Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.39491

Multimodal Feature Recognition of Traditional Vocal Performance Styles for Digital Music Education

Junhui Chen

School of Music, South China Normal University, Guangzhou, China
wellington589125@gmail.com

Abstract. In traditional vocal performance, timbre organization, melodic ornamentation, articulation, and embodiment form an inseparable combination, requiring approaches to recognition which would take into account complementary audio and visual cues. In this work, we propose a lightweight framework for recognition of Beijing opera, Kunqu opera, and Yue opera, based on 1,200 twenty-second audiovisual clips. Frozen wav2vec 2.0 embeddings, log-Mel statistics, and interpretable features extracted from vocal recordings, together with MediaPipe Holistic normalized facial and upper-body movement descriptors, make up a 384-dimensional representation, which is then classified using a small multilayer perceptron. Evaluation separated by performer produces a multimodal macro-F1 of 0.914 ± 0.018, compared to 0.851 ± 0.024 for audio-only and 0.742 ± 0.031 for visual streams. In presence of simultaneous acoustic and visual interference, macro-F1 is 0.838 ± 0.025. A prototype implementation in Streamlit applies recognition to the selection of style through role-based comparison of audiovisual features, as well as generation of phrase-specific recommendations for improvement of vocal practice. On the basis of the analysis by instructors of 30 students, style consistency improves by 0.62 ± 0.41 points. These findings support complementary audiovisual characterization while distinguishing genre recognition from artistic quality assessment.

Keywords: Traditional Vocal Performance, Multimodal Recognition, Digital Music Education, Audio-Visual Fusion, Vocal Style Classification

1. Introduction

Traditionally, Chinese vocal performance makes stylistic distinctions with the help of the interaction between vocal timbre, melodic structure, vocal articulation, and action. Beijing opera, Kunqu opera, and Yue opera share fundamental components of performance in the theatrical form but have their own repertoires and performance practices. The findings on the classification of jingju timbre show that measurable spectral features are able to distinguish dramatic roles and culturally-specific identities [1]. Visual component is also important in musical performance. Research on Karnatak singing proves the connection between gesture kinematics and vocal acoustics; thus, human movement carries an information which is complementary to the sound [2]. Facial expressions and upper body movements should be analyzed in conjunction with phrasing, role, and intentions of the performer. Digital education requires recognition to provide valuable information to the learner. The

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.39491

systematic review of the feedback in music reveals some methodological flaws and calls for a more nuanced approach to evaluation of feedback processes [3]. The recent meta-analyses of the use of AI in music education also highlight the possibilities for personalization [4]. Consequently, the classifier which assigns a correct genre is a necessity but not enough if the feedback associated with it helps learners reflect and improve their performance. Classification accuracy should also be separated from the estimate of artistic skill level. In this research, the dataset of synchronized performances and a lightweight recognition system combining pre-trained embeddings, interpretable vocal features, and movement statistics is built. The contributions of this work lie in the quality-aware feature combination, performer-separated evaluation, and role-conditioned educational feedback. The novelty of complex deep learning architecture is not claimed.

2. Literature review

2.1. Acoustic properties of traditional vocal styles

The paralinguistics investigation of singing attributes shows that it is possible to capture particular vocal characteristics based on annotated acoustic data by applying machine learning [5]. It is important to note that attributes are independent of genres since similar methods are used in different traditions and roles. In addition, general audio research using self-supervised learning proves the usefulness of representations that combine both local and global information [6]. Thus, in the case of traditional singing, one has to assess the quality of pretrained embeddings along with particular descriptors.

2.2. Visual performance features in vocal performances

Computational studies of ragas have shown that kinematic properties of the gesture might be related to the features of the melody and offer additional information apart from audio [7]. Newer multimodal studies examine differences in the gestural-vocal interactions between different performers [8]. This finding validates temporal alignment and performer-independent validation. However, visual tracking might pick up on camera motion, anatomy, and gesture habits not associated with the style.

2.3. Multimodal recognition for digital music education

Studies in automatic performance assessment also make the difference between computational metrics and their educationally meaningful interpretation [9]. Systematic research on individual performance assessment also reveals diverse metrics, scales, and evaluative judgments [10]. Consequently, accuracy of recognition and educational value should be evaluated separately. In this situation, feature combination helps to solve the problem of genre classification, while feedback distinguishes observable features from those of professional models.

3. Methodology

3.1. Dataset construction and annotation

This data set includes 1,200 clips of 22 seconds each, which are divided equally between Beijing opera, Kunqu opera, and Yue opera. There are sixty actors who recorded twenty unique clips each, which corresponds to twenty actors in each genre. The metadata of the sources include information

DOI: 10.54254/2755-2721/2027.39491

about the dramatic character, repertory, availability of accompaniment, camera angle, and recording session. The genre categorization and phrase relevance are judged independently by three vocal teachers on graduate level. In case of discrepancies, the final decision is made through adjudication before including the clips into the data set. The requirement is the presence of the visible soloist, sound, and vocal track of it. The training, validation, and test parts contain 42, nine, and nine actors correspondingly, i.e., 840, 180, and 180 clips.

3.2. Audio features extraction

Librosa 0.10 is utilized for resampling vocal stems at 16 kHz. The use of a 512-point Hann window and a hop size of 160 samples produces 128-bin log-Mel spectrogram windows, thus maintaining a twenty-second temporal window. The Hugging Face Transformers framework is used to generate frozen wav2vec 2.0 base embeddings from the same waveform; the process of freezing is performed as an attempt to minimize overfitting from the small dataset. Mean and variance pooling of such embeddings alongside with log-Mel frequency statistics and hand-crafted features is done, and then a 256-dimensional projection is performed. Parsemouth predicts fundamental frequencies on voiced windows after an octave error filter is applied; the range of pitch is expressed in semitones. Vibrato rate is calculated from the detrended sustained note contours at the 3-9 Hz bandwidth.

3.3. Visual performance feature extraction

Decoding of video is done at 25 fps and audio timestamps are retained as they were originally. Facial landmarks and shoulder, elbow, and wrist landmark coordinates are obtained using MediaPipe Holistic. Normalization with respect to shoulder and scaling with inter-shoulder distances are done to account for the distance between the camera and the body size; facial ratios are computed using interocular distance. Facial features like mouth opening, head tilt, wrist velocity, and relative gesture amplitude are collected within two-second aligned windows. Landmarks that cannot be detected due to low visibility are masked, and interpolation is done only when there are less than 0.2 seconds worth of missing data.

3.4. Multimodal feature fusion and classification

Independent quality weights aggregate acoustic and visual windows into 256-dimensional audio and 128-dimensional visual vectors, forming a 384-dimensional input. Fusion and training are defined jointly as shown in Equation (1) - (2):

$$\mu_i^m = \frac{\sum_t q_{it}^m x_{it}^m}{\sum_t q_{it}^m} \tag{1}$$

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{3} y_{ic} \log p_{ic} + \lambda \sum_{\ell=1}^{3} \|W_\ell\|_F^2 \tag{2}$$

Here, m indexes modality, q denotes observation quality. Standard deviations are elementwise; unavailable dimensions use training-set medians with retained missingness indicators. The architecture of the neural network that was implemented by PyTorch 2.2 consists of 256 and 128

DOI: 10.54254/2755-2721/2027.39491

neurons at the hidden layers, ReLU activation function and dropout equal to 0.3. The parameters of Adam optimizer include learning rate equals 0.0001, batch size 32 and weight decay 0.0001.

3.5. Recognition and educational feedback module

In Streamlit, the system displays the predicted genre, adjusted confidence, and six interpretable features, namely pitch range, vibrato frequency, energy range, articulation frequency, mouth opening range, and upper body movement range. The professional means in the training set are grouped according to genre, dramatic role, and phrase types, with intervals representing variability observed. The learner decides on his/her intention of genre and role; therefore, he/she will not select the wrong reference due to the error of the prediction. Mappings authored by the instructor map the differences in the descriptors into listening and rehearsal tips related to a phrase. Inappropriate measurements are deliberately avoided. The feedback highlights the observable differences without enforcing greater movement and vibrato. Classification decisions associated with uncertainty require a second look from the teacher instead of issuing a score of proficiency. Figure 1 illustrates how acoustic and visual feature are combined for vocal style recognition.

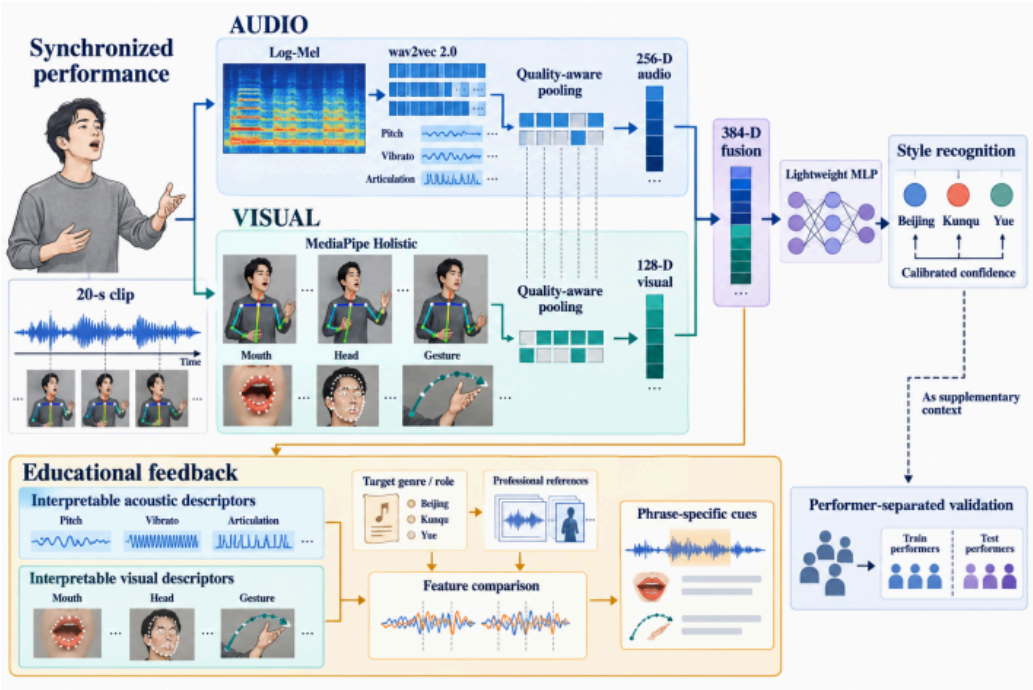

Figure 1. Synchronized audio-visual feature fusion and educational feedback framework for traditional vocal performance recognition

4. Experimental design and validation

4.1. Unimodal and multimodal comparison

Audio-only, visual-only, and hybrid classifiers employ identical performer groups, training and evaluation strategies. The main metric used is Macro-F1, along with additional metrics that include accuracy, macro-precision, macro-recall, and genre-specific confusion matrix in order to describe discrimination. Optimization differences are assessed across ten training seeds, while test set uncertainty is computed using 2,000 resamples of the performer clusters through the bootstrap

DOI: 10.54254/2755-2721/2027.39491

technique. Audio-only condition parameters utilize the same acoustic features as the hybrid condition.

4.2. Ablation and robustness experiments

Each component is ablated in isolation without embeddings, acoustic features, face information, and body movements, followed by retraining with same experimental conditions. Robustness evaluations involve addition of new background noise at 10 dB signal-to-noise ratio and removal of 20% of the frames. Corruption evaluation involves the degradation of the quality of the recording simultaneously with sequential frame losses, in addition to random frame loss. Features are re-extracted after corruption to make sure that masking and temporal pooling are done on the degraded signals.

4.3. Digital music education validation

Thirty undergraduate voice performance students performed six exercises, two for each genre, using the designated material with phrases of similar difficulty. The exercises involved making a recording, receiving automated feedback, and revising it once within a predetermined amount of practice time. The style consistency of the performances was assessed by three judges, blind to the order of the recordings and to any predictions made by the models, using five-point scales based on timbre, ornamentation, diction, and gesture. SPSS 29 was used to perform paired t-tests on the differences between the means at the individual student level, thus not producing pseudoreplication of data across tasks. Confidence intervals, paired effect sizes, and interrater agreement are provided. Participation required signed consent and IRB approval. Absence of a control group with no feedback is noted.

5. Results and discussion

5.1. Recognition performance of audio, visual, and multimodal features

Table 1 shows the strongest recognition for audiovisual fusion. Its paired macro-F1 gain over audio-only classification is 0.063 ± 0.017, with a performer-bootstrap 95% interval of 0.028–0.097. Genre-specific F1 reaches 0.938 ± 0.016 for Beijing opera, 0.892 ± 0.025 for Kunqu opera, and 0.912 ± 0.021 for Yue opera. Kunqu–Yue confusion remains the largest error concentration. Temperature scaling reduces expected calibration error from 0.081 ± 0.014 to 0.036 ± 0.009, supporting more cautious confidence display.

Table 1. Recognition performance under clean, ablated, and corrupted conditions

Model or condition	Accuracy	Macro-F1	Macro-precision	Macro-recall
Audio only	0.856 ± 0.023	0.851 ± 0.024	0.857 ± 0.022	0.850 ± 0.025
Visual only	0.750 ± 0.030	0.742 ± 0.031	0.751 ± 0.029	0.741 ± 0.032
Audio–visual fusion	0.917 ± 0.017	0.914 ± 0.018	0.918 ± 0.017	0.913 ± 0.019
Without pretrained embeddings	0.878 ± 0.022	0.873 ± 0.023	0.879 ± 0.021	0.872 ± 0.024
Without facial features	0.895 ± 0.020	0.891 ± 0.021	0.896 ± 0.020	0.890 ± 0.022
Noise plus frame loss	0.844 ± 0.024	0.838 ± 0.025	0.846 ± 0.023	0.837 ± 0.026

DOI: 10.54254/2755-2721/2027.39491

5.2. Contribution of acoustic and visual features to style discrimination

The removal of manually constructed acoustic features is associated with a decline in macro-F1 by 0.024 ± 0.010, and excluding upper body motion causes a decline in macro-F1 by 0.016 ± 0.008. Using shuffled pairs reduces macro-F1 to 0.870 ± 0.022, which may be indicative of audiovisual coordination. Noise on its own results in 0.862 ± 0.023, whereas frame loss alone results in 0.887 ± 0.020. In sum, these experiments demonstrate that there is complementary rather than equally dependent use of both modalities. Different feature sets can be described in the studied set of samples but not in terms of universal style conventions.

5.3. Educational feedback effectiveness and practical application

Instructor-assessed consistency increased from 3.08 ± 0.57 to 3.70 ± 0.49 points. The difference between the means was 0.62 ± 0.41 points with a 95% CI of 0.47–0.77, $t(29) = 8.28$, $p < 0.001$, and a paired effect size of 1.51. Inter-rater agreement attained ICC(2,3) = 0.86 with a 95% CI of 0.77–0.92. Articulation event identification resulted in an F1-score of 0.83 ± 0.04 after instructor review. The improvements can be attributed to immediate refinement; yet, repeated practice and instructor anticipation constitute viable alternative hypotheses as well.

6. Conclusion

The model includes acoustic embeddings, explainable vocal features, and normalized visual kinematics in a small classifier and educational interface. The fusion improves the performer-independent classification, whereas the quality-aware pooling allows for degradation of input data. The role-conditioned references establish the connection between numerical measures and specific training conditions and preserve the distinction between the probability of genre and the artistic proficiency.

References

[1] Liu, A. Y., & Wallmark, Z. (2024). Identifying Peking opera roles through vocal timbre: An acoustical and conceptual comparison between dan and laosheng. *Music & Science, 7*, 1–10.

[2] Pearson, L., & Pouw, W. (2022). Gesture–vocal coupling in Karnatak music performance: A neuro–bodily distributed aesthetic entanglement. *Annals of the New York Academy of Sciences, 1515*(1), 219–236.

[3] Blackwell, J., Matherne, N., & McPherson, G. E. (2023). A PRISMA review of research on feedback in music education and music psychology. *Psychology of Music, 51*(3), 716–729.

[4] Merchán Sánchez-Jara, J. F., González Gutiérrez, S., Cruz Rodríguez, J., & Syroyid Syroyid, B. (2024). Artificial intelligence-assisted music education: A critical synthesis of challenges and opportunities. *Education Sciences, 14*(11), 1171.

[5] Xu, Y., Wang, W., Cui, H., Xu, M., & Li, M. (2022). Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. *EURASIP Journal on Audio, Speech, and Music Processing, 2022*, 8.

[6] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2023). BYOL for audio: Exploring pre-trained general-purpose audio representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31*, 137–151.

[7] Nadkarni, S., Roychowdhury, S., Rao, P., & Clayton, M. (2023). Exploring the correspondence of melodic contour with gesture in raga alap singing. In *Proceedings of the 24th International Society for Music Information Retrieval Conference* (pp. 21–28).

[8] Pearson, L., Nuttall, T., & Pouw, W. (2025). The co-structuring of gesture-vocal dynamics: An exploration in Karnatak music performance. *Cognitive Science, 49*, e70137.

DOI: 10.54254/2755-2721/2027.39491

[9] Kim, H., Ramoneda, P., Miron, M., & Serra, X. (2022). An overview of automatic piano performance assessment within the music education context. In *Proceedings of the 14th International Conference on Computer Supported Education (Vol. 1)* (pp. 465–474). SciTePress.
[10] Moura, N., Dias, P., Veríssimo, L., Oliveira-Silva, P., & Serra, S. (2024). Solo music performance assessment criteria: A systematic review. *Frontiers in Psychology, 15*, 1467434.