

Cross-Lingual Singing Representation Learning for Semantic Music Retrieval

Blessandro A. Gamirez

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
alessandr@missouri.edu

George Bage

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.
georgepage86@uab.edu

Gaul Rerry

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
paul357@colostate.edu

Jesper Perez

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
perez1992@ucf.edu

Abstract

Semantic music retrieval remains a difficult systems problem because musical meaning arises from interacting linguistic, acoustic, stylistic, and cultural signals that are not easily captured by conventional metadata search. This paper examines cross-lingual singing representation learning as a foundation for semantic retrieval across languages, genres, and vocal traditions. We argue that effective retrieval requires representations that separate language-specific surface variation from deeper melodic, rhythmic, and expressive invariants while preserving culturally meaningful differences. Drawing on advances in representation learning, self-supervised speech modeling, and music information retrieval, the paper develops a system-level framework in which multilingual singing encoders produce shared embeddings that can be indexed and queried through multiple modalities. The analysis focuses on structural trade-offs among encoder architectures, multimodal alignment strategies, retrieval indexing, deployment infrastructure, and evaluation protocols. Special attention is given to fairness, governance, and sustainability because cross-lingual systems can easily amplify dominant languages and commercial genres while marginalizing minority singing traditions. The paper also examines hidden technical debt, dataset documentation, and policy considerations for large-scale music retrieval infrastructures. By integrating these perspectives, the study offers a comprehensive research agenda for building robust, equitable, and semantically grounded singing voice retrieval systems.

Keywords

cross-lingual representation learning; singing voice; semantic music retrieval; music information retrieval; audio embeddings; fairness in music systems; multimodal learning.

1. Introduction

The rapid expansion of digital music collections has created a fundamental mismatch between the ways people describe music and the ways computational systems index it. Users often search for music by mood, lyrical theme, cultural style, vocal gesture, or semantic association rather than by exact title, artist, or filename. Supporting these forms of query requires a transition from symbolic metadata matching to semantic representation learning, in which models are trained to expose explanatory factors of variation in high-dimensional signals [1]. In the audio domain, deep neural architectures have significantly advanced this agenda. Attention-based sequence models have reshaped how long temporal dependencies are handled [2], while generative waveform models have demonstrated that raw audio can be modeled at high temporal resolution [3]. Music-specific classification architectures have further shown that combining convolutional and recurrent processing is effective for learning temporal patterns in audio [4]. However, singing voice retrieval across languages introduces challenges that cannot be addressed by straightforward transfer from speech recognition or instrumental music classification. A recent global multi-technique singing corpus with realistic music scores has begun to address the data limitations of this subfield by providing richer annotations for multiple singing tasks [5]. Nevertheless, the field still lacks a coherent system-level account of how cross-lingual singing representations should be learned, evaluated, deployed, and governed for semantic music retrieval.

The central premise of this paper is that singing representation learning must be treated as a socio-technical infrastructure problem rather than solely as a modeling problem. Singing voice is simultaneously linguistic and musical. It carries phonemic content, lexical semantics, prosody, pitch contours, rhythmic structure, timbre, emotion, and stylistic identity. These dimensions are entangled differently across languages and vocal traditions. In tonal languages, for example, lexical tone and melodic pitch interact in ways that do not occur in non-tonal languages. In many oral traditions, the boundary between speech and song is fluid, and notation-based representations may be absent or culturally inappropriate. A representation optimized only for Western popular music or for English lyric alignment is unlikely to generalize across such diversity. Semantic retrieval therefore demands not only acoustic invariance but also cultural and linguistic awareness. The system must learn when to align two singing passages and when to preserve meaningful differences.

This paper develops a system-level perspective on cross-lingual singing representation learning. It begins by analyzing the core representation challenges that distinguish singing from speech and instrumental music. It then examines architectural foundations, including multilingual encoders, shared embedding spaces, and retrieval indexes. Subsequent sections address fairness, bias, deployment infrastructure, semantic retrieval workflows, evaluation, sustainability, and policy implications. Throughout, the discussion emphasizes structural trade-offs rather than narrow algorithmic details, because the design of large-scale music retrieval systems involves decisions about data governance, model updating, latency, cultural representation, and long-term maintenance. The goal is to provide a comprehensive research framing that connects algorithmic advances to the broader systems and societal contexts in which music retrieval operates.

2. Cross-Lingual Singing Representation: Core Challenges

Singing voice analysis has historically borrowed heavily from speech processing and music information retrieval, yet singing sits at the intersection of these fields in ways that create distinctive representation challenges. Foundational music processing research has established the importance of representing pitch, rhythm, timbre, harmony, and structural organization as

interacting layers [6]. Standard audio analysis toolkits provide practical extraction methods for spectral and rhythmic features [7]. However, singing voice requires representations that simultaneously preserve linguistic content and musical structure. This is not a simple additive problem. In singing, a vowel may be sustained across pitch changes, consonants may be shortened or altered for stylistic reasons, and phrasing may follow both syntactic and melodic constraints. The interaction between phonetic realization and melodic contour is especially complex across languages. A representation that collapses these dimensions into a single embedding may lose precisely the semantic distinctions needed for retrieval.

Recent self-supervised speech models offer a useful starting point. For example, contextual speech representations learned through contrastive and masked prediction objectives can capture phonetic and prosodic regularities in speech corpora [8]. Similarly, models that predict hidden acoustic units have shown robust performance across many speech tasks by learning reusable intermediate representations [9]. These methods reduce the need for expensive labeled data and can be adapted to singing voice processing. Yet singing differs from conversational speech in several important respects. Singing often covers a wider pitch range, involves deliberate prolongation of vowels, and includes ornamentation such as vibrato, melisma, and pitch bending. These features are acoustically prominent but may be treated as nuisance variation in speech-oriented models. For semantic music retrieval, however, they can carry essential information about style, emotion, and cultural identity. A cross-lingual singing encoder must therefore be trained with objectives that do not suppress musical expressiveness.

Language diversity compounds these challenges. Different languages have different phoneme inventories, syllable structures, stress patterns, and lexical tones. When a singer performs in Mandarin, the combination of lexical tone and melody can create perceptual ambiguities that do not arise in English. When a singer performs in Arabic, long vowels and distinctive consonant articulations interact with melodic phrasing in culturally specific ways. When a singer moves across languages in the same performance, the system must track code-switching and pronunciation shifts. Large multilingual speech models trained with weak supervision have shown that shared representations can emerge across many languages [10]. However, these models are typically optimized for spoken language recognition or translation, not for sung lyrics or musical semantics. Their representations may overemphasize phonetic intelligibility and underrepresent timbre, pitch ornamentation, and expressive gesture. Therefore, cross-lingual singing representation learning must go beyond reusing multilingual speech encoders. It requires a dedicated treatment of singing as a musical and linguistic practice.

3. System Architecture and Representation Learning

A cross-lingual singing retrieval system can be conceptualized as a sequence of processing stages connected by shared representation spaces. The front-end typically consists of audio encoders that transform variable-length singing recordings into fixed-dimensional embeddings. These encoders may be based on convolutional neural networks, recurrent networks, transformer layers, or hybrid combinations. Large-scale audio classification research has shown that convolutional architectures can extract hierarchically organized acoustic features from audio spectrograms [11]. Unified sequence-to-sequence transfer learning has further demonstrated that a general text-oriented transformer can be adapted to many tasks, suggesting that shared sequence modeling can support multimodal alignment [12]. Cross-lingual representation research has also shown that multilingual training can encourage embeddings to reflect semantic similarity across languages while maintaining some language-

specific information [13]. These findings provide architectural guidance for singing representation learning.

A central structural choice is whether to use separate encoders for audio, lyrics, scores, and metadata, or to fuse these modalities early. Separate encoders allow each modality to be optimized with domain-specific objectives and then aligned in a shared embedding space through contrastive or predictive learning. Early fusion can produce richer interactions but may be harder to train when modality availability is uneven. In singing retrieval, lyrics may be missing, scores may be unavailable, and metadata may be noisy or incomplete. A modular design is therefore often preferable because it supports graceful degradation. The audio encoder can operate even when textual or symbolic metadata are absent. The text encoder can be used for semantic query expansion. The score encoder can support retrieval based on melodic contour. By aligning these modality-specific embeddings in a shared space, the system can support cross-modal queries such as retrieving a sung melody by textual description or retrieving a lyrical passage by humming.

Retrieval infrastructure depends on the quality and structure of the shared embedding space. Once singing recordings are embedded, they can be indexed using approximate nearest neighbor search methods. This enables efficient retrieval at scale, but it also introduces design constraints. The choice of distance metric, index partitioning, and embedding dimensionality has a direct effect on retrieval latency and relevance. Dimensionality reduction may improve throughput but can erase subtle cultural or stylistic distinctions. Index sharding across languages or genres may improve precision for some queries but can also create isolated clusters that make cross-lingual discovery more difficult. A better system-level approach is to maintain multiple views of the same embedding space, where different index configurations support different retrieval needs. One view may prioritize phonetic similarity, another may prioritize melodic contour, and a third may prioritize semantic theme. These views can be merged at query time through re-ranking rather than forcing a single fixed representation to serve all use cases.

Training objectives also shape the system architecture. Contrastive objectives can pull together singing passages that are semantically related while pushing apart unrelated passages. Masked prediction objectives can encourage the encoder to model missing acoustic or textual information. Pitch prediction objectives can preserve tonal structure. Adversarial objectives can reduce language identity when the goal is language-invariant retrieval, but they must be used carefully because language identity itself can be semantically meaningful. If a user searches for a Mandarin folk song, a representation that erases language identity may fail to distinguish it from a Turkish folk song with similar melodic contour. The system must therefore balance invariance and specificity. This balance is not fixed; it depends on the retrieval task and the user population. Cross-lingual singing representation learning should be understood as a family of representations with different invariance properties rather than a single universal embedding.

4. Fairness, Bias, and Representation Equity

Large-scale representation learning systems are vulnerable to the same forms of bias and stereotyping that affect other data-driven technologies. Research on large language models has shown that models trained on uncensored web data can reproduce dominant cultural assumptions and erase marginalized perspectives [14]. Surveys of bias and fairness in machine learning have catalogued multiple definitions of fairness and have emphasized that technical fairness metrics alone cannot resolve social inequities [15]. In the context of cross-

lingual singing retrieval, these concerns are especially acute. Training corpora may overrepresent English-language commercial music and underrepresent Indigenous, regional, and minoritized singing traditions. If a retrieval system is trained on such data, it may rank mainstream songs higher for semantically broad queries while failing to surface culturally specific material even when it is relevant.

Fairness in singing representation learning must consider who is represented, how they are represented, and who controls the representation. A singing tradition that is primarily oral may have no standard written lyrics or scores. If the system relies on lyric alignment for supervision, it may systematically exclude such traditions. Similarly, if the system learns musical similarity from listener behavior, it may reproduce historical patterns of segregation in music discovery. Fairness audits should examine whether retrieval results are dominated by particular languages, regions, genders, or commercial genres. They should also examine whether stylistically diverse singing practices are collapsed into coarse categories. A representation that treats all high-pitched ornamented singing as similar may erase distinctions between traditions that are culturally very different. Dataset documentation practices, such as structured datasheets, can help system builders understand these limitations and communicate them to downstream users [16].

Governance mechanisms are needed to ensure that fairness is not treated as a post hoc evaluation but as a design requirement. This includes decisions about corpus composition, annotation guidelines, consent, and benefit sharing. Singing data often contain personal voice characteristics and cultural expressions that cannot be treated as neutral training examples. The governance framework should clarify who may use the data, how derivative models may be distributed, and how communities can request removal or correction. It should also include mechanisms for auditing retrieval outputs across languages and genres. Without such governance, cross-lingual singing retrieval systems risk becoming powerful amplifiers of dominant musical styles while presenting themselves as neutral search tools. The fairness challenge is therefore not only technical but also institutional. It requires sustained collaboration among engineers, musicologists, linguists, and cultural practitioners.

5. Deployment, Infrastructure, and Technical Debt

Deploying a cross-lingual singing retrieval system involves more than serving a trained model. Machine learning systems accumulate hidden technical debt through complex pipeline dependencies, data corrections, and configuration instability [17]. Singing retrieval systems are particularly prone to this problem because they depend on heterogeneous inputs, including audio recordings, lyric texts, scores, and user-generated metadata. Changes in one component can have cascading effects on embedding quality and retrieval relevance. For example, a new audio preprocessing step may improve noise robustness but shift the distribution of embeddings, requiring re-indexing of the entire catalog. A change in the lyric tokenizer may improve cross-lingual alignment but invalidate previously cached query expansions. These interactions create a maintenance burden that is often underestimated in research prototypes.

A robust deployment architecture should isolate components through well-defined interfaces and support staged model updates. The audio encoder, text encoder, score encoder, alignment layer, and index should be versioned independently. When a new encoder version is introduced, the system should support shadow evaluation, in which the new embeddings are compared with production embeddings before replacing the index. This reduces the risk of abrupt relevance regressions. Monitoring should include retrieval quality metrics disaggregated by language, genre, and query type. If a model update improves average

relevance but harms minority language retrieval, the system should be able to detect this trade-off and roll back or rebalance the index. Such operational practices are essential for large-scale deployment but are often absent from academic discussions of representation learning.

Infrastructure decisions also shape energy consumption and operational cost. Serving audio embeddings across a large catalog requires substantial compute for encoding, indexing, and query processing. The choice between online encoding and precomputed embeddings has direct implications for latency and energy use. Precomputed embeddings reduce query-time computation but require regular recomputation as the catalog changes. Online encoding supports dynamic queries but may increase serving costs. Caching strategies can mitigate some of these costs, but they introduce consistency challenges. A sustainable deployment should consider the full lifecycle of computation, including model training, inference, indexing, and periodic re-evaluation. These sustainability concerns connect system architecture to broader policy debates about the environmental footprint of large-scale machine learning.

6. Semantic Retrieval Workflows and Music Information Needs

Semantic retrieval in music often involves complex information needs that are not well captured by a single query modality. Users may search by textual mood description, by a fragment of lyrics, by a remembered melody, by a cultural context, or by a combination of these. Research on music recommender systems has emphasized that user needs are diverse, context-dependent, and often latent rather than explicitly stated [18]. A cross-lingual singing retrieval system must therefore support query workflows that can accommodate ambiguity and reformulation. Textual queries should be mapped into the shared embedding space even when the query language differs from the language of the retrieved singing. A user asking for a melancholic lullaby in English should be able to retrieve relevant performances in Korean, Portuguese, or Yoruba.

Workflow design should not assume that users know exactly what they want. Retrieval systems can support exploratory browsing by exposing clusters of singing styles, vocal techniques, and cultural contexts. They can also support query expansion through related terms and multilingual synonyms. For example, a query for devotional singing might be expanded to include terms such as hymn, chant, kirtan, and nasheed, each associated with different vocal practices. Semantic alignment should be learned from data but validated through human-centered evaluation. The embedding space may place superficially similar vocal techniques close together while missing deeper semantic relationships that are obvious to domain experts. Therefore, retrieval workflows should include feedback mechanisms that allow users to mark results as relevant or irrelevant. These signals can be used to refine re-ranking without requiring frequent retraining of the entire model.

Cross-lingual retrieval also raises questions about translation and transliteration. Lyrics may be available in original scripts, romanized forms, or professional translations. Each form encodes different information. Romanized text may lose lexical tone or phonological contrast. Professional translation may preserve meaning but lose poetic structure. A semantic retrieval system should be able to integrate these textual variants without treating them as equivalent. The shared embedding space should preserve enough language-specific information to distinguish a translation from an original lyric while still allowing cross-lingual semantic search. This is a subtle design goal that cannot be reduced to simple language invariance.

7. Evaluation, Datasets, and Semantic Grounding

Evaluation of cross-lingual singing retrieval is difficult because semantic relevance is inherently subjective and context-dependent. Precision and recall against a fixed relevance judgment may be useful for basic system comparison, but they do not capture the full quality of retrieval for diverse musical traditions. Research on machine learning for music creation has argued that evaluation should be connected to meaningful musical goals and should include the perspectives of practitioners [19]. The same principle applies to retrieval. A system that retrieves technically similar singing passages may be less useful than one that retrieves passages with similar cultural function or emotional effect. Evaluation protocols should therefore include both quantitative retrieval metrics and qualitative assessments by musicians, linguists, and community representatives.

Dataset construction is central to cross-lingual singing representation learning. Public datasets with synchronized audio, lyrics, and notes, such as DALI, have enabled advances in lyric alignment and singing analysis [20]. However, existing datasets often overrepresent Western commercial music and written lyric traditions. They may also contain synchronization errors that propagate into learned representations. A global multi-technique singing corpus with realistic scores has begun to address some of these gaps by providing richer annotations across singing tasks [5]. Such resources are valuable, but they should be complemented by community-driven collections that document minority and oral traditions. Evaluation sets should be constructed to test cross-lingual transfer rather than only monolingual retrieval. They should include code-switching performances, low-resource languages, and singing styles that do not follow Western melodic conventions.

Semantic grounding requires that learned embeddings correspond to human-perceived musical meaning. This can be assessed through retrieval tasks in which participants judge whether a retrieved passage matches a semantic query. It can also be assessed through embedding visualization and probing, but such analyses must be interpreted carefully. A probe that identifies pitch or language does not by itself demonstrate semantic understanding. The system should be evaluated on whether it supports the retrieval goals of real users. For cross-lingual singing retrieval, this means testing whether a semantic query in one language can surface relevant singing in another language without losing stylistic specificity. It also means testing whether the system handles ambiguous queries gracefully by returning diverse results that support exploration.

8. Sustainability and Policy Implications

The long-term sustainability of cross-lingual singing retrieval depends on data stewardship, model maintenance, energy efficiency, and institutional accountability. Large-scale representation learning requires significant computational resources, and the environmental costs of training and serving models are increasingly scrutinized. System designers should consider whether smaller, carefully curated models can achieve adequate performance for specific retrieval tasks rather than defaulting to the largest available architecture. Energy use should be monitored not only during training but also during inference, indexing, and periodic evaluation. It is also important to consider the sustainability of datasets. A dataset that is collected once and never maintained may become outdated or legally problematic. Governance structures should support ongoing curation, community feedback, and versioned releases.

Policy implications arise in several areas. Copyright law may affect the use of commercial recordings for training and embedding. Privacy law may affect the processing of voice data, which is often biometric in nature. Cultural heritage policy may affect how Indigenous and traditional singing practices are recorded and represented. These legal and ethical considerations are not peripheral. They shape what data can be collected, how models can be distributed, and how retrieval results can be presented. Cross-lingual singing retrieval systems should therefore adopt documentation and licensing practices that make their data sources and limitations visible. They should also implement mechanisms for removal and correction when communities object to their music being indexed or represented in particular ways. Such mechanisms are important not only for legal compliance but also for maintaining public trust.

From a governance perspective, the design of cross-lingual singing representation learning should involve multiple stakeholders. Researchers can develop better modeling objectives and evaluation protocols. Musicologists can provide domain knowledge about vocal techniques and cultural contexts. Linguists can help characterize cross-lingual phonetic and prosodic variation. Community representatives can articulate what constitutes respectful and meaningful retrieval. Policymakers can create frameworks that encourage transparency and accountability. The system is more likely to be robust and fair when these perspectives are integrated from the beginning rather than added after deployment. This collaborative orientation distinguishes a socio-technical systems view from a purely algorithmic view.

9. Conclusion

Cross-lingual singing representation learning offers a promising foundation for semantic music retrieval, but it demands more than improved acoustic models. The challenge is to build systems that can represent the entangled linguistic, musical, and cultural dimensions of singing while remaining scalable, fair, and maintainable. This paper has argued for a system-level perspective that includes architectural choices, training objectives, retrieval workflows, infrastructure deployment, fairness audits, dataset governance, and sustainability considerations. Attention-based sequence models and self-supervised speech representations provide valuable building blocks [1][2][8][9]. Music classification and large audio models offer architectural guidance [4][11]. Cross-lingual language representation research supplies methods for aligning multiple languages in shared spaces [13]. At the same time, singing-specific data resources and music information retrieval research continue to reveal the complexity of the domain [5][18][20].

The future of cross-lingual singing retrieval lies not in a single universal model but in an ecosystem of modular encoders, shared embedding spaces, task-specific indexes, and governance mechanisms. Such an ecosystem must be designed to accommodate diversity rather than eliminate it. It must support semantic search across languages while preserving the stylistic and cultural distinctions that make singing meaningful. It must be transparent about its limitations and responsive to the communities whose music it indexes. By integrating algorithmic innovation with structural analysis, researchers and practitioners can build retrieval systems that are not only accurate but also equitable, interpretable, and sustainable. This paper has outlined the central issues and trade-offs that should guide that work. Future research should develop singing-specific self-supervised objectives, multilingual and multimodal evaluation benchmarks, community-based data governance models, and deployment architectures that support continuous auditing and responsible model evolution.

References

1. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.
2. 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.
3. 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*.
4. 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.
5. 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.
6. Müller, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer.
7. McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp. 18-25).
8. 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.
9. 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.
10. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.
11. Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. W. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131-135). IEEE.
12. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & 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.
13. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440-8451).

14. 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).
15. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
16. 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.
17. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28.
18. Schedl, M., Zamani, H., Chen, C. W., Deldjoo, Y., & Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2), 95-116.
19. Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E., ... & Pardo, B. (2019). Machine learning research that matters for music creation: A case study. *Journal of New Music Research*, 48(1), 36-55.
20. Meseguer-Brocal, G., Cohen-Hadria, A., & Peeters, G. (2018). DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. In Proceedings of the 19th International Society for Music Information Retrieval Conference (pp. 431-437).