Quantum-Enhanced Analysis and Grading of Vocal Performance

Rohan Agarwal *

Posted Date: 3 September 2025

doi: 10.20944/preprints202509.0281.v1

Keywords: quantum computing; music information re-trieval; audio analysis; vocal performance; hybrid classical–quantum; neural networks

 doi:10.20944/preprints202509.0281.v1

Article

Quantum-Enhanced Analysis and Grading of Vocal Performance

Rohan Agarwal

Monta Vista High School and De Anza College (dual enrollment), Cupertino, CA, USA; agarwalrohan2@student.deanza.edu

Abstract

Vocal singing is a profoundly emotional art form possibly predating spoken language, yet evaluating a vocal track remains a subjective and specialized task. Meanwhile, quantum computing shows promise to bring about significant advances in science and art. This study introduces *QuantumMelody*, a quantum-enhanced algorithm to evaluate vocal performances through objective metrics. QuantumMelody begins by collecting a comprehensive array of classical acoustic and musical features including pitch contours, formant frequencies, Mel-spectrograms, and dynamic ranges. These features are divided into three musically categorized groups, converted into scaled angles based on statistical metrics, and then encoded into specific quantum rotation gates. Each qubit group is entangled internally, followed by intergroup entanglement, thus exploring subtle, non-linear relationships within and across feature sets. The resulting quantum probability distributions and classical features are used to train a neural network, combined with a spectrogram transformer to holistically grade each recording on a 2–5 scale. Key difference metrics like the Jensen–Shannon distance and Euclidean measures of scaled angles are used to enable nuanced comparisons of different recordings. Furthermore, the algorithm uses classical music-based heuristics to provide targeted suggestions to the user for various aspects of vocal technique. On a dataset of 168 labeled 20 s vocal excerpts, QuantumMelody achieves 74.29 % agreement with expert graders. The circuits are simulated; we do not claim hardware speedups, and results reflect a modest, single-domain dataset. We position this as an applied audio-signal-processing contribution and a feasibility step toward objective, interpretable feedback in singing assessment.

Keywords: quantum computing; music information retrieval; audio analysis; vocal performance; hybrid classical–quantum; neural networks

1. Introduction

Assessing the quality of a singing performance is a very subjective and specialized task. This is especially important in music schools, where hundreds of student recordings might need review each week and become a severe bottleneck for music teachers. The motivation for this research is rooted in the need to offer a reliable automated system for feedback with clear, consistent pointers for improvement.

Prior research in singing voice analysis provides a foundation for an objective approach. For example, humans tend to judge performances by concrete aspects such as pitch accuracy, stability of tone, timing, and vocal timbre, even if their overall impressions are subjective. Work by Ghisingh *et al.* (2017) [1] analyzed Indian classical singing by isolating the vocal track from background music and examining acoustic production features like pitch and root-mean-square energy. Other researchers have built systems to grade singing on metrics including intonation correctness, rhythm consistency, and vibrato usage. For instance, Liu and Wallmark (2024) [2] trained machine learning models on annotated singer characteristics to classify timbre and technique attributes in traditional opera singing. Meanwhile, recent surveys by Hashem *et al.* (2023) [3] document speech emotion recognition systems that infer emotions from voice signals via deep learning. These advances illustrate that modern AI and

 doi:10.20944/preprints202509.0281.v1

machine learning (ML) can decipher subtle information from vocal audio, whether it be emotional tone or singing skill.

While classical ML approaches for voice grading are promising, nascent technologies like quantum computing open new frontiers for audio analysis. Quantum computing's ability to represent and entangle high-dimensional data offers a novel way to capture the complex interdependencies of musical features. Kashani *et al.* (2022) [4] implemented a note detection algorithm based on the Quantum Fourier Transform (QFT). Miranda *et al.* (2021) [5] envisioned quantum computing's impact on music technology and introduced frameworks for quantum music intelligence. Gündüz (2023) [6] used information-theoretic metrics like Shannon entropy to quantify musical complexity. These developments set the stage for a hybrid approach to the long-standing challenge of vocal performance evaluation.

QuantumMelody is proposed as a solution that combines robust audio feature analysis with quantum computing to achieve consistent and insightful vocal grading. First, the algorithm extracts a feature vector from raw singing audio, encompassing both traditional and innovative descriptors of vocal quality. These features capture pitch and intonation accuracy, frequency stability (jitter) and amplitude stability (shimmer), LUFS energy (loudness and dynamics), and timbral characteristics such as MFCCs and formant frequencies. Expressive elements like vibrato extent and rate are also included.

2. Materials and Methods

2.1. Dataset and Preprocessing

We collect 20 s vocal recordings in seven Hindustani ragas, with 42 labeled samples each for ratings of 2, 3, 4 and 5 across the seven ragas, for a total of 168 samples. All audio is converted to mono and resampled to 22 050 Hz using Librosa's high-quality resampler. We then apply denoising and drone removal by attenuating low-frequency harmonics corresponding to a tanpura drone via a narrow band-stop filter around the drone frequency. This is followed by harmonic–percussive separation to extract the pure vocal track. After filtering, we use a short-time Fourier transform (STFT) [1] to verify that the drone, percussion, and any low-frequency noise are removed while preserving the singer's voice.

2.2. Feature Extraction

We extract a suite of time-domain and frequency-domain features for each recording as follows.

Pitch deviation (cents)

Let the instantaneous fundamental frequency be $F_0[n]$ and the tonic be F_{ref}. The per-frame pitch deviation in cents is

$$\Delta[n] = 1200 \log_2\left(\frac{F_0[n]}{F_{\text{ref}}}\right). \tag{1}$$

We compute the average absolute pitch deviation $\frac{1}{N}\sum_{n=1}^{N} |\Delta[n]|$, which quantifies pitch stability.

Jitter

With successive pitch periods $T_i = 1/F_0[i]$, the local jitter is

$$J_{\text{local}} = \frac{1}{M-1} \sum_{i=1}^{M-1} \frac{|T_{i+1} - T_i|}{T_i}, \tag{2}$$

reported as a percentage.

 doi:10.20944/preprints202509.0281.v1

Shimmer

With glottal pulse peak amplitudes A_i, the local shimmer is

$$S_{\text{local}} = \frac{1}{M-1} \sum_{i=1}^{M-1} \frac{|A_{i+1} - A_i|}{A_i}. \tag{3}$$

Loudness (LUFS) and RMS energy

ITU-R BS.1770 loudness is approximated by

$$L_{\text{LUFS}} = -0.691 + 10 \log_{10}\left(\frac{1}{N} \sum_{n=1}^{N} w[n]^2 \right), \tag{4}$$

where $w[n]$ is the K-weighted waveform. We also compute

$$E_{\text{RMS}} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x[n]^2}. \tag{5}$$

Tone-to-Noise Ratio (TNR)

Using Parselmouth, we estimate

$$\text{TNR}_{\text{dB}} = 10 \log_{10}\left(\frac{P_{\text{tone}}}{P_{\text{noise}}} \right). \tag{6}$$

MFCCs

From log-Mel energies S_m, the kth MFCC is

$$\text{MFCC}_k = \sum_{m=1}^{M} \log(S_m) \cos\left[\frac{\pi k}{M} \left(m - \tfrac{1}{2} \right) \right]. \tag{7}$$

Zero-Crossing Rate (ZCR)

The short-term zero-crossing rate per frame is

$$Z = \frac{1}{N-1} \sum_{n=1}^{N-1} \mathbf{1}\{x[n]x[n+1] < 0\}. \tag{8}$$

Spectral centroid

$$C = \frac{\sum_f f\,|X(f)|}{\sum_f |X(f)|}. \tag{9}$$

Spectral bandwidth

$$B = \sqrt{\frac{\sum_f (f - C)^2 |X(f)|}{\sum_f |X(f)|}}. \tag{10}$$

Spectral flatness

$$F = \frac{\exp\left(\frac{1}{K} \sum_{k=1}^{K} \ln P_k \right)}{\frac{1}{K} \sum_{k=1}^{K} P_k}, \tag{11}$$

where $P_k = |X(f_k)|^2$.

 doi:10.20944/preprints202509.0281.v1

Formants F1–F3

Using Burg LPC we estimate the first three formants per voiced frame and take their means (Hz).

Vibrato extent and rate

For the pitch contour $F_0[n]$, the vibrato extent (in cents) is

$$E_v = 1200 \log_2\left(\frac{\max_n F_0[n]}{\min_n F_0[n]}\right), \tag{12}$$

and the rate is the dominant modulation frequency of $F_0[n]$ (Hz).

2.3. Angle Scaling

After computing raw features, we scale them to angles in $[0, 2\pi]$ to map to the Bloch sphere and construct the quantum circuit. Representative mappings include:

$$\theta_{\text{pitch}} = 2\pi \cdot \frac{1}{1 + e^{-\frac{(d-d_0)}{k}}}, \tag{13}$$

$$\theta_{\text{jitter}} = 2\pi \cdot \frac{1}{1 + e^{-a(J-J_0)}}, \tag{14}$$

$$\theta_{\text{tempo}} = 2\pi \cdot \tanh\left(\frac{T_{\text{rel}}}{r}\right), \tag{15}$$

$$\theta_{\text{shimmer}} = 2\pi \cdot \frac{1}{1 + e^{-b(S-S_0)}}, \tag{16}$$

$$\theta_{\text{LUFS}} = 2\pi \cdot \frac{L - L_{\min}}{L_{\max} - L_{\min}}, \tag{17}$$

$$\theta_{\sigma\text{LUFS}} = 2\pi \cdot \frac{\sigma_{\text{LUFS}}}{\sigma_{\max}}, \tag{18}$$

$$\theta_{\text{MFCC}} = 2\pi \cdot \frac{\ln(1+M)}{\ln(1+M_{\max})}, \tag{19}$$

$$\theta_{\text{ZCR}} = 2\pi \cdot \frac{\min(ZCR, 0.2)}{0.2}, \tag{20}$$

$$\theta_{\text{TNR}} = 2\pi \cdot \frac{T_{\max} - \text{TNR}}{T_{\max} - T_{\min}}. \tag{21}$$

Here, $T_{\max}$ and $T_{\min}$ are the empirical bounds for TNR (dB); analogous symbols are used for other features. Unless noted, scaling parameters ($d_0, k, a, b, S_0, r, \sigma_{\max}, M_{\max}$) are fixed from the training set using robust bounds (5th–95th percentile or empirical min/max as indicated) and remain constant during evaluation.

Table 1. Features, rotation group, and empirical scaling bounds.

Feature	Rotation Group	Min	Max
Average pitch deviation (cents)	R_x (pitch stability)	0	1431.7
Average jitter	R_x (pitch stability)	0	0.3278
Std. dev. tempo (BPM)	R_x (rhythm)	30	180
Average shimmer	R_y (dynamics)	0	1.1735
Mean LUFS energy (dB)	R_y (dynamics)	-60	-10
Std. dev. LUFS energy	R_y (expression)	1	12
Std. dev. MFCC (timbre)	R_z (timbre)	0	0.25
Zero-crossing rate	R_z (clarity)	0.01	0.12
Mean tone-to-noise ratio (TNR, dB)	R_z (clarity)	5	30

 doi:10.20944/preprints202509.0281.v1

2.4. Quantum Circuit Architecture

We construct a 9-qubit circuit in Qiskit. All qubits are initialized with a Hadamard layer $H^{\otimes 9}$, then receive rotations based on grouped features: qubits 0–2 receive $R_x(\theta_i)$ (pitch-related angles), 3–5 receive $R_y(\theta_i)$ (dynamics), and 6–8 receive $R_z(\theta_i)$ (timbre). We apply intra-group CNOTs (0→1, 1→2; 3→4, 4→5; 6→7, 7→8) and cross-group CNOTs (2→3, 1→4, 0→6, 5→7), then measure all qubits with 8192 shots on the *Qiskit simulator* to obtain measurement probabilities.

2.5. Hybrid Neural Network

We combine an Audio Spectrogram Transformer (AST) [7] fine-tuned on Mel-spectrograms with parallel MLPs over classical features and quantum-derived features. The concatenated embeddings are fed to a fully-connected head that predicts a categorical grade (2–5).

2.6. Comparison Metrics and Environment

For two recordings with quantum probability distributions P and Q, the Jensen–Shannon divergence is

$$D_{\mathrm{JS}}(P\|Q) = \tfrac{1}{2}D_{\mathrm{KL}}(P\|M) + \tfrac{1}{2}D_{\mathrm{KL}}(Q\|M), \quad M = \tfrac{1}{2}(P+Q). \tag{22}$$

where $D_{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. We also compute the Euclidean distance between scaled-angle vectors $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$:

$$d = \sqrt{\sum_{i=1}^{n}\left(\theta_i^{(1)} - \theta_i^{(2)}\right)^2}. \tag{23}$$

All code is written in Python 3.12 using Librosa, Parselmouth, Qiskit, PennyLane and PyTorch on macOS.

2.7. Ethics and Reproducibility

All participants provided informed consent for recording and research use of their vocal audio. Data are anonymized; no identifying metadata are released.

Audio is mono at 22.05 kHz. Feature set: F_0 deviation (cents), jitter, shimmer, LUFS/RMS, TNR, MFCCs (first 13, μ/σ), ZCR, spectral centroid/bandwidth/flatness, formants, and vibrato extent/rate. The quantum circuit configuration and training code are available on request; a companion repository with scripts and hyperparameters will be posted after this preprint is announced.

3. Results and Discussion

The hybrid classical–quantum framework processed each vocal recording in approximately one minute on average (*Qiskit simulator*). We do not claim hardware speedups; real-time performance on quantum hardware is outside our scope. Figure 1 shows the classical metrics captured for each recording and used as inputs to the combined model.

 doi:10.20944/preprints202509.0281.v1

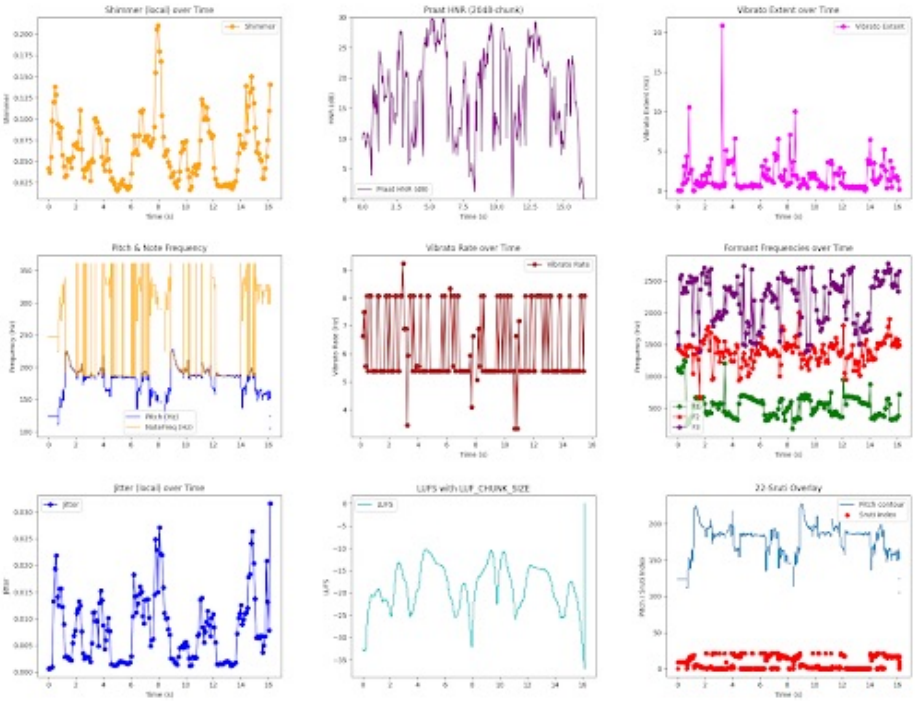

Figure 1. Classical metrics used as inputs to the combined classical–quantum model.

Quantum measurement distributions were calculated for each circuit using 8192 shots (Figure 2). Most recordings did not display a single spike but showed a clear shape, indicating intertwined features.

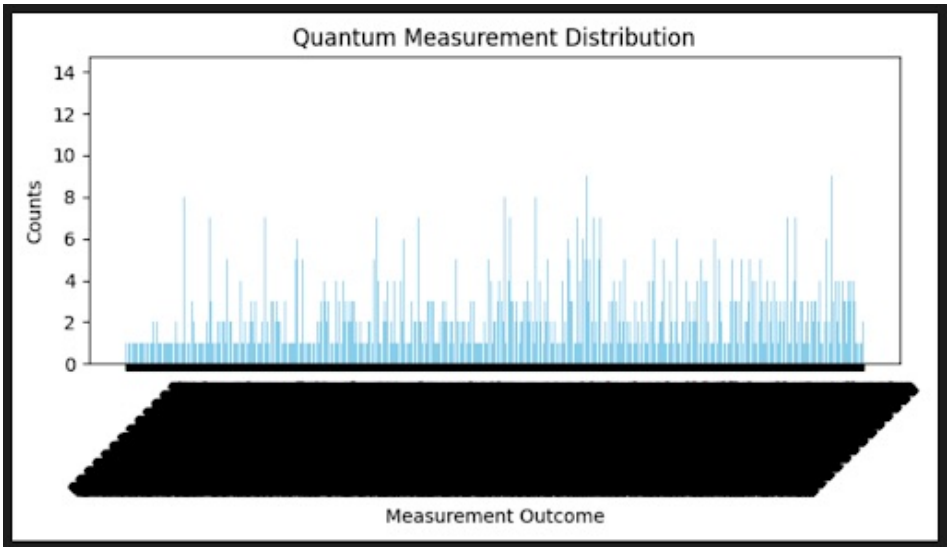

Figure 2. Quantum measurement distribution over bitstrings.

3.1. Improvement over Classical Methods

The baseline included classical features only (pitch dev., jitter/shimmer, MFCC stats, TNR, LUFS, ZCR, formants) with an MLP; the hybrid model adds AST embeddings + quantum measurement probabilities. We used an 80/20 stratified split with raga/label balance and no speaker leakage. We report *agreement*, defined as the percentage of samples whose predicted label exactly matches the expert-provided grade. We cast 2–5 grading as a four-class classification task and report agreement with expert graders. Compared to the classical-only baseline, the hybrid model achieved a +12.86-point absolute improvement in *agreement with expert graders*, reaching 74.29 % versus 61.43 % for the classical system (Figure 3). Given n=168, estimates have non-trivial variance; we report point estimates here.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2025 doi:10.20944/preprints202509.0281.v1


```
Connected to database quantummusic. (fetch-only)
Database connection closed.
Inference for Record ID=2150:
  True raga=Bhairav, True quality=4
  Predicted raga=Bhairav, predicted quality=4
```

```
=== Quality-Only: Quantum vs. No-Quantum ===
Quality Accuracy (Quantum):    0.7429
Quality Accuracy (No-Quantum): 0.6143
Difference in Accuracy = 12.86 % points
```

Figure 3. Quality-only comparison: quantum-enhanced vs. classical baseline.

3.2. Student vs. Master Comparison

Divergence measures—Jensen–Shannon and Euclidean—computed on quantum-encoded feature vectors *qualitatively aligned* with perceptual discrepancies between student and teacher recordings. Analysis showed pitch deviations of 15–25 cents from master tracks (ideal <10 cents), LUFS differences up to 3 dB, and TNR values of 12 dB to 18 dB for students vs. > 20 dB for teachers (Figure 5). Training curves and confusion matrices for the AST model are in Figure 4.

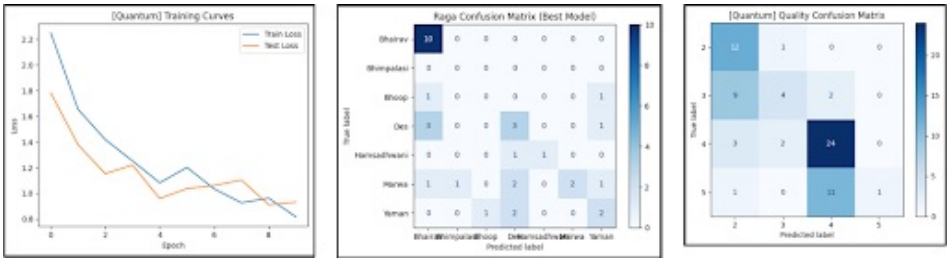

Figure 4. Hybrid AST model training curves and confusion matrices.

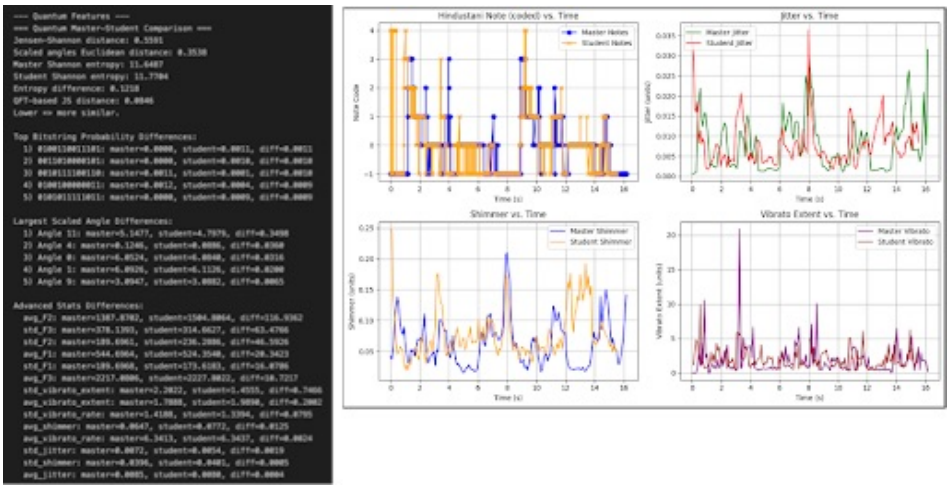

Figure 5. Student vs. master comparison across selected features.

4. Conclusion

We presented *QuantumMelody*, a hybrid method that encodes grouped vocal features in a compact quantum circuit and fuses the circuit's measurement probabilities with spectrogram-transformer embeddings to estimate a 2–5 grade and surface technique-level feedback. The circuit uses nine qubits

 doi:10.20944/preprints202509.0281.v1

with R_x, R_y, and R_z encodings aligned respectively with pitch stability, dynamics, and timbre, with intra- and inter-group entanglement to model cross-domain interactions.

On 168 labeled 20 s excerpts, the hybrid model attains 74.29 % agreement with expert graders, a +12.86-point improvement over a classical-features baseline. All quantum results are produced *on a laptop-class Qiskit simulator*; we do not claim hardware speedups, and behavior on current NISQ devices may differ. Given the modest, single-domain dataset, these findings should be interpreted as a feasibility demonstration within applied audio signal processing.

Overall, the approach provides objective, interpretable indicators for vocal technique without relying on quantum hardware performance claims. This work is most appropriately read within applied audio signal processing.

References

1. Ghisingh, S.; Sharma, S.; Mittal, V.K. Acoustic analysis of Indian classical music using signal processing methods. In Proceedings of the Proc. IEEE Region 10 Conference (TENCON), 2017. https://doi.org/10.1109/TENCON.2017.8228104.
2. Liu, A.Y.; Wallmark, Z. Identifying Peking Opera Roles Through Vocal Timbre: An Acoustical and Conceptual Comparison Between Dan and Laosheng. *Music & Science* **2024**, *7*. https://doi.org/10.1177/20592043241278450.
3. Hashem, A.; Arif, M.; Alghamdi, M. Speech emotion recognition approaches: A systematic review. *Speech Communication* **2023**, *154*, 102974. https://doi.org/10.1016/j.specom.2023.102974.
4. Kashani, S.; Alqasemi, M.; Hammond, J. A quantum Fourier transform (QFT) based note detection algorithm. arXiv preprint, 2022. https://doi.org/10.48550/arXiv.2204.11775.
5. Miranda, E.R.; Yeung, R.; Pearson, A.; Meichanetzidis, K.; Coecke, B. A quantum natural language processing approach to musical intelligence. arXiv preprint, 2021. https://doi.org/10.48550/arXiv.2111.06741.
6. Gündüz, G. Entropy, Energy, and Instability in Music. *Physica A: Statistical Mechanics and its Applications* **2023**, *609*, 128365. https://doi.org/10.1016/j.physa.2022.128365.
7. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the Proceedings of Interspeech 2021. ISCA, 2021.