support music reconstruction in live performance settings must also handle variable input conditions, including noisy microphones, missing accompaniment, and user-specific vocal ranges. Robust deployment therefore requires not only model optimization but also adaptive preprocessing and fallback strategies.

The infrastructure for singing voice generation also includes data pipelines, versioning, and evaluation systems. Because generative audio models are frequently updated with new data or architectural changes, consistent evaluation across versions is difficult. The use of realistic music scores and multi-technique annotations can improve benchmarking, but evaluation metrics must capture perceptual quality, stylistic fidelity, and lyrical intelligibility in addition to reconstruction error [7]. A well-designed deployment stack should log generated outputs, track user feedback, and support A/B testing under ethically approved conditions. Without such infrastructure, iterative improvement becomes speculative and risks degrading performance for specific vocal communities.

7. Robustness, Fairness, and Quality Assurance

Robustness in singing voice generation refers to the ability of a system to maintain acceptable quality across diverse voices, languages, genres, and recording conditions. Neural codec-based systems can be sensitive to codec quantization errors, especially for high-pitched or breathy vocalizations that may not be well represented in the training distribution. Singing with heavy ornamentation, non-standard phonation, or rapid pitch changes can expose weaknesses in both the codec and the generative model. Robustness requires stress testing across a broad range of vocal techniques and acoustic environments, including amateur recordings, live performances, and archival material with degraded audio quality.

Fairness concerns arise when systems perform unevenly across demographic and cultural groups. A singing voice generator trained predominantly on commercial pop vocals may fail to reproduce the stylistic norms of classical, folk, or non-Western singing traditions. Accent and language coverage also affect intelligibility and naturalness. These disparities are not merely technical but reflect historical asymmetries in data collection and music industry representation. Addressing them requires more than collecting additional data; it requires metric design that does not encode a single aesthetic standard as universal. Evaluation should incorporate plural notions of quality that are informed by diverse listeners and practitioners.

Quality assurance for generative audio systems must account for both objective and subjective dimensions. Objective measures such as signal-to-noise ratio or spectral distance do not fully capture the perceived naturalness of a singing voice. Perceptual studies are necessary but can be costly and difficult to standardize. Model cards provide a structured way to report intended use, training data, evaluation results, and known limitations, enabling downstream users to make informed decisions [17]. Fairness assessments should be integrated into model cards

and updated as systems evolve. However, quality assurance also requires organizational accountability, including clear procedures for responding to harms such as unauthorized voice cloning or the production of misleading synthetic performances.

8. Sustainability and Policy Implications

The environmental impact of large-scale generative audio systems is a growing concern. Training neural codecs and generative models on extensive singing corpora requires significant computational resources, contributing to carbon emissions and energy consumption. Although audio models are often smaller than large language models, the repeated training cycles required for codec refinement, generative pretraining, and downstream fine-tuning can accumulate substantial costs. Research on energy-aware machine learning has highlighted the importance of reporting efficiency alongside accuracy and perceptual quality [19, 20]. Sustainability considerations should inform model size, token sequence length, and training duration, not as afterthoughts but as design constraints.

Open and efficient model development can mitigate some sustainability burdens by reducing redundant training across institutions. The release of efficient foundation models has demonstrated that competitive performance can be achieved with carefully designed architectures and training procedures [21]. For singing voice generation, open codec models and shared tokenizers could allow research groups to fine-tune smaller models rather than retraining large systems from scratch. However, openness also raises governance questions, particularly when pretrained voice models can be used to clone specific singers without authorization. Policy responses must balance the benefits of open research with the need to prevent misuse of vocal identity.

Policy implications extend beyond energy and privacy to include copyright, performer rights, and cultural preservation. The ability to reconstruct or imitate a singing voice from compressed representations challenges existing legal categories that distinguish between a performance, a recording, and a derivative work. Some jurisdictions are beginning to address digital replicas and voice rights, but international coordination remains limited. A coherent policy framework should recognize the dual-use nature of singing voice technology, support legitimate creative and accessibility applications, and establish enforceable safeguards for non-consensual use. Engaging with singers, music educators, and cultural institutions can help ensure that technical innovation does not outpace the social norms that govern musical expression.

9. Conclusion

Neural audio codec-based multi-style singing voice generation and music reconstruction represents a significant advance in generative audio systems. The discrete token representation provided by neural codecs enables sequence-based generative models to synthesize expressive singing while supporting tasks such as style transfer, inpainting, and reconstruction. However, the system-level trade-offs are substantial. Codec fidelity, generative architecture, conditioning design, and dataset composition jointly influence controllability, acoustic realism, robustness, and deployment efficiency. No single design point is likely to satisfy all use cases, and system developers must therefore make explicit choices about what to prioritize.

This paper has argued that technical progress cannot be separated from governance, fairness, and sustainability. Singing voice systems operate at the intersection of personal identity, cultural tradition, and creative labor. Their development requires not only stronger models

and richer datasets but also improved documentation, participatory governance, and policy mechanisms that protect vocal rights. Future work should develop evaluation protocols that reflect diverse aesthetic standards, build shared infrastructure for reproducible research, and create institutional structures for ongoing oversight. By treating these socio-technical dimensions as core system requirements, the field can advance singing voice generation in ways that are both technically sophisticated and socially accountable.

References

1. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 6306-6315.
2. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., & Tagliasacchi, M. (2021). SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 495-507.
3. Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
5. Dhariwal, P., Jun, H., Payne, C., Kim, J., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *Advances in Neural Information Processing Systems*, 33, 5005-5019.
6. Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2022). DiffSinger: Singing voice synthesis via shallow diffusion mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 11020-11028.
7. Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... & Zhao, Z. (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *Advances in Neural Information Processing Systems*, 37, 1117-1140.
8. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., ... & Zeghidour, N. (2023). AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2523-2533.
9. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzett, M., Caillon, A., ... & Frank, C. (2023). MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.
10. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., ... & Adi, Y. (2023). AudioGen: Textually guided audio generation. *Proceedings of the Eleventh International Conference on Learning Representations*.
11. Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., ... & Wei, F. (2023). Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.
12. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2024). Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*.
13. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction

of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451-3460.

14. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017). Neural audio synthesis of musical notes with WaveNet autoencoders. *Proceedings of the 34th International Conference on Machine Learning*, 1068-1077.
15. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840-6851.
16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
17. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
18. Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68.
19. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645-3650.
20. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
21. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
22. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.