

Personalized Singing Voice Generation via Singing Style Transfer and Voice Cloning

Xiangfan Cao

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
xcao@buffalo.edu

Abstract

Personalized singing voice generation sits at the intersection of speech synthesis, music information retrieval, biometric identity modeling, and generative artificial intelligence. The task requires a system to synthesize novel vocal performances that preserve a target singer's timbral identity while transferring expressive attributes such as vibrato, dynamics, ornamentation, and rhythmic phrasing from a reference performance or score. This paper examines the problem as an infrastructure-level challenge rather than a narrowly defined modeling exercise. It discusses the architectural trade-offs between end-to-end and modular pipelines; the role of self-supervised, discrete, and spectral representations; the design of large-scale singing corpora; training strategies for disentanglement, robustness, and cross-lingual generalization; and the sociotechnical obligations that arise when voice cloning becomes widely deployable. The analysis draws on advances in neural text-to-speech, voice conversion, neural vocoding, music generation, and biometric privacy. It argues that sustainable progress requires coordinated attention to model architecture, data governance, evaluation beyond acoustic fidelity, identity protection, equity across languages and genres, and deployment efficiency. The paper closes with a forward-looking research agenda centered on privacy-preserving personalization, transparent provenance, and robust evaluation. Throughout the discussion, personalized singing voice generation is treated as a coupled sociotechnical system in which technical design and institutional safeguards must co-evolve.

Keywords

personalized singing voice generation; singing style transfer; voice cloning; neural vocoding; self-supervised audio representation; data governance; fairness; deployment.

1. Introduction

Personalized singing voice generation can be understood as the computational construction of a vocal performance that is not explicitly present in any single recorded source. It combines two related but distinct capabilities: voice cloning, which seeks to reproduce the recognizable timbral identity of a target singer, and singing style transfer, which seeks to transplant expressive performance features such as pitch contour, vibrato, breathiness, articulation, ornamentation, and dynamic shaping from a reference performance or symbolic score into a new rendition. Recent advances in neural text-to-speech have shown that highly natural speech can be generated from textual input using attention-based sequence models [1], and raw waveform generative models such as WaveNet [2] established the viability of neural vocoding for high-fidelity audio. Efficient generative adversarial vocoders later reduced the computational cost of waveform synthesis while preserving perceptual quality [3]. These advances were accompanied by large-scale speaker datasets that made speaker embeddings a practical mechanism for controlling timbre [4]. The extension from speech to singing, however, is not a trivial transfer of existing methods. Singing imposes stronger demands on

pitch accuracy, long-range temporal coherence, expressive variability, and the simultaneous control of phonetic content, musical note sequences, and performance style.

The research community has begun to address these demands through zero-shot voice cloning systems, controllable speech synthesis architectures, and music generation models. Transfer learning from speaker verification to multispeaker text-to-speech demonstrated that a fixed speaker encoder can provide stable identity conditioning [5], while zero-shot systems such as YourTTS showed that multi-speaker synthesis and voice conversion can be unified in a single framework [6]. Self-supervised speech representations such as wav2vec 2.0 [7] and HuBERT [8] have provided more robust phonetic and prosodic features than surface acoustic representations. In parallel, large-scale audio generation models such as MusicLM [9] illustrate how hierarchical sequence models can capture long musical structures. Nevertheless, the combination of fine-grained singer identity preservation, singing style transfer, and realistic musical expression remains an open systems problem. The present paper examines that problem from an interdisciplinary perspective, focusing on architectural trade-offs, data infrastructure, training robustness, governance, deployment, and evaluation.

2. Problem Framing: Identity, Style, and the Stability-Authenticity Tension

A core conceptual challenge in personalized singing voice generation is that singer identity and singing style are entangled in natural performances, yet they must often be controlled separately in a generation system. Timbre includes long-term anatomical and habitual spectral characteristics, while style includes local and global expressive decisions such as vibrato rate, vowel coloring, legato, rhythmic timing, pitch bending, and dynamic contrast. In a performance, these dimensions interact: a singer may alter timbre to produce breathiness or belting, and stylistic choices can change the apparent identity of the voice. Therefore, a system that enforces complete separation between identity and style may fail to capture the subtle ways in which a performer's identity is enacted through stylistic behavior. Conversely, a system that fails to separate them may leak style information into the identity embedding, causing unwanted transfer of accent, genre, or emotional manner when only timbre should be preserved.

This tension can be framed as a stability-authenticity trade-off. Systems that prioritize stable identity preservation tend to produce voices that are recognizable but may sound stiff or limited in expressive range. Systems that prioritize expressive authenticity may produce more musical performances but may fail to preserve the target singer over long utterances or across different registers. The design objective is therefore not merely to maximize a single similarity metric but to manage a multidimensional trade-off between timbral fidelity, style controllability, phonetic intelligibility, musical accuracy, and naturalness. This systems-level framing has implications for architecture, data, and evaluation. For example, an identity encoder trained only on spoken data may generalize poorly to sung phonemes and higher fundamental frequencies, while a style encoder trained only on Western popular music may not capture ornamentation in other vocal traditions. The recognition of these coupling effects is essential for building robust personalized singing voice systems.

3. System Architecture and Representation Trade-offs

Personalized singing voice generation systems can be organized into modular pipelines or more integrated end-to-end models. A modular design typically contains a text or score encoder, a content encoder, a speaker identity encoder, a style encoder, a duration or rhythm model, and a neural vocoder. Modularity improves interpretability and allows components to

be trained on different datasets, but it can also introduce cascading errors when a representation mismatch occurs between components. End-to-end systems can reduce such mismatch but require more carefully curated multi-task data and can be more difficult to debug, retrain, or adapt to new languages and musical styles. The choice between modular and end-to-end design is thus a structural trade-off involving development cost, controllability, robustness, and maintainability.

The representation of audio is another central architectural decision. Spectral representations such as mel spectrograms provide a convenient bridge between acoustic modeling and waveform synthesis but may discard phase information and fine-grained high-frequency detail. Neural vocoders have become sufficiently mature to reconstruct high-quality waveforms from compact spectral or learned representations [3], yet they still introduce a dependency on the vocoder's training distribution. Raw waveform models such as WaveNet [2] produce high-fidelity output directly but are computationally expensive. More recently, discrete neural audio codecs have emerged as a promising middle ground, enabling high-fidelity audio to be represented as sequences of discrete tokens that can be modeled with language-like architectures [14]. Self-supervised speech representations offer another path, because they encode phonetic and prosodic content in ways that are robust to recording variation [7,8]. Each representation makes a different trade-off between information density, ease of generative modeling, controllability, and computational cost.

Voice conversion architectures such as AutoVC [10] have demonstrated that zero-shot timbre transfer can be achieved with autoencoder-style bottlenecks. Subsequent work on unsupervised speech decomposition showed that pitch, content, and timbre can be disentangled through carefully designed bottlenecks and reconstruction objectives [11]. The use of discrete self-supervised representations also enables speech resynthesis with reduced dependence on text and explicit phoneme alignment [12]. Fast and controllable text-to-speech architectures have additionally shown that non-autoregressive generation can improve inference speed while retaining high naturalness [13]. For singing voice generation, these components must be coordinated with musical score information, note-level pitch constraints, and rhythmic alignment. The architecture must also handle long-range dependencies, because a sung phrase may contain melismatic passages, sustained notes, and expressive timing that differ substantially from speech. Systems that model singing as a sequence of discrete audio tokens can borrow from text-to-speech and music generation, but they must be carefully conditioned to avoid losing the singer's identity over long sequences.

4. Data Infrastructure and Corpus Governance

Personalized singing voice generation depends heavily on the availability of high-quality singing data with accurate musical annotations. Unlike speech corpora, which can be collected through read prompts, singing data require careful treatment of melody, rhythm, phoneme alignment, and performance style. The construction of large-scale singing corpora is complicated by copyright restrictions, the need for multi-track recordings, and the variability of vocal technique across genres and languages. Public singing datasets have historically been smaller and less diverse than speech datasets. A recently released global multi-technique singing corpus with realistic music scores addresses some of these gaps by providing aligned score information and broad task coverage [15]. Corpora of this kind are valuable because they allow researchers to evaluate singing synthesis, style transfer, and voice cloning under more realistic musical conditions, but they also raise important governance issues around licensing, performer consent, and the reuse of artistic performances.

The data supply chain for personalized singing voice generation must be governed more carefully than generic speech data pipelines. Singing performances often carry strong personality and creative identity, and their collection may involve contracts that restrict subsequent machine learning use. Furthermore, datasets may contain demographic, linguistic, and stylistic imbalances that bias the resulting models. A responsible data infrastructure requires documented provenance, explicit consent for voice cloning, versioning of annotation schemas, and mechanisms for correcting or withdrawing data. The introduction of synthetic data and data augmentation can reduce dependence on scarce recordings, but synthetic data must not inadvertently reproduce the style of an unconsenting reference performer. Corpus governance should also include technical measures such as embedding-level privacy filters, watermarking, and access controls that limit how singer identity representations can be extracted and reused.

5. Training Strategies, Robustness and Generalization

Training a personalized singing voice generation system is not simply a matter of scaling up a speech synthesis model. The model must learn to align linguistic content with musical scores, preserve a target singer's timbre under pitch and register changes, and transfer style without transferring unintended identity attributes. Transfer learning from speech can provide useful initial representations, because speech synthesis models have benefited from large amounts of transcribed audio and speaker-labeled data. However, singing includes sustained vowels, wide pitch ranges, and expressive timing that are poorly represented in read speech. Fine-tuning on singing data is necessary, and zero-shot voice cloning requires that the identity encoder generalizes beyond the speakers seen during training.

Disentanglement is often encouraged through architectural bottlenecks, adversarial training, and auxiliary objectives that encourage a representation to predict one attribute while remaining invariant to another. These objectives are conceptually related to the information bottlenecks used in voice conversion [10,11]. In singing systems, the style encoder may be trained to capture global performance attributes while the identity encoder is trained to remain insensitive to score, language, and emotional expression. The resulting representations are rarely perfectly independent, and the residual entanglement must be monitored through leakage tests. Robustness to noisy reference audio, reverberation, and varying microphone quality is equally important for deployment, especially when users provide short clips recorded on consumer devices. Augmentation with room impulse responses, equalization, compression, and background noise can improve generalizability, but excessive augmentation may degrade the very timbral details that define a singer's identity. A balanced training curriculum and careful validation across diverse recording conditions are therefore essential.

Specialized singing synthesis pipelines have also shown that large-scale singing data can improve note-level control and pronunciation accuracy [16]. Hierarchical generative models for music such as Jukebox [17] demonstrated that raw audio can be generated as discrete tokens while preserving long-range musical structure. AudioLM [18] extended this direction by combining semantic and acoustic token modeling, and controllable music generation frameworks such as MusicGen [19] illustrate how compact conditioning can guide coherent audio without a separate text-to-music alignment stage. These music-oriented systems are not directly designed for voice cloning, but their tokenization strategies, long-sequence modeling techniques, and conditioning mechanisms inform the design of personalized singing voice generation.

Cross-lingual and cross-genre generalization remains challenging. Singing techniques differ substantially across Western classical, pop, opera, folk and various non-Western traditions, and the same acoustic feature may carry different stylistic meaning in different repertoires. A system optimized for Western pop may rely on pitch contours and vibrato patterns that are absent or performed differently in other traditions; similarly, a model trained primarily on studio recordings may fail on field recordings with reverberation. This points to the need for training data that spans not only many speakers but many recording conditions, languages, and musical genres. It also implies that robustness cannot be achieved solely by increasing model capacity. Instead, robustness must be designed through representation choices, augmentation strategies, and evaluation protocols that explicitly test out-of-distribution scenarios.

Moreover, the transfer from speech to singing introduces a set of acoustic phenomena that are less prominent in speech. Sustained notes require stable pitch control over hundreds of milliseconds, and the model must avoid unnatural drift or flutter. The presence of a musical score provides a strong prior on pitch and rhythm, but the score does not determine timbre or expressive microtiming. Therefore, a singing system must balance score adherence with the freedom to produce ornamentation, portamento, and rubato. This balance is especially delicate in zero-shot style transfer, where the reference performance may contain stylistic devices that conflict with the target singer's natural habits. The model must decide whether to preserve the reference's melodic contour, the target's habitual vibrato, or some negotiated intermediate. These decisions are not purely technical; they reflect assumptions about what constitutes identity, style, and artistic intent.

Finally, the availability of richer singing corpora and open research pipelines has made it possible to investigate these trade-offs more systematically. Nevertheless, the field still lacks a unified evaluation framework that captures the multidimensional nature of singing quality. As a result, improvements in one dimension, such as speaker similarity, may be achieved at the expense of another, such as musical expressiveness. Future training strategies should therefore include multi-objective evaluation and human-in-the-loop refinement, rather than optimizing a single acoustic metric.

6. Evaluation, Fairness, and Long-Term Robustness

Evaluation of personalized singing voice generation has often borrowed from text-to-speech and voice conversion. Objective measures include mel cepstral distortion, root mean square error in fundamental frequency, speaker embedding cosine similarity, and automatic speech recognition word error rate. Subjective listening tests typically assess naturalness, similarity to target speaker, intelligibility, and musicality. These measures, however, capture only part of the system's performance. A generated phrase may be highly natural yet fail to preserve the singer's distinctive vowel coloration under a high note; it may be pitch-accurate yet rhythmically rigid; it may score well on speaker similarity yet leak stylistic attributes from the reference performance. The field needs evaluation protocols that explicitly measure style transfer accuracy, identity leakage, score adherence, and cross-lingual intelligibility.

Fairness concerns arise from the distribution of training data and the design of evaluation benchmarks. Singing datasets are often dominated by a small number of languages, predominantly English and other widely resourced languages, and by popular music genres. Singers with voices outside the typical range, including those with lower or higher registers, older voices, or voices shaped by regional accent, may be underrepresented. As a result, such voices may be synthesized with lower fidelity or may be unintentionally normalized toward

the dominant training distribution. This raises equity concerns for users who depend on personalized voice synthesis for assistive communication or for creative self-representation. Fairness in this context is not merely a matter of demographic balance; it also requires attention to linguistic tone, musical scale systems, and culturally specific ornamentation. A system that cannot reproduce non-Western melodic modes without distortion fails to serve global users equitably.

Long-term robustness is another dimension that is often neglected in publication-oriented evaluation. A model may perform well on clean, studio-recorded reference clips but degrade sharply on smartphone recordings with background noise, reverberation, or compression artifacts. Personalized systems intended for consumer deployment must be robust to such variation, because users are unlikely to provide professionally recorded samples. Robustness can be improved through data augmentation, self-supervised pretraining, and noise-invariant representation learning, but these techniques must be validated under realistic evaluation conditions. Furthermore, robustness has a temporal dimension. As more users interact with the system and provide feedback, the system should be able to adapt without catastrophic forgetting of previously learned identities. This calls for continual learning strategies that are compatible with privacy-preserving constraints.

Human evaluation itself is subject to cultural and perceptual biases. Listeners may rate a synthesized voice as more natural if it conforms to familiar styles, and the same voice may be judged differently by listeners from different linguistic backgrounds. Crowdsourced listening panels often lack the diversity needed to detect localized failures. Therefore, evaluation should include expert musicians, native speakers of low-resource languages, and listeners with hearing differences. Community-based evaluation and participatory design can help surface failure modes that standardized benchmarks miss.

7. Deployment, Infrastructure, and Sustainability

Deployment of personalized singing voice generation systems involves significant computational trade-offs. High-fidelity neural vocoding, diffusion-based generative models, and large sequence models with discrete audio tokens can produce impressive output, but they may require substantial memory, specialized accelerators, or cloud-based inference. Real-time applications such as live performance assistance, interactive music production, and assistive voice output demand low latency and reliable scheduling. Edge deployment offers lower latency and stronger privacy by keeping voice data on the user's device, but it may require aggressive model compression, quantization, or distillation that can reduce timbral fidelity. Cloud deployment can support larger models and more frequent updates but introduces data transmission risks and ongoing operational costs. A sustainable deployment architecture often requires a hybrid approach, with a lightweight on-device encoder for identity and style extraction and a more powerful cloud or on-premise generator for high-quality synthesis, depending on the user context and consent preferences.

The environmental sustainability of large generative audio models is a growing concern. Training a state-of-the-art model on large-scale audio datasets consumes considerable energy and produces carbon emissions, especially when multiple ablations and hyperparameter searches are performed. Personalized fine-tuning can multiply this cost if each user requires separate adaptation. Foundation models trained once and adapted efficiently through low-rank methods, adapter modules, or zero-shot conditioning can reduce the marginal energy and data cost per user. However, even inference at scale can become significant for a service with millions of users. System designers should therefore adopt efficiency metrics alongside

quality metrics, report energy consumption, and explore model architectures that provide graceful quality degradation under computational constraints. The goal is not to sacrifice quality but to make the relationship between quality, compute, and environmental cost explicit.

From an infrastructure perspective, personalized singing voice generation requires more than a trained model. It demands identity registration workflows, consent management, secure storage of voice profiles, versioned model artifacts, monitoring for distribution shift, and mechanisms for user correction. The voice profile itself is a sensitive biometric that should be encrypted at rest and in transit, with access limited to authorized subsystems. Logging and audit trails are necessary for accountability, but they must be designed to minimize the retention of raw audio or embedding data. A well-governed system also needs fallback mechanisms when a user withdraws consent or when a voice profile is suspected of being used without authorization. These infrastructure requirements are often underestimated in academic research but are essential for any real deployment.

8. Governance, Policy, and Sociotechnical Implications

Personalized singing voice generation sits at the center of several legal and ethical regimes. A singer's voice can be protected under right of publicity, personality rights, or unfair competition law in some jurisdictions, while biometric privacy regulations may classify voiceprints as sensitive personal data. The collection and use of singing recordings for model training therefore require layered consent that distinguishes between the original performance, the extraction of a voice profile, and subsequent synthetic reuse. The fact that singing voice models can generate new performances in a recognizable style creates tension with copyright, not only in the underlying composition but also in the performer's interpretive choices. Policymakers have begun to respond with proposals for transparency, consent, and labeling of synthetic media, but the global legal landscape remains fragmented. A system that is lawful in one jurisdiction may be unlawful in another, especially when the target singer did not explicitly consent to voice cloning.

Technical safeguards can support governance objectives. Watermarking and provenance tracking can help identify synthetic audio, although current watermarking methods are not fully robust to editing, compression, or adversarial attacks. Speaker verification can be used to ensure that a user enrolling a voice profile has access to a sufficient number of samples from the claimed identity, but it is not foolproof against replay or deepfake-based enrollment. Differential privacy and federated learning can reduce the exposure of raw voice data during training and personalization, but they may also degrade the fine-grained timbral detail that makes a synthesized voice convincing. The design challenge is to integrate these safeguards without making the system unusable for legitimate creative and assistive purposes. A modular architecture can support governance by allowing different components, such as the identity encoder or vocoder, to be audited, replaced, or restricted independently.

Governance must also address power asymmetries. Professional singers and voice actors may have the legal and technical resources to negotiate contracts and detect misuse, but ordinary users may not. A widely available voice cloning tool can be used to create credible impersonations of private individuals without their knowledge, leading to harassment, fraud, or reputational harm. Conversely, restrictive regimes that limit voice cloning to licensed entities may concentrate creative power in large platforms and exclude independent artists or community-based projects. A balanced approach requires transparent terms of service, accessible reporting mechanisms, clear opt-out pathways, and independent oversight. It also

requires that researchers and developers avoid treating voice data as a freely available resource simply because a dataset has been publicly posted. Responsible data use extends beyond legal compliance to include respect for the original performer's artistic and personal identity.

Looking forward, the policy landscape for personalized singing voice generation will likely be shaped by broader developments in synthetic media regulation, biometric privacy, and platform accountability. Technical researchers can contribute by developing standardized provenance formats, by reporting the limitations of their models, and by engaging with legal scholars and civil society. The long-term viability of the field depends not only on producing ever more convincing audio but on building trustworthy systems that users and the public can understand, contest, and govern. This is especially important as personalized singing voice generation moves from research demonstrations to commercial music production, virtual performers, and assistive communication tools.

9. Conclusion

Personalized singing voice generation is best understood as a sociotechnical infrastructure problem rather than a single modeling task. The ability to preserve a target singer's timbre while transferring expressive style from a reference performance requires careful coordination of representation learning, sequence modeling, neural vocoding, data governance, and deployment architecture. The stability-authenticity tension cannot be resolved by a single objective function; it demands continual negotiation between identity preservation and expressive flexibility.

The systems discussed in this paper illustrate that progress in speech synthesis, voice conversion, self-supervised learning, and music generation provides essential building blocks, but singing-specific data and evaluation remain critical bottlenecks. Large-scale singing corpora with realistic scores and broad style coverage are necessary but not sufficient. They must be governed by consent, provenance, and equitable access. At the same time, evaluation must expand beyond acoustic fidelity to include style transfer accuracy, identity leakage, cross-lingual intelligibility, fairness across voices and genres, and robustness to real-world recording conditions.

Future research should pursue privacy-preserving personalization, singing-specific self-supervised representations, sustainable model architectures, transparent provenance mechanisms, and standardized multi-attribute evaluation frameworks. These directions are interdependent. A technically superior model that violates privacy or fails across linguistic communities will not be sustainable. Conversely, strong governance without technical feasibility may stifle creative and assistive applications. The maturation of personalized singing voice generation as a field will therefore depend on deeper collaboration between engineers, music scholars, legal experts, and diverse user communities.

References

1. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.
2. Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

3. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 17022–17033.
4. Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of Interspeech 2017* (pp. 2616–2620).
5. Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., ... & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 31, 4485–4495.
6. Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., Ponti, M. A., ... & Barbosa, L. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2709–2720). PMLR.
7. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.
8. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhota, 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.
9. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzett, M., Caillon, A., ... & Zeghidour, N. (2023). MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.
10. Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019). AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 5210–5219). PMLR.
11. Qian, K., Zhang, Y., Chang, S., Cox, D., & Hasegawa-Johnson, M. (2020). Unsupervised speech decomposition via triple information bottleneck. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 7836–7846). PMLR.
12. Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhota, K., Hsu, W.-N., ... & Dupoux, E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. In *Proceedings of Interspeech 2021* (pp. 3615–3619).
13. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 3171–3180.
14. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., & Tagliasacchi, M. (2021). SoundStream: An end-to-end neural audio codec. *arXiv preprint arXiv:2107.03312*.
15. 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.
16. 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.

17. Dhariwal, P., Jun, H., Payne, C., Kim, J., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341.
18. 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.
19. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., ... & Adi, Y. (2023). Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 47704–47720.