

Multilingual Singing Recognition and Music Semantic Understanding for Cross-Cultural Music Analysis

Bohit Gaman

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

rohitraman94@ku.edu

Deepak Ahuja

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.

deepak1994@uab.edu

Abstract

The computational analysis of singing has traditionally developed along separate trajectories from multilingual speech processing and music semantic understanding. Speech-oriented foundation models have produced increasingly reliable lyric transcription for several high-resource languages, while music information retrieval systems have focused heavily on Western instrumental repertoires. However, cross-cultural music analysis requires more than the concatenation of these capabilities. It demands an integrated system architecture that can represent phonetic content, melodic structure, performance style, cultural context, and semantic intent in mutually interpretable ways. This paper provides a system-level examination of multilingual singing recognition and music semantic understanding for cross-cultural music analysis. It discusses architectural choices, including modular pipelines and end-to-end audio-language models, and analyzes trade-offs between scalability, controllability, interpretability, and cultural specificity. The paper further examines data infrastructure, corpus governance, robustness, fairness, deployment sustainability, and policy implications. It argues that cross-cultural singing analysis should be understood not primarily as a statistical pattern recognition problem, but as a socio-technical infrastructure challenge in which data sovereignty, annotation quality, evaluation validity, and deployment context shape what computational systems can responsibly claim to understand. The discussion draws on speech processing, music information retrieval, generative audio systems, and critical dataset scholarship to outline a forward-looking research agenda that integrates technical design with cultural and institutional considerations.

Keywords

multilingual singing recognition, music semantic understanding, cross-cultural music analysis, audio-language models, corpus governance, fairness, cultural data sovereignty.

1. Introduction

Recent advances in large-scale audio processing have substantially changed the possibilities for analyzing vocal music. Weakly supervised speech recognition models now provide robust multilingual transcription across many languages, demonstrating that large pretrained systems can absorb substantial acoustic variability when exposed to sufficiently diverse training data

[1]. In parallel, generative audio models have shown that text-conditioned music generation can produce stylistically coherent outputs, revealing a growing capacity to link natural language descriptions with musical structure [2]. Contrastive audio-language embedding methods have extended this linkage by aligning music audio and textual descriptions in shared representational spaces, supporting retrieval and classification tasks without requiring fully supervised category structures [3]. These developments rest on a longer tradition of large music metadata collections that have enabled reproducible experimentation and benchmarking [4]. Together, they suggest that a more integrative approach to multilingual singing recognition and music semantic understanding is technically plausible.

Yet singing presents problems that neither speech recognition nor music tagging systems fully address. Singing combines linguistic content with melodic pitch, rhythm, vocal timbre, ornamentation, and expressive timing. Across cultures, these components are organized according to different tonal systems, tuning practices, performance conventions, and poetic forms. A system that merely maps sung audio onto lyrics in a majority language may fail to capture the semantic functions of vocal style, pronunciation, register, and repetition. Similarly, a music semantic model trained largely on Western popular or classical recordings may impose categories that are inappropriate for maqam-based improvisation, Hindustani raga exposition, Peking opera vocal roles, or West African griot performance. Consequently, cross-cultural analysis requires a system-level perspective that treats multilingual singing recognition not as an isolated transcription task, but as one component of a broader infrastructure for music semantics.

2. From Monolingual Singing Recognition to Cross-Cultural Music Semantics

The development of self-supervised audio representation learning has reduced the dependence of speech systems on manually transcribed corpora. Learned representations from masked acoustic modeling have been shown to transfer effectively across languages and domains, including noisy and expressive speech [5]. Multilingual speech corpora have extended this capability by providing aligned recordings and text across many languages, enabling researchers to evaluate cross-lingual transfer and language-specific error patterns [6]. Neural architectures for music classification have similarly benefited from convolutional and recurrent structures that capture local timbral patterns and longer temporal dependencies [7]. Yet automatic music transcription remains a difficult problem even for Western staff notation because of polyphony, overlapping instruments, and timbral ambiguity [8]. Singing transcription adds the complexities of co-articulated syllables, pitch glides, vibrato, and pronunciation variation that differ from both read speech and instrumental performance.

Cross-cultural music semantics further complicates the relationship between acoustic features and meaningful categories. A transcription model may correctly recognize phonemes without understanding whether a particular vocal technique carries ritual significance, marks a role shift in a theatrical work, or signals emotional intensification. A semantic system may assign mood tags or genre labels that are technically plausible but culturally incomplete. Therefore, cross-cultural music analysis must move beyond monolingual accuracy metrics and toward architectures that can represent multiple levels of musical meaning, including lyrical denotation, musical grammar, performance practice, and social function. This shift has implications for how corpora are designed, how models are evaluated, and how deployment systems are governed.

3. Architectural Foundations and System-Level Trade-offs

System architectures for multilingual singing recognition can be broadly organized along a spectrum from modular pipelines to integrated end-to-end models. A modular design might separate vocal source separation, pitch tracking, phonetic decoding, lyric alignment, and semantic tagging. This approach offers interpretability and permits component-level diagnostics. However, it introduces cascading errors in which mistakes made during source separation propagate into phonetic decoding and semantic inference. An integrated model, by contrast, can optimize acoustic and linguistic representations jointly, potentially improving robustness to noisy recordings and singing styles. The trade-off is that integrated models often require substantially more data and can be difficult to audit because latent representations may not correspond cleanly to musically meaningful units. In practice, hybrid designs have become common, with self-supervised audio encoders feeding task-specific heads for singing voice detection, lyric transcription, and semantic projection.

Beyond architectural choices, system-level design must address documentation and accountability. Dataset documentation frameworks have been proposed to make training data more transparent and to expose assumptions about collection, annotation, and intended use [9]. Critical analyses of large language and audio models caution that scale alone does not guarantee appropriateness for marginalized communities, since representational harm can be amplified through unexamined training distributions [10]. In the music domain, analyses of artificial intelligence systems have raised questions about how generative models relate to copyright, style imitation, and the engineering processes that encode normative assumptions [11]. These concerns are particularly acute for multilingual singing systems because the categories of language, genre, and cultural tradition are themselves contested and historically constructed.

4. Multilingual Lyric Transcription and Semantic Alignment

A recent multilingual singing corpus with realistic music scores has been introduced to support singing transcription, synthesis, and related tasks across multiple techniques [12]. Such a resource contributes to the unification of lyric annotation, note alignment, and acoustic modeling by providing multi-task supervision under realistic performance conditions [12]. Its design illustrates a broader shift toward corpora that treat singing as a distinct modality rather than as a variant of speech or instrumental music. Nevertheless, corpus availability alone does not resolve semantic alignment. Even when lyric and note annotations are aligned, the cultural interpretation of a sung phrase may require knowledge of language-specific poetic conventions, regional pronunciation norms, and performance contexts that are not encoded in the score.

Multicultural perspectives in music information research have long emphasized that Western tonal and rhythmic assumptions are not universally applicable [13]. Open music collections have improved reproducibility and enabled large-scale experimentation, but they continue to exhibit genre and language skew because of platform biases and regional access patterns [14]. Generative systems that accompany singing demonstrate the value of jointly modeling vocal acoustics and symbolic accompaniment structure, yet they also reveal how difficult it is to preserve stylistic nuance across performance traditions [15]. These observations suggest that multilingual lyric transcription and music semantic understanding should be developed together, with cultural metadata treated as a first-class component rather than an optional post hoc label.

5. Music Semantic Understanding Across Cultural Repertoires

Music semantic understanding encompasses more than lyric meaning. It includes recognition of formal structure, modal organization, rhythmic cycles, expressive timbre, and performance roles. Scaling speech technology to more than one thousand languages has shown that multilingual acoustic modeling can be extended to many low-resource settings, although the practical difficulty of obtaining sufficient high-quality data remains considerable [16]. Multimodal audio-text embedding models have demonstrated that shared semantic spaces can support open-vocabulary audio classification and cross-modal retrieval [17]. Singing voice detection, a foundational component for downstream vocal analysis, remains challenging when recordings contain overlapping instruments, non-Western tuning, and ornamented vocal production [18]. These limitations matter for cross-cultural music analysis because they affect which musical traditions can be analyzed reliably and which acoustic behaviors become visible to computational systems.

The operationalization of music semantics across cultures also raises conceptual difficulties. Universal tagging vocabularies may flatten culturally specific distinctions, while highly specialized vocabularies may limit cross-corpus comparability. A system trained to detect Western chord progressions may misrecognize heterophonic texture as noise, just as a lyric model trained on syllable-timed languages may struggle with mora-timed or tonal languages. Semantic understanding must therefore be sensitive to the interaction between language, tuning, rhythm, and performance context. This requires evaluation protocols that include cultural domain experts and that report disaggregated performance across language families, musical systems, and recording conditions.

6. Data Infrastructure, Annotation, and Corpus Governance

Building sustainable multilingual singing corpora requires infrastructure that goes beyond waveform storage and metadata tables. The FAIR principles for scientific data management emphasize that data should be findable, accessible, interoperable, and reusable [19]. For music data, these principles imply standardized identifiers, versioned annotations, clear licensing, and documented provenance. In cross-cultural settings, annotation cannot be reduced to automatic lyric alignment or crowd-sourced tags. It must involve performers, musicologists, linguists, and community stewards who can verify that transcriptions and semantic descriptions are appropriate. Corpus governance arrangements need to specify who may access sensitive cultural recordings, how derivatives may be used, and how attribution and benefit sharing will be handled.

Annotation pipelines for multilingual singing often combine automatic pre-alignment with human correction. This hybrid model can improve efficiency, but it introduces dependency on automatic systems that may be biased toward majority language phonetics or Western rhythmic notation. Documentation frameworks should therefore record not only final labels but also disagreement patterns, annotator backgrounds, and revision histories. Such practices make it possible to distinguish model error from legitimate variation in performance practice. They also support longitudinal corpus maintenance, which is essential for cross-cultural research because musical traditions change, and the meaning of a recording may shift as its community context evolves.

7. Robustness, Fairness, and Bias in Cross-Cultural Deployment

Ethical analyses of music information retrieval have emphasized that technical systems do not operate outside social relations and that algorithmic outputs can shape how musical traditions are represented and valued [20]. Fairness in multilingual singing recognition cannot be

evaluated solely by aggregate word error rates. Instead, performance must be disaggregated by language, gender, age, vocal register, recording quality, and musical style. A system may perform well on studio-recorded pop songs in English and Mandarin while failing on field recordings of Indigenous vocal music or religious chant. Such failure is not merely a technical limitation; it can reinforce cultural hierarchies by making some traditions appear less legible to digital platforms and archival systems.

Bias can enter at every stage of the system lifecycle. Training corpora may overrepresent commercially available music and underrepresent ritual, ceremonial, and community-based practices. Evaluation benchmarks may use metrics derived from Western staff notation that do not apply to oral traditions. Deployment interfaces may encourage users to query music using rigid genre labels that ignore hybridity and migration. Robustness must therefore include robustness to distribution shift across cultural contexts, but also organizational robustness through documentation, auditing, and dispute resolution. Fairness requires interdisciplinary participation in model design and evaluation, especially from scholars and practitioners who understand the meaning systems of the traditions being analyzed.

8. Deployment Architectures and Sustainability

Deployment architectures for cross-cultural singing analysis must balance computational cost, latency, privacy, and cultural appropriateness. Large cloud-based models can provide high throughput and frequent updates but may be inappropriate for recordings governed by community access protocols. On-device models can preserve local control and reduce connectivity dependence, but they may have limited capacity for rare languages or complex musical structures. Hybrid architectures can support local feature extraction with selective cloud querying for high-uncertainty cases, although this approach requires careful privacy engineering and bandwidth planning. Sustainability also depends on avoiding repeated retraining from scratch; modular adaptation and parameter-efficient fine-tuning can reduce energy use and allow communities to update models without replacing entire systems.

Long-term deployment requires more than technical maintenance. It involves building institutional relationships with archives, universities, cultural organizations, and local practitioners. Systems that treat communities as mere data sources are unlikely to produce sustainable or trusted research infrastructure. Conversely, systems that include community governance mechanisms may be slower to develop but more resilient to contestation and more capable of representing cultural nuance. Sustainability therefore includes organizational continuity, open documentation, and ongoing dialogue about the purposes of analysis. This is especially important when the same infrastructure may serve ethnomusicological research, music education, archival indexing, and commercial recommendation systems.

9. Policy, Intellectual Property, and Cultural Data Sovereignty

Multilingual singing recognition intersects with several policy domains, including copyright, data protection, intangible cultural heritage, and Indigenous data sovereignty. Recordings may contain multiple rights holders, including composers, lyricists, performers, and record labels. Lyrics are often separately protected as literary works, and traditional songs may be governed by community protocols rather than individual copyright. Systems that train on large web-crawled corpora may inadvertently reproduce copyrighted material or sacred songs without appropriate permission. Policy frameworks must therefore distinguish between private intellectual property and collective cultural rights. They must also address whether model

outputs, such as generated lyrics or synthesized singing, are derivatives that require licensing or cultural consent.

Cultural data sovereignty implies that communities have the right to determine how their music is collected, described, shared, and used. This principle challenges the default assumption that publicly accessible audio is automatically available for training. In practice, culturally appropriate governance may involve tiered access, dynamic consent, or local review boards. Technical architectures can support these policies through access logs, usage restrictions, and model provenance records. However, policy cannot be solved by technology alone. It requires negotiation among researchers, platform operators, collecting institutions, and cultural authorities. Cross-cultural music analysis can become a productive site for developing more accountable artificial intelligence only if these governance questions are treated as central rather than secondary.

10. Conclusion

Multilingual singing recognition and music semantic understanding are converging as audio foundation models, multilingual corpora, and music-text alignment techniques mature. Yet the path from technical capability to responsible cross-cultural analysis is not automatic. It depends on system architectures that can accommodate diverse vocal and musical practices, data infrastructures that respect culturally specific meanings and access conditions, and evaluation protocols that disaggregate performance across traditions. The required GTSinger corpus represents one important step toward multi-task singing research with realistic scores, but corpora alone do not produce semantic understanding. Multilingual singing systems must be designed as socio-technical infrastructures that connect acoustic modeling with humanistic knowledge, community governance, and policy frameworks. Future work should develop architectures that retain interpretability while scaling across languages, annotation methods that document cultural context, and deployment models that support local ownership. Only then can computational music analysis begin to engage seriously with the plurality of the world's vocal traditions.

References

1. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (PMLR Vol. 202, pp. 28492–28518). PMLR.
2. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., ... & Roberts, A. (2023). MusicLM: Generating music from text. arXiv preprint arXiv:2301.11325.
3. Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., & Ellis, D. P. W. (2022). MuLan: A joint embedding of music audio and natural language. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (pp. 559–566). ISMIR.
4. Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (pp. 591–596). ISMIR.
5. 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.

6. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A large-scale multilingual corpus for speech research. In *Proceedings of Interspeech 2020* (pp. 1043–1047). ISCA.
7. 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* (pp. 2392–2396). IEEE.
8. Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic music transcription: Breaking the glass ceiling. In *Proceedings of the 20th International Society for Music Information Retrieval Conference* (pp. 100–107). ISMIR.
9. 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.
10. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
11. Sturm, B. L., Iglesias, M., Ben-Tal, O., Miron, M., & Gómez, E. (2019). Artificial intelligence and music: Open questions of copyright law and engineering process. *Arts*, 8(3), 115.
12. 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.
13. Serra, X. (2011). A multicultural approach in music information research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 151–156). ISMIR.
14. Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference* (pp. 316–323). ISMIR.
15. Donahue, C., Caillon, A., Roberts, A., Manilow, E., Esling, P., Agostinelli, A., ... & Engel, J. (2023). SingSong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*.
16. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ... & Auli, M. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(19), 1–52.
17. Guzhov, A., Raue, F., Hees, J., & Dengel, A. (2021). AudioCLIP: Extending CLIP to image, text and audio. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 976–980). IEEE.
18. Schlüter, J., & Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *Proceedings of the 16th International Society for Music Information Retrieval Conference* (pp. 121–126). ISMIR.
19. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

20. Holzapfel, A., Sturm, B. L., & Coeckelbergh, M. (2018). Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval*, 1(1), 44–55.