

Singing Voice–Melody Co-Generation and Automatic Arrangement for Intelligent Music Composition

Leif C. Russell

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
lrussell@unh.edu

Liangxuan Xia

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
liangxuanwork@colostate.edu

Darren Ramos

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
darrenwork@uab.edu

Abstract

The integration of singing voice synthesis with melodic co-generation and automatic arrangement represents a substantial shift in intelligent music composition. This paper presents a system-level examination of coupled vocal and musical generation, focusing on architectural trade-offs, multimodal representation, training infrastructure, evaluation, fairness, and governance. Rather than treating voice synthesis and melody generation as separate downstream tasks, the paper argues for a coupled modeling paradigm in which symbolic musical structure, linguistic content, and acoustic vocal detail are aligned through hierarchical representations. The study examines how speech synthesis architectures, audio language models, symbolic music generators, and arrangement systems can inform this paradigm. Attention is given to the structural tensions between end-to-end integration and modular engineering, between high-fidelity acoustic modeling and interpretable symbolic control, and between expressive diversity and deterministic consistency. The paper further discusses data requirements for singing-specific corpora, the role of automatic arrangement as constrained generation, and the robustness risks that arise when generative models are deployed in creative workflows. Infrastructural concerns include compute demands, dataset governance, licensing, voice identity protection, and the environmental footprint of large-scale training. Evaluation frameworks are considered in terms of objective metrics and human musical judgment, with particular attention to the inadequacy of speech-oriented or instrument-oriented metrics for sung expression. The discussion extends to policy implications, including consent, provenance, cultural representation, and the need for transparent model documentation. The paper concludes that sustainable progress in singing voice and melody co-generation requires not only advances in generative architectures but also institutional support for open benchmarks, ethical data collection, and interdisciplinary collaboration among engineers, musicians, and legal scholars.

Keywords

intelligent music composition, singing voice synthesis, melody generation, automatic arrangement, generative modeling, music information retrieval, socio-technical systems.

1. Introduction

Intelligent music composition has historically developed along two relatively separate trajectories: symbolic modeling of musical structure and acoustic modeling of performance. Attention-based sequence models [1] and autoregressive waveform synthesis [2] have created common computational foundations for both trajectories. However, singing voice and melody co-generation introduces new integration requirements because the voice is not merely another instrument; it carries linguistic, prosodic, and expressive constraints that must be aligned with harmonic and rhythmic structure. Early neural vocoders and efficient synthesis architectures [3], [4] showed that high-fidelity audio can be produced from compact intermediate representations, but they were developed primarily for speech. Their adaptation to singing requires addressing pitch contours, phoneme timing, vibrato, and style transfer in a unified generation process.

Speech synthesis has offered useful design patterns for this challenge. Fast and controllable text-to-speech systems [5] demonstrated that non-autoregressive architectures can be robust for long sequences. Conditioning waveform models on predicted spectrograms [6] established a modular separation between linguistic-acoustic mapping and waveform reconstruction. End-to-end speech synthesis [7] further simplified system pipelines, but singing voice generation remains more complex because the melody is simultaneously an input, an output, and a structural constraint. In creative composition, a system may need to invent a melody while synthesizing a vocal line, or harmonize a given melody while maintaining natural voice production. These coupled objectives require not only improved acoustic models but also new forms of multimodal control and evaluation.

Parallel work in music generation has produced large generative models that can create audio with broad stylistic range [8], [9], [10]. Symbolic music generation has also advanced through multi-track generative adversarial networks [11] and long-range transformer models [12]. Surveys of deep learning for music generation [13], [14], [15] emphasize that representation choice, temporal coherence, and evaluation remain central problems. Public datasets such as MAESTRO [16] enabled controllable piano music modeling, while evaluation frameworks [17] provide metrics for pitch, rhythm, and harmonic consistency. Yet the singing domain has lacked data and benchmarks that combine realistic music scores with high-quality vocal performances across multiple techniques. The present paper examines how co-generation and automatic arrangement can be organized into a system-level research program that connects these lines of work.

2. Representation and Multimodal Alignment

A critical bottleneck in singing voice melody co-generation is the alignment of symbolic and acoustic representations. Recent work has introduced a large-scale multi-technique singing corpus with realistic music scores, covering multiple languages, techniques, and expressive annotations [18]. This resource changes the data landscape by enabling models to learn relationships among scores, phoneme sequences, pitch contours, and recorded vocals in a more systematic manner. The availability of such aligned data is important because co-generation cannot be reduced to simple concatenation of a melody generator and a vocoder. The system must resolve timing differences, local ornamentation, syllable stress, and expressive deviations that arise when a singer interprets a score.

Audio language models have demonstrated that discrete token representations can support high-level semantic and low-level acoustic generation within a single sequence modeling framework [19]. For singing voice and melody co-generation, a comparable tokenization strategy may be used to represent notes, phonemes, duration, dynamics, and spectral features

in a shared vocabulary. The trade-off is between compression and controllability: coarse tokens are easier to model but may lose the fine detail of vocal timbre, while fine tokens preserve detail but require more compute and larger datasets. Multitrack resources such as MedleyDB [20] illustrate how detailed annotations can support research on source separation and automatic arrangement, although singing-specific corpora provide tighter coupling between score and voice.

Alignment also involves governance of missing or uncertain information. In real-world music creation, a user may provide only lyrics, only a melody, or only a chord progression. The system must infer plausible completions without violating musical constraints or producing phonetically awkward vocal lines. This type of inference requires a compositional representation in which musical events, linguistic units, and acoustic descriptors are aligned at multiple temporal resolutions. Cross-domain comparisons with speech synthesis suggest that hierarchical representations improve controllability, but singing introduces additional degrees of freedom such as pitch drift, register shifts, and breath placement. System-level design therefore favors modular encoders and decoders that can be trained separately on symbolic and acoustic data, then fine-tuned jointly on paired singing corpora.

3. System Architecture and Co-Generation Trade-Offs

Co-generation architectures can be organized along a spectrum from tightly coupled end-to-end models to modular pipelines. End-to-end models promise simpler training and more consistent latent representations, but they often require very large datasets and can produce unpredictable failures when a single component degrades. Modular pipelines permit independent improvement of melody generation, vocal synthesis, and arrangement, but they introduce error propagation and require carefully designed interfaces. Automatic accompaniment research such as MySong [21] demonstrates that even relatively constrained systems can support usable creative interaction by modeling harmonic compatibility rather than attempting complete musical generation. However, singing voice co-generation demands tighter integration because vocal plausibility depends on both melodic contour and phonetic articulation.

Neural audio synthesis of musical notes [22] provides insight into the trade-offs between autoencoder-based compression and generative quality. In a singing co-generation system, a similar autoencoder can compress vocal recordings into interpretable latents that encode pitch, loudness, and timbre, while a separate sequence model generates the symbolic melody and phonetic content. The decoder then reconstructs a singing voice conditioned on both the acoustic latent and the generated musical score. This separation has practical advantages: it allows the system to reuse pretrained vocal decoders, to control voice identity separately from melodic content, and to apply adversarial or diffusion-based training to improve realism. The main structural cost is that the model must learn to resolve mismatches between inferred acoustic latents and newly generated symbolic constraints, which can lead to instability if not regularized.

Hierarchical generative modeling for speech [23] offers a partial solution by separating high-level linguistic content from low-level acoustic detail. Applying this approach to singing co-generation suggests a three-layer architecture in which the highest layer models global structure such as form, key, and emotional contour; the middle layer models phrase-level melody and lyrics; and the lowest layer models fine-grained vocal acoustics. Such a hierarchy reduces the computational burden on any single component and enables conditional interventions at different levels. The trade-off is that errors at the global layer can cascade

through lower layers, and training a hierarchical model with musical and linguistic objectives requires careful balancing. System engineers must decide whether to optimize all layers jointly or to use teacher forcing and scheduled sampling, each of which affects sample diversity and inference stability.

4. Automatic Arrangement as Constrained Generation

Automatic arrangement is often formulated as the generation of additional musical parts that are coherent with a lead melody and harmonic progression. In intelligent composition, arrangement must consider instrument roles, rhythmic density, voice leading, and production constraints. The problem becomes more complex when the lead line is sung because the vocal already contains expressive timing and timbral variation. A system that treats arrangement as constrained generation can condition on vocal features such as pitch contour, syllable boundaries, and phrase structure. This is different from arranging for a monophonic instrument because the vocal line imposes phonological and perceptual constraints that are not represented in a purely symbolic score.

System-level arrangement architectures may be compared to multi-track generative models but with a lead voice as an explicit conditioning stream. The arranger must decide not only what notes to add but also when to leave space for the vocal, how to avoid spectral masking, and how to support the emotional arc of the song. These requirements suggest the use of separate encoders for lead vocals, harmonic context, and instrumental timbre, followed by a shared decoder that generates multiple instrumental tracks. The decoder may operate on symbolic tokens or on neural audio codecs, each offering different trade-offs. Symbolic arrangement is easier to edit and integrate with digital audio workstations, while audio-domain arrangement can capture performance nuance but is more difficult to control and evaluate.

A significant challenge is the evaluation of arrangement quality. Objective metrics for pitch and harmony can be derived from music information retrieval tools, and classification models trained on musical audio [24] can provide coarse style and instrumentation signals. However, such metrics do not fully capture musical appropriateness or creative value. Expressive performance modeling [25] shows that human musicality involves subtle timing, dynamics, and articulation that are difficult to formalize. Therefore, automatic arrangement should be treated as a human-centered optimization problem rather than a purely statistical prediction task. The best system-level designs incorporate iterative user feedback, explicit constraint definitions, and interpretive controls for density, mood, and arrangement style.

5. Infrastructure and Training Data

Training singing voice and melody co-generation systems requires careful infrastructure planning because the data are multimodal, temporally complex, and often protected by rights. Singing-specific corpora with realistic scores [18] reduce the alignment burden but raise new questions about language coverage, recording quality, and the representation of vocal techniques. A production-oriented infrastructure must support ingestion of audio, symbolic scores, lyric annotations, and speaker metadata while preserving provenance. The data pipeline should also include robust alignment checks, because a small timing error between a syllable and a note can produce unnatural results. These requirements resemble the data engineering challenges in speech synthesis, but the musical dimension adds harmonic, stylistic, and expressive metadata that speech corpora typically do not include.

Compute infrastructure is another major concern. Waveform-level generation and long-sequence modeling are computationally expensive, and singing voice generation often requires high sampling rates to preserve breath and harmonic detail. System designers must decide whether to generate in a compressed spectral or discrete token domain and then reconstruct audio, or to generate waveforms directly. The compressed approach lowers memory and latency requirements but introduces codec artifacts that may be perceptible in sustained vocal tones. Direct waveform generation can produce higher fidelity but is less practical for interactive applications. These trade-offs are well known from speech synthesis [2], [4], but singing voice imposes stricter demands on pitch stability and timbral richness.

Deployment infrastructure must also support versioning, monitoring, and rollback. Generative music systems are nondeterministic by design, which complicates quality assurance and debugging. A modular architecture with separate melody, vocal, and arrangement components can facilitate targeted updates, but it also requires clear contracts between modules. Infrastructure teams must track not only model weights and code versions but also training data provenance, hyperparameter configurations, and evaluation results. Such practices are increasingly necessary for responsible deployment of creative artificial intelligence, particularly when systems may be used in commercial music production or public performance.

6. Robustness, Fairness, and Evaluation

Robustness in singing voice and melody co-generation is complicated by the subjective nature of musical quality. A model may produce technically correct pitch and rhythm while failing to produce a convincing vocal performance. Evaluation therefore cannot rely solely on objective metrics such as pitch accuracy or mel-cepstral distortion. Speech-oriented metrics may penalize the very expressiveness that defines singing, while instrument-oriented metrics may ignore lyric intelligibility and vocal timbre. Robust evaluation requires a layered approach that includes objective measures for pitch and timing, perceptual tests for naturalness, and musical assessments by trained listeners. The use of music information retrieval and music classification models [24] can provide quick diagnostic signals, but these signals should be interpreted cautiously because they are themselves trained on incomplete and culturally specific categories.

Fairness is another important consideration. Singing voice models are often trained on datasets that overrepresent certain languages, vocal styles, and gender categories. This can lead to systems that perform well on mainstream commercial styles but poorly on traditional, regional, or underrepresented singing traditions. The choice of language and phoneme encoders also affects fairness, because a model trained primarily on one language may produce incorrect syllable stress or vowel quality in another. Fairness in this context is not only about demographic parity but also about cultural representation and the preservation of musical diversity. System designers should audit training data and model outputs across languages, registers, and vocal techniques, and should document limitations openly.

Another robustness risk is the propagation of errors across coupled modules. In a modular pipeline, a subtle error in melody generation may cause the vocal synthesizer to produce an unnatural pitch contour, which in turn misleads the arrangement model. End-to-end systems may reduce this problem by sharing gradients, but they can also obscure the source of an error. Robust architectures therefore need diagnostic tools that can trace failures back to specific modules or training stages. In deployed systems, this might involve logging intermediate

representations, running component-level unit tests, and maintaining a curated set of difficult cases for regression testing.

7. Governance, Deployment, and Sustainability

Governance of singing voice and melody co-generation must address intellectual property, voice identity, and consent. Singing voice models can be used to mimic real singers, which raises legal and ethical questions that are more acute than in instrument-oriented music generation. A system that generates a new melody in the style of a particular vocalist may not directly clone that vocalist, but the boundary between stylistic imitation and identity replication is often ambiguous. Governance mechanisms should include clear terms for voice data collection, consent for training, and restrictions on voice conversion or cloning. The use of large singing corpora [18] requires robust licenses that separate lyric, composition, performance, and recording rights, because these rights may be held by different parties.

Deployment in creative industries requires interfaces that support human agency. Professional composers, producers, and vocalists are unlikely to adopt systems that operate as black boxes. They need mechanisms for controlling melodic contour, pronunciation, vocal register, ornamentation, arrangement density, and harmonic palette. Systems that expose hierarchical controls aligned with musical concepts are more useful than those that only accept abstract text prompts. At the same time, deployment should account for user skill levels, from novice creators who require guidance to experts who require fine-grained editorial control. This requirement influences both model design and interface engineering, and it reinforces the value of modular architectures that can be adapted to different user roles.

Sustainability also deserves attention. Large generative models consume substantial energy during training and inference, and the music sector is not exempt from environmental concerns. The trade-off between high-fidelity waveform generation and efficient token-based generation has direct implications for carbon footprint. Sustainable deployment may involve smaller specialized models, distillation, caching, and on-device inference where feasible. Governance policies should encourage transparency about compute usage and promote shared benchmarks that compare efficiency as well as quality. Finally, long-term sustainability depends on open academic collaboration and reproducible evaluation, which can prevent the concentration of generative music technology in a small number of proprietary platforms.

8. Conclusion

Singing voice and melody co-generation with automatic arrangement is not simply an extension of text-to-speech synthesis or symbolic music generation. It is a coupled modeling problem that requires simultaneous control of linguistic, musical, and acoustic variables across multiple timescales. This paper has argued that system-level design choices are central to progress in the field. Architectural decisions about modularity, representation, and hierarchical control shape both technical performance and usability. Data infrastructure must support multimodal alignment and provenance, while evaluation must move beyond narrow objective metrics toward musical and cultural validity. Robustness and fairness are not secondary concerns but core design requirements in systems that generate vocal performance.

Future work should develop structured benchmarks that include diverse languages, vocal techniques, and arrangement styles. It should also investigate governance models that protect singers and composers while enabling creative use. Advances in generative architectures will remain important, but they must be accompanied by careful attention to infrastructure, evaluation, and policy. A sustainable research agenda for intelligent music composition will

depend on collaboration among computer scientists, musicians, linguists, and legal scholars. Such collaboration can help ensure that singing voice and melody co-generation systems are expressive, controllable, fair, and accountable.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
2. van den Oord, A., 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. Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., ... & Kavukcuoglu, K. (2018). Efficient neural audio synthesis. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2410-2419). PMLR.
4. 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.
5. 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.
6. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779-4783). IEEE.
7. 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. In *Interspeech 2017* (pp. 4006-4010). ISCA.
8. Dhariwal, P., Jun, H., Payne, C., Kim, J., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.
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. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., ... & Adi, Y. (2024). Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.
11. Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2018). MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 34-41.
12. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., ... & Eck, D. (2019). Music Transformer: Generating music with long-term structure. In *International Conference on Learning Representations*.
13. Briot, J.-P., Hadjeres, G., & Pachet, F.-D. (2020). Deep learning techniques for music generation: A survey. *Neural Computing and Applications*, 32(20), 16171-16200.

14. Herremans, D., Chuan, C.-H., & Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Computing Surveys*, 50(5), 1-30.
15. Ji, S., Yang, X., & Luo, J. (2023). A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, 56(1), 1-39.
16. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., ... & Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.
17. Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., ... & Ellis, D. P. W. (2014). mir_eval: A transparent implementation of common music evaluation metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference* (pp. 367-372).
18. 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.
19. 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.
20. Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014). MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference* (pp. 155-160).
21. Simon, I., Morris, D., & Basu, S. (2008). MySong: Automatic accompaniment generation for vocal melodies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 725-734). ACM.
22. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017). Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1068-1077). PMLR.
23. Hsu, W.-N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Cao, Y., & Wang, Y. (2019). Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations*.
24. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392-2396). IEEE.
25. Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32(4), 955-967.