

Self-Supervised Singing Representation Learning for Low-Resource Music Information Retrieval

Yoss Martin

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
martin214@missouri.edu

Leiliang Tan

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.
leiliang1986@uab.edu

Abstract

The representation of singing voice in machine-readable form has become a central problem in music information retrieval, especially in settings where annotated data are scarce or unevenly distributed across languages, genres, and vocal techniques. This paper presents a system-level examination of self-supervised singing representation learning as a framework for low-resource music information retrieval. It argues that singing voice presents distinctive challenges that generic speech and audio models do not fully address, including wide pitch ranges, expressive timing, ornamentation, phoneme prolongation, and strong coupling between linguistic and musical structures. The paper discusses contrastive, masked, and multimodal pre-training strategies, and analyzes their architectural and infrastructural implications. Rather than presenting a single technical model, the study focuses on structural trade-offs among data curation, model capacity, downstream adaptation, evaluation validity, robustness, fairness, and deployment sustainability. It explores how low-resource conditions reshape corpus design, transfer learning, and inference infrastructure. The paper further addresses governance and policy concerns, including informed consent, cultural representation, energy consumption, and the risk of homogenizing vocal expression. By integrating technical architecture with social and institutional considerations, the paper offers a broad research perspective on building singing representations that are not only accurate but also maintainable, equitable, and responsive to diverse musical practices.

Keywords

self-supervised learning; singing voice; low-resource music information retrieval; representation learning; audio infrastructure; fairness; sustainability.

1. Introduction

Singing voice is one of the most information-rich signals in music. It simultaneously encodes language, melody, rhythm, timbre, expressive dynamics, and cultural style. Music information retrieval systems that can robustly analyze singing voice are therefore valuable for tasks such as lyric transcription, singer identification, vocal technique classification, pitch tracking, music recommendation, and computational musicology. However, the construction of supervised systems for singing analysis has traditionally required large quantities of precisely aligned labels, including phoneme boundaries, note events, pitch contours, and singer metadata. These labels are expensive to produce, difficult to verify, and often

unavailable for minority languages, regional vocal styles, or historical recordings. Self-supervised representation learning offers a compelling alternative because it can exploit unlabeled or weakly labeled audio to learn representations that transfer across tasks and domains. Early contrastive predictive frameworks demonstrated that useful representations could emerge from predicting future latent states [1]. Speech-oriented models built on this idea substantially improved low-resource speech recognition [2]. Masked prediction over learned speech units further showed that rich acoustic and linguistic structure could be learned without transcribed supervision [3]. These developments have encouraged the music information retrieval community to investigate whether similar methods can produce robust singing representations.

This paper approaches the problem from a system-level perspective. It does not propose a narrow architectural modification. Instead, it examines how self-supervised learning interacts with the larger infrastructure of music data, evaluation, deployment, and governance. The central argument is that low-resource music information retrieval is not only a machine learning problem but also an institutional and infrastructural challenge. Effective singing representation systems must balance model expressiveness with data scarcity, computational cost with local usability, and technical performance with cultural fairness. By analyzing these interdependencies, the paper aims to clarify the design space for researchers and practitioners who seek to build singing-aware systems under constrained resources.

2. Self-Supervised Learning in Audio and Music

Self-supervised learning in audio has developed along several major lines. Contrastive methods train models to distinguish augmented views of the same audio segment from other segments. This approach has been shown to produce general-purpose audio representations that transfer to diverse downstream tasks [4]. In the music domain, contrastive learning has been adapted to exploit music-specific structure, including harmonic and timbral similarity, and has improved performance on classification and tagging tasks when labeled data are limited [5]. Another line of work treats audio as a sequence of discrete events and learns representations through codified audio language modeling [6]. More recently, multimodal contrastive objectives have aligned music audio with natural language descriptions, producing representations that support zero-shot retrieval and classification [7]. Related audio-language models have demonstrated that text can act as a flexible supervisory signal for music understanding [8]. These models indicate that self-supervision in music can be strengthened by combining acoustic context with external semantic information.

Despite these advances, supervised baselines remain important for understanding the value of pre-training. Convolutional recurrent architectures have long provided strong results for music classification and tagging [9]. Open source tools such as librosa have supported reproducible feature extraction and evaluation pipelines [10]. Annotation-intensive datasets have enabled detailed comparisons across tasks [11]. Source separation models have illustrated the importance of learning fine-grained time-frequency structure from mixtures [12]. Waveform-domain separation methods further showed that end-to-end architectures can learn useful representations without hand-designed spectral features [13]. In music tagging, pre-trained convolutional networks have become widely used as feature extractors [14]. These supervised systems build on general architectural advances such as deep residual learning [15]. However, their dependence on labeled data limits their reach in low-resource singing scenarios, where annotations are fragmented or absent.

3. Singing Voice Representation and Architectural Foundations

Singing voice differs from speech and instrumental audio in ways that have direct consequences for representation learning. Singing often involves sustained vowels, large pitch intervals, vibrato, intentional timing deviations, and simultaneous constraints from lyrics and melody. Vocal timbre varies widely across individuals and traditions, and the same song may be performed with very different ornamentation. A representation that captures only phonetic content may ignore essential musical information, while a representation that captures only pitch may fail to preserve lyric intelligibility. Recent resources have begun to address this complexity by providing large multi-technique singing corpora with realistic music scores for singing tasks [16]. Such resources are valuable because they connect audio recordings with musical structure rather than treating singing as an extension of speech. Nevertheless, the availability of a large corpus does not automatically solve the low-resource problem. It mainly provides a foundation from which more targeted and culturally diverse datasets can be constructed.

Modern representation learning architectures rely heavily on self-attention and deep bidirectional context. The transformer architecture provides a flexible mechanism for modeling long-range dependencies in sequential data [17]. Bidirectional masked modeling, originally developed for text, has been adapted to speech and music to learn representations by predicting hidden segments from surrounding context [18]. For singing, masked modeling may encourage the model to integrate phonetic, prosodic, and harmonic cues because the model must infer missing frames using multiple sources of structure. However, transformer-based models are computationally demanding. Their benefits must be weighed against the cost of pre-training and fine-tuning in low-resource environments. A system-level analysis therefore requires careful attention not only to what a model can learn but also to whether the required compute is available and maintainable.

4. Low-Resource Music Information Retrieval

Low-resource conditions in music information retrieval take many forms. They may involve a small number of recordings, limited labels, under-resourced languages, scarce metadata, or insufficient computational infrastructure. In such settings, transfer learning from pre-trained models is often the most practical strategy. The usefulness of pre-trained representations depends on the alignment between the source domain and the target singing style. For example, a model pre-trained on Western pop singing may transfer poorly to traditional vocal practices with different tuning systems, vocal registers, or rhythmic conventions. Evaluation therefore needs to be carefully contextualized. Transparent metric implementations such as `mir_eval` have provided consistent evaluation for many music tasks [19]. Broad audio event datasets have supported large-scale pre-training for general audio understanding [20]. However, these resources may not capture the particular features of singing voice that matter in low-resource communities.

Low-resource transfer also raises questions about how models should be adapted. Full fine-tuning may be impossible when the target dataset is very small. Frozen feature extraction or lightweight adaptation layers may be more robust but can limit the model’s ability to learn new vocal timbres or languages. System designers must consider whether the pre-trained model contains sufficient invariance to irrelevant factors such as recording conditions and sufficient sensitivity to relevant factors such as pitch ornaments and phonetic contrasts. The interplay between model capacity, adaptation strategy, and target data size is a structural trade-off that cannot be resolved by architectural change alone. It requires careful empirical study and the development of culturally situated benchmark tasks.

5. Robustness, Fairness, and Evaluation

Robustness and fairness are increasingly central to the evaluation of learned representations. Large-scale pre-training datasets often overrepresent commercially prominent music, which can lead to representations that perform well on mainstream genres but poorly on marginalized vocal traditions. Critical scholarship has shown that large models can amplify biases present in their training data and can produce fluent but socially harmful outputs [21]. Dataset documentation practices such as datasheets provide a mechanism for making data collection and intended use more transparent [22]. The broader ethics of algorithmic systems highlights the need to examine not only average performance but also differential impacts across communities [23]. In singing representation learning, fairness concerns include language coverage, gender representation, stylistic diversity, and the accurate treatment of non-Western vocal techniques. A model that misclassifies a traditional singing style as poor pitch accuracy is not merely inaccurate; it encodes a normative judgment about musical correctness.

Evaluation protocols must therefore move beyond aggregate metrics. Disaggregated evaluation by language, gender, age, and musical genre can reveal hidden failures. Listening tests and qualitative musician feedback can complement automatic metrics. In low-resource settings, evaluation data may be as scarce as training data, which increases the risk of overfitting to small test sets. System designers should document the limitations of evaluation datasets and avoid overstating generality. Fairness in representation learning is not a final property that can be achieved once. It requires ongoing monitoring, community engagement, and revision of both data and model objectives.

6. Infrastructure, Deployment, and Sustainability

The deployment of singing representation systems in low-resource environments depends heavily on infrastructure. Large self-supervised models require significant energy for pre-training and inference. Energy and policy analyses have demonstrated that deep learning in language processing can impose substantial environmental costs [24]. Similar concerns apply to audio and music, especially when models are pre-trained repeatedly on large corpora. System-level design should therefore consider model distillation, quantization, and selective activation so that useful singing representations can be served on modest hardware. Edge deployment may be necessary in community archives, mobile recording projects, or educational settings without stable cloud access. In these contexts, smaller models with acceptable performance may be preferable to larger models with marginal gains.

Sustainability also involves the maintenance of datasets and tools. A low-resource singing corpus may be collected through a short-term project but then remain inaccessible due to broken links, unclear licenses, or obsolete formats. Sustainable infrastructure requires persistent identifiers, versioned metadata, open documentation, and community stewardship. It also requires computational reproducibility. If a pre-trained model cannot be run on available hardware, its scientific value is limited. The design of singing representation systems should therefore treat sustainability as a first-class requirement rather than an afterthought.

7. Governance and Policy Implications

Governance of singing representation systems involves decisions about data collection, model release, and downstream use. Singing voice data often contains personally identifiable information because the human voice can be used to identify individuals. Informed consent

must account for the fact that singers may agree to share a performance without anticipating that their voice will be used to train a general-purpose model. Dataset documentation practices support accountability but are not sufficient without enforceable consent mechanisms [22]. The ethics of algorithmic decision-making further emphasizes the need for transparency and contestability when models are used in cultural or educational settings [23]. Policy frameworks should consider how vocal data are collected from communities with different norms about ownership, performance, and intellectual property.

There is also a risk that self-supervised models contribute to cultural homogenization. If pre-training is dominated by a few commercial music traditions, the resulting representations may push diverse singing practices toward a narrow set of acoustic norms. Policymakers, archivists, and researchers should consider how to support local data sovereignty and community-controlled model training. Governance should not only restrict harmful uses but also invest in the creation of diverse and accessible training resources. The long-term value of self-supervised singing representation learning depends on whether it enables a wider range of musical practices to be studied and preserved, rather than further erasing those that are already underrepresented.

8. Conclusion

Self-supervised singing representation learning has significant potential for low-resource music information retrieval. By learning from unlabeled or weakly labeled audio, it can reduce dependence on expensive annotations and improve transfer across tasks. However, the technical promise of self-supervision cannot be separated from the systems in which it is embedded. Architectural choices, data curation, evaluation practices, fairness constraints, computational infrastructure, and governance all shape whether a representation system is effective and sustainable. Singing voice is not merely another audio category. It is a carrier of linguistic, musical, and cultural meaning, and its representation requires careful attention to variability and context. Future research should combine singing-specific pre-training objectives with community-centered data practices, transparent evaluation, and efficient deployment. By treating low-resource music information retrieval as a system-level challenge, the field can build representation learning methods that are technically rigorous and socially responsible.

References

1. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.
2. 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.
3. 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.
4. Saeed, A., Grangier, D., & Zeghidour, N. (2021). Contrastive learning of general-purpose audio representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3875–3879). IEEE.

5. Spijkervet, J., & Burgoyne, J. A. (2021). Contrastive learning of musical representations. In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR).
6. Castellon, R., Donahue, C., & Liang, P. (2021). Codified audio language modeling learns useful representations for music information retrieval. In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR).
7. 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 (ISMIR).
8. Manco, I., Benetos, E., Quinton, E., & Fazekas, G. (2022). Contrastive audio-language learning for music. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR).
9. 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.
10. 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).
11. 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 (ISMIR).
12. Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR).
13. Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). Music source separation in the waveform domain. arXiv preprint arXiv:1911.13254.
14. Pons, J., & Serra, X. (2019). musicnn: Pre-trained convolutional neural networks for music audio tagging. In Proceedings of the 20th International Society for Music Information Retrieval Conference Late-Breaking Demos.
15. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778).
16. 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.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
18. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

19. 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 MIR metrics. In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR).
20. Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 776–780). IEEE.
21. 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).
22. 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.
23. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.
24. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).