

Knowledge-Enhanced Large Language Models for Joint Lyrics Understanding and Singing Voice Generation

Gylen Bailey

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
dylanbailey@ucf.edu

Fernando Barnett

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
barnettfernando@ku.edu

Abstract

Large language models have established new capabilities in conditional text generation, but singing voice generation requires simultaneous control over semantic content, prosody, pitch, rhythm, voice identity, and emotional expression. This paper examines knowledge-enhanced large language model architectures for joint lyrics understanding and singing voice generation from a systems perspective. The discussion analyzes how structured musical knowledge, multilingual lyric representations, retrieval augmentation, and domain-specific tokenization shape the trade-offs among intelligibility, expressiveness, and computational cost. Architectural alternatives including autoregressive token prediction, latent diffusion over acoustic representations, and hybrid encoder-decoder designs are compared with attention to scalability, latency, and controllability. The paper further examines deployment infrastructure constraints, licensing and provenance of singing corpora, robustness to noisy or adversarial lyrics, fairness across languages and vocal identities, and sustainability implications of large-scale training and inference. By synthesizing insights from language modeling, speech synthesis, music information retrieval, and socio-technical governance, the paper argues that joint lyrics understanding and singing voice generation should be treated as a coupled sociotechnical system rather than a purely acoustic optimization problem. Evaluation paradigms, data documentation, and policy mechanisms must evolve alongside model architectures to support culturally diverse, accountable, and sustainable music generation.

Keywords

large language models; singing voice generation; lyrics understanding; knowledge-enhanced architectures; generative audio; data governance; music information retrieval.

1. Introduction

The emergence of Transformer-based sequence modeling has reshaped language technology by replacing recurrent and convolutional inductive biases with attention mechanisms that scale across long contexts [1]. Pretrained bidirectional encoders demonstrated that self-supervised objectives over large text corpora can produce contextual representations useful for many downstream language understanding tasks [2]. Subsequent autoregressive large language models showed that increases in parameter and data scale yield substantial few-shot capabilities across translation, summarization, question answering, and reasoning benchmarks

[3]. Instruction tuning and reinforcement learning from human feedback then aligned these models more directly with user intent, enabling conversational and task-oriented deployments [4]. Unified text-to-text frameworks further established that many language tasks can be recast as sequence transduction problems, simplifying multi-task training and transfer [5]. These developments have made large language models attractive as central controllers for multimodal generation tasks beyond text, including audio and music.

Singing voice generation is a particularly demanding multimodal problem because it couples linguistic meaning with musical structure. A model must not only understand the semantic, syntactic, and emotional content of lyrics but also produce acoustic sequences that respect melody, rhythm, phonation, voice identity, and cultural style. Unlike text-to-speech, where prosody is relatively constrained by sentence meaning, singing requires deliberate synchronization of syllables with notes, pitch contours, and expressive dynamics. Joint lyrics understanding and singing voice generation therefore sits at the intersection of language modeling, symbolic music processing, and neural audio synthesis. Knowledge-enhanced large language models are promising because external musical and linguistic knowledge can compensate for limited paired singing data and improve systematic generalization. However, integrating such knowledge raises structural trade-offs involving modality alignment, inference latency, memory footprint, data provenance, and evaluation validity. This paper addresses these system-level issues rather than proposing a single model. It analyzes architectural strategies, infrastructure constraints, governance, robustness, fairness, and sustainability with the goal of informing both research and responsible deployment.

2. Background and Related Systems

Recent generative audio and music systems illustrate two broad conditioning strategies. Latent diffusion models compress continuous signals into a lower-dimensional latent space and learn denoising steps conditioned on text or other modalities [6]. These models have been successfully extended to high-fidelity image synthesis and more recently to music and audio generation, where the latent representation reduces the memory and compute burden of operating directly on waveforms. Music generation systems such as MusicLM [7] and MusicGen [8] demonstrate that text descriptions can control high-level musical attributes including genre, instrumentation, mood, and tempo. However, text-music generation often treats lyrics as one conditioning stream among many, which may not guarantee phoneme-level alignment between the sung words and the acoustic realization.

Earlier work on raw audio generation with models such as Jukebox [9] highlighted both the promise and the cost of large-scale generative audio. Jukebox showed that hierarchical VQ-VAE representations and autoregressive transformers can generate long, stylistically diverse musical excerpts with singing, but it also exposed challenges in intelligibility, coherence over minutes-long contexts, and substantial computational requirements. These systems generally separate semantic understanding from acoustic generation: an upstream language model or text encoder processes lyrics, while a downstream decoder produces audio frames or neural codec tokens. This separation creates an interface problem because linguistic features, musical features, and acoustic features evolve at different time scales and have different local stationarity properties. Joint modeling requires architectures and training objectives that can align these heterogeneous representations without sacrificing the scalability benefits of large language models.

3. Knowledge-Enhanced Architectures for Joint Modeling

Modern neural text-to-speech architectures provide important baselines for singing voice generation. FastSpeech 2 decomposes text-to-speech into encoder, variance adaptor, and decoder stages, predicting duration, pitch, and energy as explicit intermediate variables [10]. Glow-TTS uses normalizing flows and monotonic alignment search to learn a probabilistic mapping between text and acoustic representations [11]. These architectures demonstrate that separating linguistic content, prosodic features, and acoustic decoding improves controllability and data efficiency. Singing-specific models such as DiffSinger further adapt shallow diffusion mechanisms to generate mel spectrograms conditioned on music scores and lyrics, showing that iterative refinement can produce expressive singing without overly complicated autoregressive dependencies [12].

A knowledge-enhanced large language model for joint lyrics understanding and singing voice generation can build on these ideas by introducing explicit musical knowledge modules. The language model may consume lyrics, phoneme sequences, note pitch, note duration, key, tempo, and style tags as a unified token sequence. Structured knowledge graphs encoding language-specific phonotactics, rhyme conventions, melodic idioms, and singer characteristics can be retrieved or injected through adapter layers. Music Transformer provides evidence that relative attention and long-sequence modeling can capture periodic musical structure over long horizons, which is useful for preserving global musical form while generating local acoustic details [13]. Data remains a critical bottleneck, and a recent global multi-technique singing corpus with realistic music scores has been developed to cover diverse singing tasks, languages, and recording conditions [14]. Such resources enable researchers to evaluate whether knowledge-enhanced architectures can move beyond single-singer and single-language demonstrations toward robust and generalizable systems.

Architecturally, three configurations deserve attention. The first uses a single autoregressive sequence model that interleaves lyric tokens, phoneme tokens, note tokens, and quantized acoustic tokens. Vector quantized variational autoencoders established that discrete latent codes can compress audio into sequences that language models can predict [15]. AudioLM later showed that hierarchical token prediction can preserve semantic and acoustic consistency over long spans [16]. This unified design supports rich cross-modal attention and straightforward conditional generation, but it requires careful tokenization so that lyric syllables and musical notes align at the correct temporal resolution. Without alignment constraints, a model can generate plausible local phonemes and plausible local pitches while producing mismatched syllable-to-note correspondences. Knowledge-enhanced variants can inject alignment constraints as additional tokens or as attention biases derived from music scores, reducing the search space.

A second configuration uses a pretrained encoder for linguistic and phonetic features and a separate acoustic decoder or vocoder. Self-supervised speech representations such as HuBERT [17] and wav2vec 2.0 [18] capture speaker-independent phonetic content and prosodic cues. The encoder can be frozen or fine-tuned while the decoder learns to synthesize singing-specific features. HiFi-GAN provides an efficient generative adversarial network for converting mel spectrograms into waveforms, achieving both high perceptual quality and low inference latency [19]. In this design, the language model performs lyrics understanding and symbolic musical reasoning, while neural vocoding handles the final audio synthesis. The principal trade-off is that encoder and decoder objectives may drift apart if the singing data distribution differs significantly from the speech data used for pretraining. Domain adaptation on singing corpora becomes essential.

A third configuration adopts latent diffusion or iterative refinement over acoustic latents, where a language model acts as a semantic planner and a diffusion model or flow model refines the acoustic representation. AudioGen illustrates how discrete audio tokens can be generated from textual conditions using an autoregressive model [20]. Singing voice generation can extend this approach by conditioning the diffusion process on score-informed features, key signatures, voice identity embeddings, and emotion labels. This design often provides better diversity and robustness to over-smoothing than deterministic regression. However, iterative refinement increases computational cost during inference, especially for long songs. Hybrid systems may use autoregressive planning for coarse structure and a small number of refinement steps for fine acoustic detail. The choice among these configurations is not purely acoustic; it depends on deployment targets, latency budgets, controllability requirements, and the availability of paired lyrics-score-audio data at scale. A unified token model may be preferable for research on compositional generalization, while an encoder-decoder pipeline may be more practical for studio-assisted workflows requiring high controllability and low latency.

4. Deployment Infrastructure and System-Level Trade-Offs

Deployment of knowledge-enhanced singing voice generation models involves decisions about cloud versus edge inference, memory footprint, quantization, caching, and multi-tenant serving. The computational profile differs from text-only language models because audio token sequences can be orders of magnitude longer than lyric token sequences. A three-minute song sampled at 44.1 kHz might require tens of thousands of acoustic tokens depending on the codec and hop length. Autoregressive decoding across such sequences can incur substantial latency, making on-device real-time generation difficult without aggressive quantization, distillation, or speculative decoding. Knowledge modules add further memory and retrieval costs. A retrieval-augmented architecture must maintain an index of musical knowledge, which may include phonetic dictionaries, score databases, genre-specific patterns, and singer profiles. The index must be updated and versioned alongside the model to ensure reproducibility and consistency. Serving systems must therefore treat the knowledge base as a first-class infrastructure component, not an auxiliary artifact.

Caching and batching introduce additional trade-offs. Text generation serving often benefits from key-value cache accumulation across tokens. Audio generation with long sequences imposes larger cache memory demands. Multi-user systems serving different songs may need to switch voice identities and musical styles frequently, reducing the benefits of cross-request caching. Model parallelism and tensor parallelism can reduce latency but increase inter-device communication. Quantization to 8-bit or 4-bit precision has proven effective for text models, but its impact on acoustic token prediction and vocoder quality requires separate validation. Perceptual audio metrics may degrade nonlinearly under quantization, especially in high-frequency regions that contribute to timbre and breathiness. System designers must balance bit rate, codec quality, model precision, and energy consumption.

The separation between lyrics understanding and acoustic decoding can also facilitate deployment. A lightweight on-device model may perform lyric alignment and symbolic pitch timing while a larger cloud model performs waveform generation. This split protects user privacy because raw lyrics and voice characteristics may remain on-device, while only symbolic features or anonymized embeddings are sent to the cloud. However, split inference introduces network latency and potential security risks if embeddings can be inverted. A

system-level account must therefore include not only model accuracy but also service reliability, user consent, data minimization, and edge-cloud coordination.

5. Robustness, Fairness, and Cultural Representation

Joint lyrics understanding and singing voice generation inherits robustness challenges from both natural language processing and generative audio. Lyrics often contain code-switching, abbreviated spellings, dialectal phonetics, nonstandard orthography, and creative punctuation. A model trained predominantly on standardized languages may produce unintelligible or inappropriate phoneme sequences when encountering these forms. Knowledge-enhanced systems can mitigate this by consulting phonotactic dictionaries and multilingual pronunciation models. Yet the knowledge base itself may encode prescriptive norms that erase dialectal variation. Robustness is therefore not only a technical property of the model but also a question of whose linguistic and musical conventions are represented.

Fairness concerns arise across multiple axes, including language, gender, accent, vocal register, and musical genre. Singing corpora have historically been skewed toward a small number of languages and professional singers in controlled studio conditions. If a generative model is optimized for average intelligibility or similarity to reference vocals, it may systematically underperform on minority languages or non-normative voice types. This can reproduce exclusionary dynamics in music production tools, where users who do not conform to the dominant vocal aesthetic receive lower quality outputs. Knowledge-enhanced architectures can in principle improve fairness by incorporating diverse priors and retrieval examples. However, without explicit fairness criteria in data curation, evaluation, and deployment, additional knowledge may amplify the norms embedded in the knowledge base.

Large language models have been criticized for reproducing stereotypes and for treating linguistic variation as noise [22]. Similar risks apply to singing voice generation. A model may learn to associate certain genres with particular accents, genders, or emotional expressions in ways that reinforce cultural stereotypes. For example, a knowledge module that retrieves style templates based on genre labels might assign a narrow range of vocal timbres to a genre such as hip-hop or country. Datasheets for datasets [23] and foundation model transparency mechanisms [24] can help practitioners document the intended use, composition, and limitations of singing corpora and models. Such documentation is essential for downstream users who wish to audit bias or adapt a model to a new cultural context. However, documentation alone does not resolve representational harms; it must be combined with participatory design, community consultation, and independent evaluation.

The concept of singer identity and voice likeness raises additional ethical and legal concerns. Singing voice generation can be used to clone a specific singer's timbre, raising questions about consent, publicity rights, and artistic attribution. The same capability that enables personalized assistive music creation can also enable unauthorized impersonation. Governance mechanisms must distinguish between voice conversion for parody, accessibility, or historical restoration and voice cloning for deceptive purposes. Provenance tracking, watermarking, and access controls are infrastructural safeguards, but their effectiveness depends on adoption across platforms and legal jurisdictions. A joint lyrics understanding system may also infer emotional states or sensitive attributes from lyrics and singing style, creating privacy risks that differ from text-only models.

6. Evaluation and Governance Implications

Evaluation of joint lyrics understanding and singing voice generation cannot rely solely on text-based metrics such as BLEU or perplexity. Systems require multidimensional assessment that separates lyric intelligibility, pitch accuracy, rhythmic alignment, voice similarity, emotional expressiveness, and listener preference. Subjective listening tests remain important but are costly, difficult to reproduce, and sensitive to listener demographics. Objective metrics such as mel cepstral distortion, F0 correlation, and phoneme error rate provide partial information but do not fully capture perceived naturalness or cultural appropriateness. Knowledge-enhanced models may perform well on objective metrics while failing human judgments because they produce technically aligned but expressively flat singing. Evaluation protocols should therefore include both human listening panels and context-specific automatic metrics, with detailed reporting of rater recruitment and task instructions.

Generalization across languages and singing styles is another evaluation challenge. Cross-lingual transfer may succeed for phoneme prediction but fail for prosody due to language-specific tone, stress, and syllable timing. A model trained on Mandarin and English singing may produce intelligible Mandarin phonemes but incorrect lexical tones if the musical pitch contour conflicts with tonal requirements. Song-level coherence also matters. A model may generate plausible local phrases while losing global key, tempo, or emotional arc over a full song. Knowledge-enhanced architectures that explicitly represent musical form can improve long-range coherence, but this requires benchmark tasks evaluating entire songs rather than short clips.

Sustainability considerations enter both training and inference. Training large generative audio models is computationally expensive and energy-intensive. Prior analyses of natural language processing have shown that training large Transformer models can produce substantial carbon emissions and that these costs are not evenly distributed across institutions [21]. Singing voice generation models often require additional audio preprocessing, vocoder training, and high sampling rates, further increasing energy demands. Deploying such models at scale, especially for consumer-facing real-time applications, can shift costs from one-time training to continuous inference. System designers should report energy consumption, support quantization and distillation, and consider whether smaller task-specific models may be preferable for many use cases. Similarly, dataset storage and knowledge base maintenance require long-term infrastructure investment. Open-access corpora may become inaccessible or unmaintained, leading to reproducibility challenges. Archiving and versioning practices must be part of the research lifecycle.

Governance of generative music systems intersects intellectual property, cultural policy, and platform regulation. Training on copyrighted lyrics or singing recordings may raise legal questions that differ across jurisdictions. Compensation mechanisms for singers, lyricists, and composers are underdeveloped relative to the scale of generative music deployment. Policymakers and platform operators may require model developers to disclose training data, implement opt-out mechanisms, and provide output provenance. At the same time, overly restrictive governance may hinder research and limit access for independent artists. A balanced approach would distinguish between commercial deployment and noncommercial research, mandate transparency without requiring full public release of proprietary data, and support community-based governance for culturally specific musical traditions. International coordination remains difficult but necessary, because singing data and generative outputs flow across borders through cloud infrastructure and streaming platforms.

7. Conclusion

Knowledge-enhanced large language models offer a promising direction for joint lyrics understanding and singing voice generation, but the problem demands more than scaling model parameters. The system-level perspective advanced in this paper shows that architectural choices must be evaluated alongside deployment constraints, data governance, robustness, fairness, cultural representation, and environmental cost. Unified autoregressive token prediction, encoder-decoder separation, and diffusion-based refinement each embody different trade-offs between controllability, latency, and expressiveness. External musical and linguistic knowledge can improve alignment and generalization, but it also introduces new infrastructure dependencies and the risk of encoding prescriptive norms.

Robustness and fairness are not secondary concerns that can be delegated to post-hoc audits. They are constitutive properties of how training data are collected, how knowledge bases are structured, how evaluation metrics are defined, and how models are deployed across cloud and edge environments. The singing voice is deeply tied to identity, culture, and artistic agency. Systems that generate singing must therefore be governed with attention to consent, provenance, and equitable participation. Evaluation protocols must expand beyond acoustic fidelity to include linguistic diversity, long-form coherence, and socio-cultural appropriateness.

Future research should develop open benchmarks that reflect multilingual singing traditions, document training data through datasheets and model cards, and measure energy consumption across the full lifecycle. Researchers should also investigate hybrid architectures that combine the compositional strengths of discrete token modeling with the perceptual quality of neural vocoders while maintaining interpretability and controllability. As generative audio moves from research demonstrations to creative production tools, the boundaries between language models, music information retrieval, and audio synthesis will continue to blur. Interdisciplinary collaboration among engineers, musicians, linguists, legal scholars, and policy researchers will be essential to ensure that knowledge-enhanced singing voice generation serves diverse creative communities responsibly and sustainably.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shinn, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

5. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.
7. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.
8. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*.
9. Dhariwal, P., Jun, H., Payne, C., Kim, J., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.
10. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. *International Conference on Learning Representations*.
11. Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33, 8067–8077.
12. Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2021). DiffSinger: Singing voice synthesis via shallow diffusion mechanism. *arXiv preprint arXiv:2105.02446*.
13. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music Transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*.
14. 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.
15. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 6306–6315.
16. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., & Zeghidour, N. (2023). AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2523–2533.
17. 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.
18. 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.

19. 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.
20. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., & Adi, Y. (2023). AudioGen: Textually guided audio generation. *International Conference on Learning Representations*.
21. 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.
22. 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.
23. 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.
24. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.