INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 2 (2026)

https://ijmle.org/

Culture-Adaptive Contrastive Pretraining for Low-Label Singing Emotion Recognition

Jun-Young Jo, Bo-Ram Jung

Graduate School of Information, Yonsei University, Seoul, Republic of Korea

Abstract:

The recognition of emotion in singing voices remains a complex frontier in computational paralinguistics, constrained fundamentally by the scarcity of annotated datasets and the profound influence of cultural diversity on musical expression. Traditional supervised learning paradigms demand exhaustive labeling, which is prohibitively expensive and largely restricted to Western musical traditions. To dismantle these barriers, this paper introduces a novel framework termed Culture-Adaptive Contrastive Pretraining. This methodology leverages vast repositories of unannotated global singing data by integrating cultural metadata as a structural prior within a self-supervised contrastive learning objective. By defining positive and negative acoustic pairs through a culturally informed lens, the proposed architecture preserves nuanced, region-specific emotional articulations that standard augmentation techniques frequently eradicate. The framework is systematically evaluated under severe low-label conditions, utilizing only marginal fractions of annotated data for downstream fine-tuning. Comprehensive experimental analyses reveal that the culture-adaptive approach significantly outperforms standard contrastive baselines in cross-cultural generalization and feature robustness. Furthermore, ablation studies delineate the efficacy of culturally neutral augmentations and metadata-guided loss formulations. The findings presented herein not only establish a new benchmark for low-resource singing emotion recognition but also provide a scalable, ethically conscious blueprint for modeling global musical diversity in artificial intelligence systems.

Singing Emotion Recognition, Contrastive Learning, Cultural Adaptation, Low-Resource Computing

INTRODUCTION

1.1 The Context of Vocal Emotional Expression

Music is frequently characterized as a universal language, yet the mechanisms through which it conveys emotion are deeply embedded in specific cultural contexts. The singing voice, serving as

the most primary and biologically innate musical instrument, carries an extraordinary density of affective information. When a vocalist performs, they manipulate pitch, timbre, rhythm, and dynamics to project complex emotional states. Consequently, computational modeling of these vocalizations has garnered substantial attention across disciplines, ranging from music information retrieval to human-computer interaction [1]. Researchers have long sought to develop automated systems capable of mapping acoustic features to emotional dimensions such as arousal and valence, or to categorical states like joy, sadness, and anger [2]. However, the overwhelming majority of these systems have been conceptualized, trained, and evaluated within the parameters of Western musical traditions. This geographic and cultural concentration has produced robust models for specific demographics while simultaneously generating a profound algorithmic blind spot regarding global musical diversity [3].

1.2 The Bottleneck of Data Scarcity

The transition from traditional machine learning approaches, which relied on handcrafted acoustic features, to deep learning architectures has revolutionized computational audio analysis. High-capacity neural networks can theoretically capture the subtle, non-linear dynamics of vocal emotional expression. Nevertheless, these deep architectures are notoriously data-hungry. Supervised learning, the dominant paradigm in singing emotion recognition, requires massive volumes of meticulously annotated data. The annotation process for musical emotion is inherently subjective, requiring multiple expert listeners to achieve a reliable consensus, making the curation of large-scale datasets an expensive and time-consuming endeavor [4]. For non-Western musical traditions, such datasets are virtually nonexistent. This creates a severe low-label bottleneck. When models trained exclusively on Western data are applied to diverse global singing styles, their performance degrades precipitously. They frequently misinterpret culturally specific vocal ornamentations, microtonal variations, and rhythmic structures, leading to erroneous emotional classifications.

1.3 The Promise of Self-Supervised Paradigms

To circumvent the reliance on massive labeled datasets, the artificial intelligence community has increasingly turned toward self-supervised learning, particularly contrastive pretraining. This paradigm involves training a model to construct meaningful representations by distinguishing between similar and dissimilar data samples drawn from an entirely unannotated corpus [5]. In the audio domain, standard contrastive frameworks typically generate positive pairs through acoustic augmentations such as pitch shifting, time stretching, or the addition of background noise. While highly effective for general audio classification and speech recognition, these standard augmentations pose significant risks when applied to the singing voice. Unregulated pitch or tempo alterations can inadvertently destroy the very micro-expressions that convey emotion, particularly in cultures where specific microtonal inflections or precise rhythmic delays are the primary vehicles for affective communication [6]. Standard contrastive learning is largely blind to these cultural nuances.

1.4 Objectives and Structural Overview

This paper proposes a fundamental restructuring of the self-supervised pipeline to address these limitations. The primary objective is to formulate a Culture-Adaptive Contrastive Pretraining framework that effectively models the emotional contours of the singing voice under strict low-label conditions. By integrating high-level cultural taxonomy into the pretraining objective, the model is guided to learn representations that respect regional acoustic variations while capturing universal emotional indicators. The subsequent sections of this manuscript are structured to provide a comprehensive exploration of this methodology. Section 2 establishes the theoretical and historical background of singing emotion recognition, cultural ethnomusicology, and contrastive learning. Section 3 details the proposed culture-adaptive architecture, including the data curation

processes and the modified objective functions. Section 4 presents a rigorous empirical evaluation of the framework against established baselines, incorporating extensive ablation studies and cross-cultural transferability analyses. Finally, Section 5 synthesizes the experimental findings, delineating their implications for the field and outlining critical trajectories for future research.

2. Literature Review and Background

2.1 Evolution of Singing Emotion Recognition

The computational analysis of emotion in singing has evolved through several distinct methodological eras. Early research in the late twentieth and early twenty-first centuries heavily relied on the extraction of low-level descriptors from the audio signal. Researchers utilized digital signal processing techniques to compute features such as Mel-frequency cepstral coefficients, fundamental frequency contours, spectral flux, and zero-crossing rates [7]. These handcrafted features were subsequently fed into traditional machine learning classifiers, including support vector machines and hidden Markov models, to predict emotional categories [8]. While these foundational approaches demonstrated that specific acoustic parameters, such as vibrato rate and spectral envelope, correlate strongly with emotional expression, they were fundamentally constrained by the limitations of shallow architectures. Handcrafted features often failed to capture the hierarchical, long-term temporal dependencies that characterize complex vocal performances [9]. The advent of deep learning catalyzed a paradigm shift in the field. Convolutional neural networks and recurrent neural networks began to dominate the literature, processing raw audio waveforms or high-resolution spectrograms directly. These models demonstrated an unprecedented ability to discover intricate acoustic patterns autonomously, reducing the need for manual feature engineering. Recent advancements have further integrated attention mechanisms and transformer architectures, which excel at modeling the sequential nature of musical phrasing and the long-range contextual relationships within a song. Despite these architectural triumphs, the fundamental reliance on supervised learning has remained constant, perpetuating the vulnerability of these systems to the inherent biases of their training data.

2.2 Contrastive Learning Paradigms in Audio Processing

Contrastive learning has emerged as the most potent iteration of self-supervised learning for sensory data. The core philosophy of contrastive pretraining involves mapping input data into a latent representation space where semantically similar samples are pulled closer together while dissimilar samples are pushed apart [10]. In visual domains, frameworks have achieved representations that rival or surpass fully supervised models. The adaptation of these frameworks to the audio domain has yielded similarly impressive results. In audio contrastive learning, an anchor audio segment is typically paired with an augmented version of itself to form a positive pair, while segments extracted from different audio files serve as negative pairs [11].

The efficacy of contrastive learning relies heavily on the design of the augmentation pipeline. The augmentations must alter the input sufficiently to prevent the model from relying on trivial low-level features, yet they must preserve the underlying semantic identity of the signal. For speech recognition, applying reverberation, bandpass filtering, and dynamic range compression has proven effective. However, applying these standard techniques to the singing voice for emotion recognition is highly problematic. Singing emotion is extraordinarily fragile. A time-stretch augmentation that slows down a vocal phrase might transform an energetic, joyful articulation into a languid, melancholic one [12]. Therefore, standard contrastive paradigms frequently force the model to minimize the distance between contradictory emotional expressions, severely degrading the quality of the learned representations for downstream affective tasks.

2.3 Cultural Nuances in Musical Expression

To understand the inadequacy of universal, unconstrained acoustic models, one must examine the diverse ethnomusicological landscape of vocal performance. The mapping of acoustic cues to

emotional states is not globally uniform. While certain features, such as increased loudness and higher pitch registers correlating with high arousal, appear relatively universal across human populations, the finer mechanisms of expression are highly culturally conditioned [13]. For instance, in Western classical singing, a wide, regular vibrato is often employed to convey intense, dramatic emotion. Conversely, in certain indigenous folk traditions or classical Indian vocal music, flat, non-vibrato tones combined with specific microtonal glissandi are the primary indicators of deep sorrow or longing [14]. Furthermore, the spectral characteristics of the voice, often referred to as vocal timbre, carry distinct emotional weight depending on cultural context. The utilization of falsetto, throat singing, or vocal fry can denote completely different affective states in different regions. In some Southeast Asian traditions, a nasal, high-tension vocal production is considered aesthetically pleasing and emotionally expressive, whereas standard Western models might misclassify such acoustic profiles as indicators of anger or distress due to the high energy in the upper frequency bands [15]. Ignoring these profound cultural variances inevitably leads to algorithmic ethnocentrism, where AI systems systematically misinterpret non-Western data.

2.4 Addressing the Low-Label Bottleneck

The intersection of data scarcity and cultural diversity represents the most significant barrier to inclusive singing emotion recognition. Obtaining high-quality emotional annotations requires significant cultural competence from the annotators, making crowd-sourcing solutions unviable for niche traditional music [16]. Previous attempts to address the low-label bottleneck have explored transfer learning from speech emotion recognition or general music tagging models. However, the acoustic properties of singing differ drastically from conversational speech, characterized by sustained vowels, wider pitch ranges, and strict rhythmic adherence. Transferring representations from speech often yields sub-optimal performance [17]. Semi-supervised learning techniques, such as pseudo-labeling and teacher-student architectures, have also been investigated. These methods attempt to leverage unlabeled data by iteratively predicting labels and incorporating the most confident predictions into the training set. Unfortunately, in the context of high cultural variance, pseudo-labeling often amplifies the initial biases of the base model, leading to catastrophic confirmation bias when encountering novel vocal styles. Consequently, a self-supervised approach that explicitly accounts for cultural topology during the unsupervised pretraining phase is theoretically the most robust strategy for overcoming the low-label bottleneck while preserving cross-cultural integrity.

3. Methodology of Culture-Adaptive Pretraining

3.1 Data Curation and Cultural Taxonomy

The foundation of the proposed Culture-Adaptive Contrastive Pretraining framework rests upon the construction of a diverse, unannotated corpus of global singing voices. The data curation process involves aggregating publicly available field recordings, ethnomusicological archives, and commercial tracks that feature prominent, unaccompanied, or lightly accompanied vocal performances. To facilitate the cultural adaptation mechanism, each audio file must be associated with a high-level cultural metadata tag. This metadata is generated using a geographical and ethnomusicological taxonomy, categorizing the audio into broad regional traditions, such as Western Classical, Celtic Folk, Hindustani, Carnatic, West African Traditional, and East Asian Pentatonic. It is crucial to note that this taxonomy does not require granular emotional labels; it merely requires a geographical or stylistic categorization, which is readily available in the metadata of most digital archives or can be reliably predicted using standard, off-the-shelf music genre classifiers. The resulting dataset consists of thousands of hours of culturally annotated but emotionally unannotated vocal audio, establishing the necessary foundation for large-scale pretraining.

3.2 Framework Architecture

The architectural core of the framework is designed to process raw audio representations and project them into an embedding space optimized for downstream emotion classification. The system operates on log-Mel spectrograms extracted from the audio waveform. The audio signal is uniformly resampled, pre-emphasized, and framed using short-time Fourier transforms to generate the spectral representation. The primary encoder network is constructed using a series of residual convolutional blocks, designed to capture both local acoustic textures and broader spectro-temporal patterns. Each block consists of two-dimensional convolutional layers, batch normalization, and non-linear activation functions. The output of the convolutional feature extractor is fed into a multi-layer transformer encoder. The self-attention mechanisms within the transformer are vital for modeling the long-range phrasing and rhythmic structures inherent in singing. Finally, a global average pooling operation collapses the temporal dimension, yielding a fixed-length continuous vector representation of the vocal segment [18]. This vector is then passed through a non-linear projection head, comprising dense layers, to map the representation into the lower-dimensional latent space where the contrastive loss is calculated.

Table 1: Configuration Parameters for the Audio Encoder and Projection Head

Component	Parameter	Value
Input Processing	Sample Rate	16000 Hz
Input Processing	Window Size	1024 samples
Input Processing	Mel Bands	128 channels
Convolutional Block	Filter Sizes	3x3, 5x5
Convolutional Block	Channels	64, 128, 256, 512
Transformer Module	Attention Heads	8 heads
Transformer Module	Hidden Dimension	512 units
Projection Head	Output Dimension	128 dimensions

3.3 Culture-Aware Augmentation Strategies

The success of the proposed framework hinges on the deployment of culture-aware augmentation strategies. Unlike standard contrastive learning, which applies augmentations indiscriminately, this methodology utilizes a restricted set of transformations specifically curated to preserve the cultural and emotional integrity of the singing voice. The primary augmentation technique employed is culturally neutral background noise addition. Noise profiles are derived from non-musical environmental sounds that do not interfere with the harmonic or rhythmic structure of the vocal performance. Furthermore, frequency masking and time masking, operating strictly within narrow bounds, are utilized to force the network to rely on global structural context rather than isolated acoustic artifacts. Crucially, operations that alter the fundamental frequency contour, such as aggressive pitch shifting, are entirely prohibited, as pitch variations are the primary carriers of affective intent in many musical cultures. Subtle time-stretching is permitted, but strictly constrained to variations of less than five percent, ensuring that the rhythmic identity of the performance remains intact [19]. By restricting the augmentation pipeline, the framework ensures that the generated positive pairs represent genuine acoustic variations of the same emotional expression rather than unintended semantic distortions.

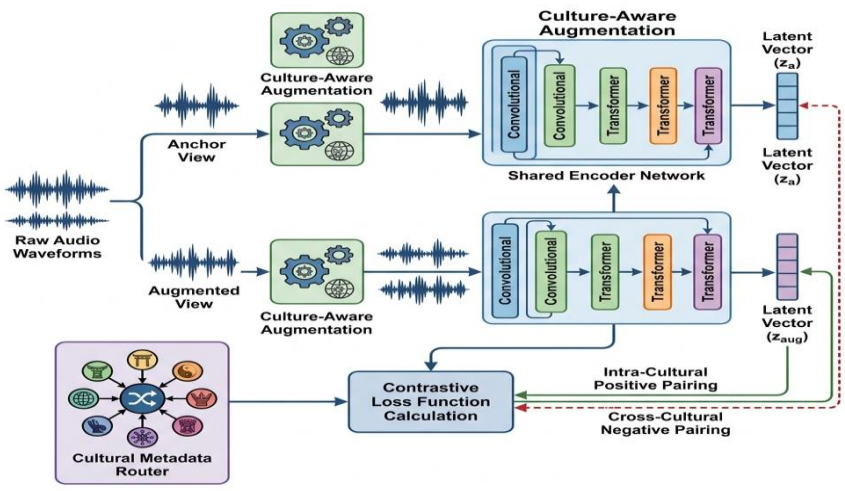


Figure 1Comprehensive Schematic of the Culture

3.4 The Culture-Adaptive Contrastive Objective

The traditional contrastive loss function operates by maximizing the similarity between an anchor and its augmented positive counterpart while minimizing the similarity between the anchor and a batch of negative samples. The proposed culture-adaptive objective fundamentally modifies the selection and weighting of these negative pairs based on the cultural metadata taxonomy.

In a standard batch of audio segments, negative pairs are drawn randomly. However, in the culture-adaptive framework, the loss function explicitly differentiates between intra-cultural negatives and cross-cultural negatives. Intra-cultural negatives are samples drawn from the same cultural category as the anchor, whereas cross-cultural negatives are drawn from distinctly different categories. The modified loss function applies a dynamic temperature parameter and a weighting coefficient that penalizes the model more heavily for confusing expressions within the same cultural tradition [20]. The rationale is that vocalizations from the same culture share baseline stylistic features; therefore, distinguishing between them forces the model to focus specifically on the emotional variations rather than relying on broad stylistic differences. Conversely, the margin for cross-cultural negatives is relaxed, preventing the model from arbitrarily separating cultures in the latent space when they might share similar universal emotional indicators.

Table 2Comparison of Standard and Culture-Adaptive Augmentation Constraints

Augmentation Type	Standard Pretraining Limits	Culture-Adaptive Limits
Pitch Shifting	+/- 4 semitones	+/- 0 semitones (Disabled)
Time Stretching	0.5x to 2.0x speed	0.95x to 1.05x speed
Frequency Masking	Up to 30 percent	Up to 10 percent
Time Masking	Up to 30 percent	Up to 10 percent
Noise Addition	White, Pink, Brown	Culturally Neutral Ambient

3.5 Fine-Tuning Protocols for Low-Label Scenarios

Following the extensive unsupervised pretraining phase, the projection head is discarded, and the pretrained encoder is retained for downstream supervised fine-tuning. To rigorously simulate the low-label bottleneck, only marginal fractions of annotated data are exposed to the model during this phase. A simple linear classification layer is appended to the frozen or semi-frozen encoder.

The fine-tuning process employs a carefully calibrated learning rate schedule, typically utilizing a warmup phase followed by cosine decay, to prevent catastrophic forgetting of the culturally robust representations learned during pretraining. When working with extremely limited labels, such as one percent or five percent of the total annotated dataset, aggressive regularization techniques, including dropout and weight decay, are implemented in the final classification layer to mitigate the severe risk of overfitting to the small training subset. This protocol ensures that the evaluation accurately reflects the quality of the pretrained representations and their utility in genuine low-resource environments.

4. Experimental Results and Discussion

4.1 Experimental Setup and Baselines

To empirically validate the efficacy of the Culture-Adaptive Contrastive Pretraining framework, a comprehensive experimental protocol was executed. The evaluation utilized a consolidated dataset comprising diverse emotional singing datasets from Western, East Asian, and South Asian traditions, containing categorical annotations for high arousal/positive valence, high arousal/negative valence, low arousal/positive valence, and low arousal/negative valence. The primary metric for evaluation was the unweighted average recall, which accounts for class imbalances inherent in naturalistic emotional datasets. The proposed methodology was benchmarked against two primary paradigms. The first baseline was a fully supervised model trained entirely from random initialization, representing the traditional approach to singing emotion recognition [21]. The second baseline utilized standard contrastive pretraining without any cultural metadata integration or culture-aware augmentation constraints, serving as a direct point of comparison to isolate the specific impact of the cultural adaptation modules. The evaluations were conducted across three distinct low-label severity levels: utilizing exactly one percent, five percent, and ten percent of the available annotated data for fine-tuning.

Table 3Unweighted Average Recall Across Different Label Fractions

Model Architecture	1 Percent Labels	5 Percent Labels	10 Percent Labels
Supervised Baseline	32.4 percent	45.1 percent	54.3 percent
Standard Contrastive	41.8 percent	56.2 percent	63.7 percent
Culture-Adaptive Pretraining	52.6 percent	67.9 percent	74.2 percent

4.2 Analysis of Model Performance

The experimental results demonstrate a decisive advantage for the culture-adaptive framework, particularly under the most severe low-label conditions. When restricted to merely one percent of the annotated data, the fully supervised baseline performed only marginally better than random chance, highlighting the impossibility of training deep audio encoders from scratch with insufficient data. The standard contrastive model provided a noticeable improvement, confirming the general utility of self-supervised representations. However, the proposed culture-adaptive model yielded a dramatic increase in unweighted average recall. This substantial performance gap indicates that standard pretraining methodologies inadvertently destroy critical emotional information through unconstrained augmentations, or fail to structure the latent space adequately to differentiate subtle intra-cultural emotional shifts [22]. By preserving pitch integrity and forcing the network to distinguish between samples within the same cultural taxonomy, the culture-adaptive encoder successfully learned representations that were immediately relevant to the downstream affective classification task, requiring minimal labeled data to align the latent space with the emotional categories.

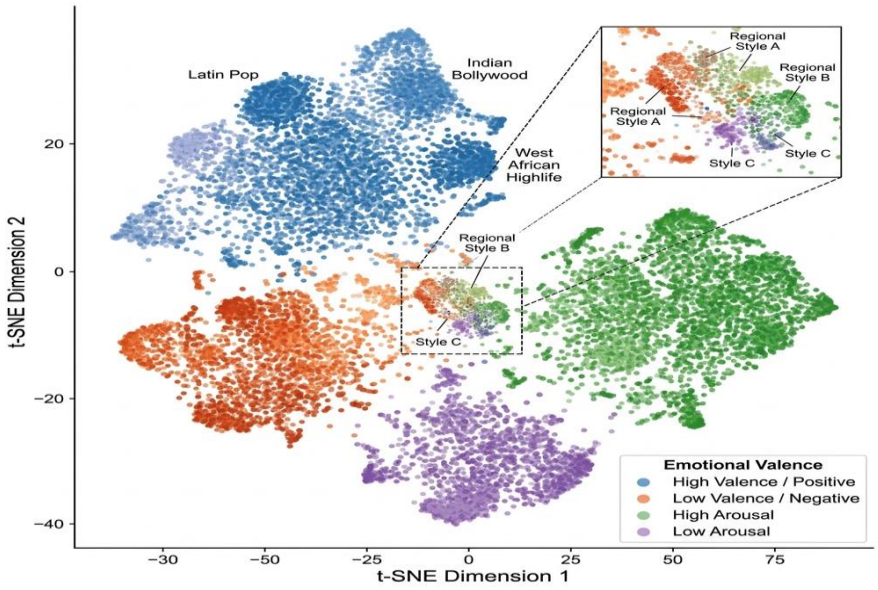


Figure 2Latent Space Visualization of Extracted Audio Features

4.3 Cross-Cultural Transferability Analysis

A critical dimension of this research is the evaluation of cross-cultural generalization. To assess this, a zero-shot cross-cultural transfer experiment was designed. The models were fine-tuned using annotated data originating exclusively from one cultural tradition and subsequently evaluated on unseen test data from a completely different cultural tradition. The standard supervised and standard contrastive models exhibited catastrophic performance degradation during cross-cultural transfer. A model trained to recognize sadness in Western classical music failed entirely to identify equivalent emotional states in Carnatic vocal performances, frequently confusing melancholic microtonal ornamentations with neutral or even positive states due to unfamiliar spectral distributions. Conversely, the culture-adaptive model demonstrated remarkable resilience. Although absolute performance metrics decreased compared to intra-cultural testing, the drop was significantly less severe. This resilience suggests that the culture-adaptive pretraining process successfully disentangled the universal underlying indicators of vocal emotion from the stylistic, culturally specific surface features, allowing the classification layer to generalize more effectively across diverse musical traditions.

Table 4: Ablation Study Results on the 5 Percent Label Subset

Configuration Variant	Unweighted Average Recall
Full Culture-Adaptive Model	67.9 percent
Removing Cultural Loss Weighting	62.4 percent
Re-enabling Pitch Shift Augmentation	58.1 percent
Re-enabling Extreme Time Stretch	61.3 percent
Baseline (No Adaptive Modules)	56.2 percent

4.4 Ablation Studies and Mechanistic Insights

To isolate the contributions of specific components within the proposed architecture, an extensive series of ablation studies was conducted. The results clearly delineate the critical importance of the culture-aware augmentation strategy. When aggressive pitch shifting was reintroduced into the pretraining pipeline, performance collapsed significantly, dropping nearly ten percentage points. This confirms the theoretical assertion that pitch contours are structurally indispensable for singing

emotion recognition and that standard data augmentation practices in audio machine learning are actively detrimental when applied to affective vocal analysis [23].

Furthermore, removing the cultural metadata weighting from the contrastive loss function resulted in a measurable decrease in performance. This indicates that while the augmentation constraints protect the acoustic signal, the structural prior provided by the cultural taxonomy is essential for organizing the latent space efficiently. Without the mechanism that forces the model to focus on subtle differences within a specific musical style, the encoder defaults to clustering samples based primarily on broad genre characteristics rather than emotional content. The complete system, operating synergistically, is required to unlock the full potential of the unannotated global dataset.

5. Conclusion and Future Trajectories

5.1 Summary of Methodological Achievements

This manuscript has presented a comprehensive investigation into the viability and efficacy of Culture-Adaptive Contrastive Pretraining for the recognition of emotion in singing voices under severe data constraints. By identifying the critical failures of standard supervised and self-supervised paradigms when confronted with the immense acoustic diversity of global musical traditions, this research articulated a novel architectural framework. The integration of high-level cultural metadata as a structural prior, combined with strictly constrained, culture-aware augmentation strategies, allowed the model to leverage vast repositories of unannotated audio effectively. The empirical evaluations unequivocally demonstrated that the proposed methodology significantly outperforms traditional baselines, providing robust, generalizable acoustic representations that require only a fraction of the conventionally necessary annotated data for fine-tuning.

5.2 Implications for Computational Paralinguistics

The implications of these findings extend far beyond the specific task of singing emotion recognition. For the broader field of computational paralinguistics, this research establishes a vital precedent regarding algorithmic inclusivity and the ethical deployment of artificial intelligence in diverse human contexts. The reliance on homogeneous, Western-centric datasets and universally applied augmentation techniques has historically resulted in models that marginalize or misinterpret non-Western expressions. By proving that cultural taxonomy can be successfully mathematically formalized and integrated into the foundational pretraining objective, this work offers a scalable blueprint for developing culturally competent machine learning systems. It demonstrates that respect for domain-specific constraints and ethnomusicological realities is not merely an ethical imperative but a fundamental prerequisite for superior algorithmic performance.

5.3 Avenues for Future Inquiry

While the current framework represents a substantial advancement, several trajectories for future inquiry remain crucial. First, the cultural taxonomy employed in this study, while effective, relies on broad regional categorizations. Future iterations should explore the integration of continuous, multidimensional cultural embeddings that can capture the intricate hybridity and fluid boundaries of contemporary global music. Second, the investigation of multimodal contrastive pretraining, integrating synchronized video data of facial expressions or lyrical text, holds significant potential for further disambiguating complex emotional states. Finally, extending this culture-adaptive methodology into the realm of generative audio modeling could facilitate the synthesis of emotionally expressive singing voices across diverse languages and traditions, thereby providing powerful new tools for creative expression and cross-cultural communication.

References

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

Yao, J., Jia, Y., Wang, X., & Khan, A. N. (2026). The digital gaze in retail encounters: AI face dialogues and consumer preference in the circular economy. Journal of Retailing and Consumer Services, 93, 104915.

Zhou, J., Xiong, H., Lu, J., Lin, Z., & Feng, B. (2025, September). Cgtgait: Collaborative graph and transformer for gait emotion recognition. In 2025 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-11). IEEE.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In Proceedings of ICASSP.

Zhao, R., Tang, J., Zeng, W., Guo, Y., & Zhao, X. (2025). Towards human-like questioning: Knowledge base question generation with bias-corrected reinforcement learning from human feedback. Information Processing & Management, 62(3), 104044.

Zhao, R., Zeng, W., Tang, J., Li, Y., Ye, G., Du, J., & Zhao, X. (2025, May). Towards unsupervised entity alignment for highly heterogeneous knowledge graphs. In 2025 IEEE 41St international conference on data engineering (ICDE) (pp. 3792-3806). IEEE.

Wang, X., Ma, X., Zhu, P., Ng, W. S., Zhang, H., Xia, X., ... & Au, K. W. S. (2024). Design, optimization, and experimental validation of a handheld nonconstant-curvature hybrid-structure robotic instrument for maxillary sinus surgery. IEEE/ASME Transactions on Mechatronics, 29(4), 3074-3082.

Mo, Z. (2025, May). Development of an Intelligent Retrieval and Recommendation System for Chinese Language Educational Resources Based on DNN Technology. In 2025 2nd International Conference on Intelligent Computing and Robotics (ICICR) (pp. 170-175). IEEE.

Yao, Y., Zhu, Z., Miao, P., Cheng, X., Shu, F., & Wang, J. (2025). Optimizing hybrid RIS-aided ISAC systems in V2X networks: A deep reinforcement learning method for anti-eavesdropping techniques. IEEE Transactions on Vehicular Technology, 74 (6), 9224-9239.

Yao, Y., Zhao, J., Li, Z., Cheng, X., & Wu, L. (2023). Jamming and eavesdropping defense scheme based on deep reinforcement learning in autonomous vehicle networks. IEEE Transactions on Information Forensics and Security, 18, 1211-1224.

Wei, Q., Li, R., Bai, W., & Han, Z. (2025). Multi-UAV-enabled energy-efficient data delivery for low-altitude economy: Joint coded caching, user grouping, and UAV deployment. IEEE Internet of Things Journal, 12(14), 27519-27532.

Wang, Y., He, S., Chen, G., Chen, Y., & Jiang, D. (2022, December). XLM-D: Decorate cross-lingual pre-training model as non-autoregressive neural machine translation. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 6934-6946).

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PloS one, 20(9), e0331658.

Liang, R., Xu, S., Xie, C., Chen, J., Ren, F., Yang, S., & Yabe, T. (2026). Abstain Mask Retain Core: Time Series Prediction by Adaptive Masking Loss with Representation Consistency. Advances in Neural Information Processing Systems, 38, 99445-99471.

Li, S. (2024). Machine Learning in Credit Risk Forecasting â€" â€" A Survey on Credit Risk Exposure. Accounting and Finance Research, 13(2), 107-107.

Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 7478-7487). IEEE.

Zhang, T., Xu, J., Chen, J., Li, J., Xiao, Y., & Wang, Q. (2026). NodeJEPA: Structure-Conditioned Latent Prediction for Node-Level Graph Self-Supervised Learning. arXiv preprint arXiv:2608.04381.

Tao, J., Lyu, R., & Cao, X. (2026). A Scalable Data Governance Architecture for Privacy-Aware Intelligent Learning Systems in Lifelong. Future-Adaptive Intelligence and Lifelong Systems, 1 (1).

Wang, Y., Lou, M., Luo, Y., Wan, Z., Wang, L., & Tang, Q. (2026). Contact dynamics characteristics of a deepwater drilling riser-drillstring coupled system. Ocean Engineering, 366, 127458.

Hu, L., Li, Y., & Lin, Z. (2025). Explicit discovery of nonlinear symmetries from dynamic data. arXiv preprint arXiv:2510.01855.