AI-Assisted Vocal Performance Assessment Through Style Recognition and Fine-Grained Acoustic Feature Analysis

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

Ekaraj Cherian
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
contactekaraj@buffalo.edu

Pierina Tagliafierro
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
pierina.tagliafierro542@binghamton.edu

Abstract

Vocal performance assessment is central to music pedagogy, auditioning, and production, yet automated systems often reduce evaluation to pitch and rhythmic accuracy while neglecting style, timbre, and expressive technique. This reduction creates a mismatch between machine scores and expert judgment, especially when legitimate stylistic choices are interpreted as errors. The manuscript develops a methodological framework for AI-assisted vocal performance assessment that couples style recognition with fine-grained acoustic feature analysis. The framework extracts frame-level descriptors of pitch, loudness, spectral shape, harmonicity, phonation, and onset behavior; aggregates them into performance-level summaries; infers a probabilistic style posterior; and compares observed descriptors with style-conditioned reference expectations. These deviations are mapped to interpretable rubric dimensions through monotone, parameterized functions, and the resulting dimension scores are fused with an uncertainty-aware decision rule. A central insight is that style should not be treated as a post hoc label but as a conditioning variable that changes the interpretation of acoustic deviations. A second insight is that reliability reporting must accompany scoring because style ambiguity and descriptor instability can otherwise produce overconfident judgments. The framework specifies calibration, regularization, and boundary conditions without assuming the availability of a particular corpus or benchmark. Its significance lies in providing an auditable route from acoustic evidence to feedback, which can support transparent assessment, targeted vocal coaching, and more defensible human-machine collaboration in performance evaluation.

Keywords

vocal performance assessment; style recognition; fine-grained acoustic features; uncertainty calibration; interpretable machine learning; singing voice analysis

1. Introduction

Vocal performance assessment combines perceptual judgment with technical evidence. Expert listeners attend to intonation, rhythmic placement, timbral balance, resonance, breath management, articulation, and stylistic appropriateness, and they integrate these dimensions into a holistic judgment. The scientific study of the singing voice has long shown that such judgments depend on measurable phonatory and resonatory behavior rather than on pitch alone [1]. Acoustic parameter sets developed for voice research and affective computing provide compact descriptions of spectral, prosodic, and voice-quality variation [2], while open-source audio analysis libraries make reproducible extraction practical [3]. Melody extraction research further demonstrates that pitch trajectories can be estimated and evaluated in realistic musical signals [4], and timbre perception research shows that spectral and temporal cues shape the perceived identity and quality of sounds [5]. Together, these lines of work support fine-grained acoustic analysis, but they do not by themselves specify how such evidence should be interpreted in style-sensitive performance assessment.

The central problem is that the same acoustic deviation can carry different meanings across vocal styles. A delayed onset, a narrowed vibrato, a sharp spectral centroid, or aperiodic phonation may be an error in one style and an intentional expressive device in another. Many automated assessment systems avoid this ambiguity by using fixed, style-agnostic targets, which rewards conformity to a single vocal template and can penalize legitimate stylistic variation. Other systems emphasize classification accuracy for style labels but do not connect style recognition to rubric-level assessment. The resulting scores are difficult to audit: a learner may be told that a dimension is weak without being told which acoustic evidence, under which stylistic expectation, produced that judgment.

This manuscript develops a methodological framework that addresses that gap. The framework treats style recognition as an intermediate inferential layer rather than as an end task. Fine-grained acoustic descriptors are extracted and aggregated; a probabilistic style posterior is estimated; style-conditioned reference distributions convert descriptors into standardized deviations; and monotone rubric functions map those deviations into interpretable dimension scores. Uncertainty from style ambiguity and rubric dispersion is propagated into a conservative decision rule. The contribution is not a claim of superior empirical accuracy but a technically specific architecture and analytical account of how style conditioning, acoustic granularity, and uncertainty reporting can be combined coherently. The framework is appropriate because it preserves the interpretive role of style while keeping the evidential path from signal to score inspectable.

2. Related Work and Problem Formulation

2.1 Style, synthesis, and acoustic evidence

Recent singing voice synthesis has advanced zero-shot style transfer and multi-level style control, showing that style is not a single global attribute but a structured set of controllable factors [6]. This development is relevant to assessment because it implies that style identity and local expression should be represented separately. A system that collapses style into one label risks confusing broad genre conventions with moment-to-moment expressive decisions. Pitch estimation methods such as CREPE provide robust frame-level fundamental frequency representations that can support fine-grained melodic analysis [7]. Voice production research explains how respiratory, laryngeal, and resonant mechanisms jointly shape acoustic output [8], which motivates descriptors that go beyond pitch and duration. Performance assessment

research has also emphasized the need to examine rater reliability, construct validity, and the multidimensional nature of musical judgment [9]. Expressive communication depends on cue utilization and listener inference [10], and auditory expectation shapes how deviations are perceived within a musical context [11]. Neural parametric singing synthesis further illustrates how timbre and expression can be modeled as separable but interacting dimensions [12]. These works provide the conceptual basis for style-conditioned assessment, but they leave open how to translate style-sensitive expectations into auditable rubric scores.

2.2 Formal problem statement

Let x denote a vocal performance recording, represented as a sequence of audio frames. Let S be a finite set of style categories relevant to the assessment context, and let J be the number of rubric dimensions, such as intonation, rhythmic precision, timbral suitability, phonation stability, and expressive control. The assessment problem is to produce, for each dimension j, a score and an uncertainty estimate that are conditioned on the inferred style distribution rather than on a single fixed template. The system must also provide sufficient acoustic evidence to explain why a dimension received its score. This formulation excludes purely holistic scoring because holistic judgments cannot be traced back to feature-level evidence. It also excludes style classification without assessment because a style label alone does not tell a singer which technical or expressive choices require attention.

3. Methodological Framework

3.1 Acoustic representation and fine-grained descriptors

The first stage converts the waveform into a frame-level representation that preserves both periodic and aperiodic information. For frame t, a compact descriptor vector is defined as

$$\varphi_t = [\log f0_t, \Delta\log f0_t, E_t, \Delta E_t, C_t, H_t, J_t, O_t]^T$$

where f0_t is the fundamental frequency estimate, Δlog f0_t is its first difference, E_t is frame energy or loudness, ΔE_t is the energy difference, C_t is a spectral centroid or related spectral balance measure, H_t is a harmonic-to-noise or harmonicity indicator, J_t is a local perturbation measure associated with phonation stability, and O_t is an onset-strength or temporal-flux descriptor. The notation ^T denotes transposition. The vector is deliberately heterogeneous: pitch and timing descriptors capture melodic and rhythmic control, energy descriptors capture dynamic shaping, spectral descriptors capture resonance and timbre, and harmonicity and perturbation descriptors capture voice quality. This choice follows voice-science evidence that singing quality arises from coupled phonatory and resonant mechanisms [8] and timbre-perception evidence that spectral and temporal cues jointly shape sound identity [5]. Frame-level values are not scored directly because local fluctuations are often normal; instead, they are aggregated into segment-level and performance-level summaries such as medians, interquartile ranges, slope, and proportion of time outside style-conditioned tolerance bands. Aggregation reduces sensitivity to isolated artifacts while retaining information about stability and control.

3.2 Style recognition and style-conditioned expectations

Style recognition is formulated as a probabilistic inference over the style set. Given a performance-level embedding z obtained by pooling frame descriptors over voiced segments, the style posterior is

$$P(s|x) = exp(g_s(z)/\tau) / \sum_\{s' \in S\} exp(g_\{s'\}(z)/\tau)$$

where g_s(z) is the logit assigned to style s by a classifier or similarity model, τ is a temperature parameter controlling posterior sharpness, and S is the style set. The softmax form provides a normalized distribution rather than a hard label, which is important because real performances may blend styles or occupy an ambiguous position between categories. The embedding z may include pooled pitch statistics, rhythmic density, spectral balance, phonation stability, and segment-level expressive variability. The classifier can be replaced by a distance-based or prototype-based model, but the output must remain a calibrated distribution if downstream uncertainty is to be meaningful. This design aligns with the separation of style identity and local expression suggested by multi-level style control in singing synthesis [6] and by parametric modeling of timbre and expression [12].

For each style s and descriptor k, reference central tendency m_{s,k} and dispersion σ_{s,k} are defined for the assessment context. A standardized deviation is then

$$d_\{s,k\}(x) = (\mu_k(x) - m_\{s,k\}) / sqrt(\sigma_\{s,k\}\text{^}2 + \varepsilon)$$

where μ_k(x) is the performance-level summary of descriptor k, m_{s,k} is the style-conditioned reference center, σ_{s,k} is the corresponding dispersion, and ε is a small positive stabilizer. The deviation is signed, so the framework can distinguish sharp from flat intonation, excessive from insufficient loudness variation, and bright from dark spectral balance. The style-conditioned reference statistics are not universal constants; they are context parameters that must be justified for the repertoire, singer population, and assessment purpose. This distinction matters because a style-agnostic z-score would treat all deviation from a global mean as deficiency, whereas a style-conditioned z-score asks whether the deviation is atypical relative to the inferred stylistic expectation. The approach does not assume that stylistic norms are prescriptive; it uses them as explicit, inspectable references that can be revised.

3.3 Rubric mapping and fusion

The rubric mapping converts standardized deviations into dimension scores. For dimension j, let K_j be the set of descriptors relevant to that dimension, let w_{j,k} be nonnegative weights with ∑_{k∈K_j} w_{j,k} = 1, let a_j > 0 control the slope, and let b_j specify a tolerance offset. A bounded dimension score under style s is

$$r_j(s,x) = \frac{1}{\left(1 + \exp\left(a_j\left(\sum_{k\in K_j} w_{j,k} d_{s,k}(x)^2 - b_j\right)\right)\right)}$$

The squared deviation treats positive and negative departures symmetrically unless descriptor-specific asymmetry is introduced by separate weights or piecewise terms. The score lies between zero and one. When the weighted squared deviation is small relative to b_j, the exponent is negative and the score is high; when the deviation is large, the score falls toward zero. The parameters a_j and b_j make the rubric explicit: a_j determines how quickly the score declines with deviation, and b_j determines how much deviation is tolerated before the score is substantially penalized. These parameters should be set from assessment guidelines, expert consensus, or ordinal calibration data rather than chosen for superficial score spread. The monotone form is intentional. It preserves the interpretation that, all else equal, larger style-conditioned deviation lowers the dimension score, while still allowing noncompensatory effects through separate dimensions and descriptor weights.

Because style is uncertain, dimension scores are not evaluated under a single hard label. The style-expected dimension score is

$$R_j(x) = \sum_{s \in S} \mathrm{P}(s|x) r_j(s, x)$$

where P(s|x) is the style posterior from the recognition stage and r_j(s,x) is the dimension score under style s. This expectation is a marginalization over style uncertainty. It prevents an ambiguous performance from being judged entirely by an arbitrary top-1 style. The cost is that a high-probability style with a strict rubric can still dominate if the posterior is sharp, which is why calibration and uncertainty reporting are necessary. The resulting vector R(x) = [R_1(x), …, R_J(x)] is the rubric profile before reliability adjustment. It provides dimension-level feedback that can be traced back to the descriptors in K_j and to the style-conditioned deviations d_{s,k}(x).

3.4 Uncertainty, calibration, and decision logic

Uncertainty is not a secondary reporting detail; it changes the meaning of a score. For dimension j, a style-sensitive uncertainty index is defined as

$$U_j(x) = H\big(\mathrm{P}(\cdot|x)\big) + \sum_{s \in S} \mathrm{P}(s|x)\big(r_j(s, x) - R_j(x)\big)^2$$

where H(P(·|x)) is the entropy of the style posterior and the second term is the posterior-weighted dispersion of dimension scores around the style-expected score. The entropy term captures ambiguity about which style-conditioned expectations should apply. The dispersion term captures disagreement among styles about the dimension score itself. If the style posterior is sharp and all plausible styles yield similar dimension scores, U_j(x) is small. If the posterior is diffuse or if plausible styles yield very different rubric outcomes, U_j(x) is large. This decomposition is useful because it separates two failure modes that require different remedies: a diffuse style posterior may call for more discriminative evidence or a blended-style rubric, while high score dispersion under a sharp posterior may call for finer style subcategories or revised reference statistics.

Calibration is required because neural classifiers and regressors often produce probabilities and scores that are not aligned with observed frequencies or expert ratings [14]. Temperature scaling, isotonic regression, or related post hoc methods can be applied to the style posterior, and ordinal calibration can be applied to dimension scores. Predictive uncertainty can also be estimated through deep ensembles or approximate Bayesian methods [13], [15], but these techniques do not remove the need for calibration against assessor judgments. A conservative decision rule can then combine the expected rubric profile with a reliability penalty:

$$A(x) = \sum_\{j=1\}\^\{J\} \alpha_j R_j(x) - \lambda \sum_\{j=1\}\^\{J\} \alpha_j \, sqrt(U_j(x))$$

where α_j are nonnegative dimension weights with ∑_{j=1}^{J} α_j = 1, and λ ≥ 0 controls the strength of the reliability penalty. The term A(x) is not intended to replace the dimension profile; it is a conservative aggregate for ranking, screening, or triage. The square root compresses large uncertainty values so that one unstable dimension does not dominate the aggregate. When λ is set to zero, the rule reduces to a weighted expected score. When λ is positive, performances with ambiguous style evidence are penalized in the aggregate but can still receive informative dimension-level feedback. This separation between profile and

aggregate is essential for pedagogical use, because a learner needs to know both what was weak and how confident the system is in that judgment.

4. Analytical Assessment and Technical Discussion

4.1 Interpretability and trade-offs

The framework's interpretability comes from its intermediate quantities. The style posterior shows which stylistic expectations were activated; the deviations $d_{s,k}(x)$ show which acoustic descriptors departed from those expectations; the weights $w_{j,k}$ show how descriptors were combined within a rubric dimension; and the uncertainty indices show where the evidence was ambiguous. This is more auditable than an end-to-end score because each stage has a defined input, output, and assumption. The trade-off is that the framework requires explicit choices: the style set S, the descriptor set, the aggregation statistics, the reference statistics $m_{s,k}$ and $\sigma_{s,k}$, and the rubric parameters a_j, b_j, and $w_{j,k}$. These choices are not neutral. A coarse style set may force blended performances into inappropriate categories, while an overly fine style set may produce sparse reference statistics and unstable posteriors. A small descriptor set may be robust but miss important vocal qualities, whereas a large descriptor set may capture nuance but increase collinearity and overfitting risk. The design therefore favors hierarchical regularization and dimension-specific descriptor subsets rather than a single large feature vector for all rubric dimensions.

The choice of a monotone rubric function also involves a trade-off. Monotonicity supports explanation and reduces the risk of paradoxical feedback, but it cannot represent every expert judgment. Some expressive devices are beneficial only within a range and become detrimental when excessive; a simple squared deviation cannot capture such nonmonotonic preferences. This limitation can be addressed by descriptor-specific piecewise terms, interaction terms within a dimension, or ordinal models learned from expert ratings, but each addition complicates interpretation. The framework treats these extensions as options rather than defaults. The baseline monotone form is appropriate when the assessment goal is transparent screening and formative feedback; more complex forms are justified when the rubric explicitly defines optimal ranges rather than one-sided deviations.

4.2 Failure modes and boundary conditions

Several failure modes are predictable. First, style misclassification can propagate into every rubric dimension. If the posterior is sharp but wrong, the expected score may be confidently inappropriate. Calibration and entropy-based abstention reduce this risk but do not eliminate it. Second, reference statistics may encode narrow institutional preferences rather than broad stylistic norms. If $m_{s,k}$ and $\sigma_{s,k}$ are estimated from a small or unrepresentative group, the system may penalize legitimate vocal traditions. Third, acoustic descriptors can be confounded by recording conditions. Microphone proximity, room reverberation, compression, and accompaniment masking alter spectral balance, loudness, and harmonicity. These effects can mimic or hide voice-quality differences. A practical system must therefore include quality control, accompaniment separation or masking-aware features, and recording-condition metadata before acoustic deviations are interpreted as performance deficiencies.

Fourth, the uncertainty index in Eq. (6) captures style ambiguity and rubric dispersion but not all sources of uncertainty. It does not directly represent annotator disagreement, model parameter uncertainty, or distribution shift. These sources require separate calibration and validation. Fifth, aggregate scores such as $A(x)$ can obscure dimension-level information if

used alone. For high-stakes assessment, the dimension profile and uncertainty values should be reported together rather than collapsed into a single rank. Sixth, the framework assumes that style categories are meaningful for the assessment context. If a performance deliberately crosses or rejects available style categories, the posterior may be diffuse, and the system should signal that the style-conditioned rubric is not fully applicable rather than forcing a low score. This boundary condition is important for artistic assessment, where innovation may be misread as error by a style-conditioned model.

4.3 Evaluation logic and implementation considerations

Evaluation should test more than classification accuracy. Style recognition can be assessed with macro-averaged precision, recall, and F1 to avoid dominance by frequent classes [17], but assessment quality requires dimension-level agreement with expert raters, calibration error, and uncertainty discrimination. Ordinal agreement measures are appropriate for rubric scores, and calibration plots or proper scoring rules are appropriate for probabilistic style outputs. Ablation studies should separate the contribution of style conditioning, descriptor granularity, and uncertainty adjustment. Error analysis should examine whether failures cluster around blended styles, poor recordings, or particular rubric dimensions. These analyses are necessary because a model can improve aggregate agreement while producing less useful feedback if its errors are concentrated in the dimensions that matter most to learners.

Implementation can use standard audio feature extraction and deep learning components, but the framework does not depend on a particular toolkit. Open-source feature extractors can provide initial spectral, prosodic, and voice-quality descriptors [18], while pitch trackers and onset detectors supply the melodic and temporal components. A multi-task training objective would combine style classification with rubric regression or ordinal classification, with regularization to prevent the acoustic encoder from overfitting to recording artifacts. Optimization can use adaptive gradient methods such as Adam [16], but the choice of optimizer is secondary to data design, calibration, and error analysis. The more consequential implementation decisions are the construction of style-conditioned reference statistics from defensible populations, the documentation of rubric parameter choices, and the separation of formative feedback from summative decisions.

5. Conclusion

This manuscript has developed an AI-assisted vocal performance assessment framework that integrates style recognition with fine-grained acoustic feature analysis. The framework extracts heterogeneous frame-level descriptors, aggregates them into performance-level summaries, infers a probabilistic style posterior, and evaluates acoustic deviations against style-conditioned reference expectations. It maps those deviations to interpretable rubric dimensions through monotone parameterized functions and propagates style ambiguity and rubric dispersion into an uncertainty-aware decision rule. The central argument is that style should condition assessment rather than follow it, because the same acoustic deviation can be an error, an expressive choice, or a neutral stylistic feature depending on context. A second argument is that uncertainty must be reported alongside scores, because style ambiguity and recording confounds can produce confident but inappropriate judgments. The framework's value lies in its auditable evidence path: from acoustic descriptor to style-conditioned deviation, to rubric score, to uncertainty estimate. This path supports formative feedback and human-machine collaboration without requiring the system to replace expert judgment. Future development should focus on defensible style reference populations, calibration against expert

ratings, robustness to recording conditions, and explicit handling of blended or emergent vocal styles.

References

1. Sundberg, J. (1987). The science of the singing voice. Northern Illinois University Press.
2. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190–202.
3. 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). SciPy.
4. Salamon, J., & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6), 1759–1770.
5. Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. Journal of the Acoustical Society of America, 118(1), 471–482.
6. Zhang, Y., Jiang, Z., Li, R., Pan, C., He, J., Huang, R., ... & Zhao, Z. (2024, November). Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 1960-1975).
7. Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). CREPE: A convolutional representation for pitch estimation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 161–165). IEEE.
8. Titze, I. R. (1994). Principles of voice production. Prentice Hall.
9. Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. Music Perception, 21(1), 21–41.
10. Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of Experimental Psychology: Human Perception and Performance, 26(6), 1797–1813.
11. Pearce, M. T., & Wiggins, G. A. (2012). Auditory expectation: The information dynamics of music perception and cognition. Topics in Cognitive Science, 4(4), 625–652.
12. Blaauw, M., & Bonada, J. (2017). A neural parametric singing synthesizer modeling timbre and expression from natural songs. Applied Sciences, 7(12), 1313.
13. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems 30 (pp. 6402–6413). Curran Associates.
14. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330). PMLR.

15. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059). PMLR.

16. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

17. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

18. Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia (pp. 835–838). ACM.