Supplementary Material for Listening Like a Judge: A Music-Aware Framework for Automatic Singing Performance Evaluation

Neelam Saini[1], *Sourav Ghosh* [1]

[1] Samsung R&D Institute Bangalore, India

neelam.saini@samsung.com, sourav.ghosh@samsung.com

1. Modality-Guided Fine-Tuning Loss for Singing ASR

To adapt Whisper for singing voice transcription, we augment the standard sequence-to-sequence objective with modality-aware regularizers that incorporate pitch, duration, monotonic alignment, and onset cues.

1.1. Overall Objective

Let $x_v(t)$ denote the separated vocal waveform, $y = (y_1, \ldots, y_N)$ the ground-truth lyric token sequence, and $\hat{y}$ the predicted token sequence. Let θ denote model parameters.

The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ASR}} + \lambda_d \mathcal{L}_d + \lambda_p \mathcal{L}_p + \lambda_a \mathcal{L}_a + \lambda_o \mathcal{L}_o \tag{1}$$

Here, $\lambda_\star$ are scalar hyperparameters controlling the contribution of each modality-guided regularizer.

1.2. Cross-Entropy Loss

$$\mathcal{L}_{\text{ASR}} = -\sum_{i=1}^{N} \log P_\theta(y_i \mid y_{<i}, x_v) \tag{2}$$

This is the standard autoregressive token-level cross-entropy.

1.3. Pitch Smoothness Loss

Let $F_0(t)$ denote the fundamental frequency contour estimated from $x_v(t)$. Let $B = \{b_1, \ldots, b_K\}$ be predicted token boundary times. Define local pitch variance around boundary b_k over window $\mathcal{W}_k$:

$$\sigma^2_{F_0}(b_k) = \frac{1}{|\mathcal{W}_k|} \sum_{t \in \mathcal{W}_k} \left(F_0(t) - \mu_k\right)^2 \tag{3}$$

$$\mu_k = \frac{1}{|\mathcal{W}_k|} \sum_{t \in \mathcal{W}_k} F_0(t) \tag{4}$$

The pitch loss discourages token boundaries in acoustically smooth regions:

$$\mathcal{L}_p = \sum_{k=1}^{K} \exp\left(-\sigma^2_{F_0}(b_k)\right) \tag{5}$$

Low pitch variance (sustained notes) yields higher penalty.

Table 1: *Qualitative Alignment between LLM-Generated NL Feedback and Expert Judges for an Exemplary Performance. Highlighted phrases indicate semantically aligned observations.*

Aspect	**Excerpt from Human Expert Judge Comments**	**NL feedback generated by `gpt-oss-120b` based on scores predicted by** MusicJudge
Pitch Accuracy	**Minor pitch instability** in higher notes; slight deviation in sustained segments.	Detected **inconsistent pitch control** during high-register phrases; recommends stabilizing sustained notes.
Rhythm Alignment	**Occasional lag behind beat** in second stanza.	Performance shows **mild rhythmic delay** relative to instrumental backing.
Pronunciation	**Unclear articulation** of conjunct consonants.	Some **articulatory imprecision** observed in complex phonetic segments.
Expressiveness	Good emotional projection but **range could improve**.	**Greater dynamic contrast** would enhance impact.

1.4. Duration Stability Loss

Let $\Delta_k = b_k - b_{k-1}$ be predicted token duration. Let $\bar{\Delta}_k$ be expected duration inferred from CTC alignment or proportional lyric timing.

$$\mathcal{L}_d = \sum_{k=1}^{K} \left(\Delta_k - \bar{\Delta}_k\right)^2 \cdot \exp\left(-\sigma^2_{F_0}(b_k)\right) \tag{6}$$

This penalizes unstable token durations especially in sustained-pitch regions.

1.5. Monotonic Alignment Loss

Let b_k denote the predicted boundary time of the k-th token. Since singing ASR should preserve the natural temporal progression of lyrics, predicted boundaries are encouraged to follow a monotonic ordering.

The monotonic alignment loss is defined as:

$$\mathcal{L}_a = \sum_{k=1}^{K-1} \max\left(0,\, b_k - b_{k+1}\right) \tag{7}$$

This penalizes temporal ordering violations and encourages monotonic progression of token boundaries throughout the vocal sequence.

1.6. Onset Alignment Loss

Let $O = \{o_1, \ldots, o_M\}$ denote detected vocal onset times (e.g., from spectral flux peaks). For each predicted boundary b_k, define distance to nearest onset:

$$\Omega(b_k) = \min_m |b_k - o_m| \tag{8}$$

The onset alignment loss is:

$$\mathcal{L}_o = \sum_{k=1}^{K} \Omega(b_k)^2 \tag{9}$$

This encourages boundaries to align with vocal attack structure.

2. Natural Language (NL) Feedback Alignment

Table 1 presents results from an exemplary performance, demonstrating that the key feedback inferred by the LLM (gpt-oss-120b), when conditioned on the block-wise sequence of modular evaluation scores produced by our system, closely reflects the observations provided by expert judges. Our unified scoring framework first computes quantitative scores across multiple singing aspects, including pitch accuracy, rhythm alignment, pronunciation clarity, and expressiveness. These scores are then provided to the LLM, which generates natural language feedback describing the strengths and weaknesses of the performance.

As illustrated in Table 1, the feedback generated by the LLM exhibits strong semantic consistency with expert commentary. For instance, pitch-related score deviations lead the model to highlight pitch instability in higher notes, while rhythm alignment scores reflecting slight temporal offsets result in feedback indicating mild rhythmic delays relative to the instrumental accompaniment. Similarly, pronunciation-related scores guide the model to identify articulatory imprecision in complex phonetic segments, and expressiveness scores capture limitations in dynamic variation.

These observations suggest that the proposed score-conditioned feedback mechanism enables the system to translate quantitative evaluation signals into interpretable, human-like feedback. Such alignment indicates that the generated explanations remain grounded in the underlying evaluation scores while maintaining semantic similarity to expert assessments.

3. Further Discussion on Related Work

Early research on automatic singing assessment primarily relied on handcrafted acoustic descriptors and shallow learning frameworks. Gupta *et al.* [1] explore automatic singing quality evaluation using pitch stability, vibrato, and timbral features, and propose a reference-free framework that predicts perceptual quality directly from acoustic cues. Subsequent neural approaches incorporate temporal modeling, such as the multi-branch MB-Net architecture [2] and the BiGRU–CapsNet framework in [3], which capture hierarchical temporal dependencies in vocal signals. These works establish the feasibility of automatic singing evaluation, but largely operate at the acoustic level without tightly coupling musical structure and linguistic content.

A parallel line of work emphasizes pitch, tonality, and rhythmic structure while largely omitting explicit lyric modeling. Representation learning approaches such as MARBLE [4] and HCLAS [5] focus on structured and hierarchical audio (or audio–text) embeddings, capturing rich musical attributes. Tonality-aware accompaniment-guided modeling [6] further integrates harmonic context into singing-related tasks, demonstrating the value of music-theoretic conditioning. However, these methods primarily target music understanding or representation learning rather than fine-grained lyric-aligned transcription or evaluation, and do not explicitly address melismatic tokenization challenges in singing ASR.

Conversely, recent work on singing transcription centers on lyric recognition while often treating pitch and rhythmic variation implicitly. SongTrans [7] adapts transformer-based ASR models for lyric transcription under musical variability, and SingMOS-Pro [8] provides a benchmark for robust singing transcription evaluation. While these efforts advance lyric modeling under expressive vocals, they largely rely on generic ASR objectives and do not explicitly regularize token boundaries using pitch continuity or onset cues, leaving melisma-induced fragmentation insufficiently constrained.

Some recent studies move toward integrating multiple modalities for automatic singing assessment. For instance, Narang *et al.* [9] leverages modern self-supervised audio representations to improve robustness in singing evaluation tasks. Although such approaches benefit from richer embeddings, they typically treat acoustic representation learning and linguistic transcription as separate stages, without a unified training objective that jointly constrains pitch behavior, temporal structure, and lexical decoding.

In contrast to prior work, our approach explicitly incorporates singing-specific modalities – pitch contour, duration stability, and onset alignment – directly into the fine-tuning objective of a large-scale ASR backbone. Rather than treating pitch–rhythm cues and lyric modeling as disjoint components, we introduce modality-guided regularization that constrains token boundary formation under melismatic and expressive conditions. This enables segmentation-aware transcription that is both linguistically accurate and musically coherent, bridging the gap between music-informed modeling and lyric-level ASR adaptation. Furthermore, we leverage the semantic boundaries derived from transcription pipeline to aid in the assessment of acoustic features like pitch and rhythmic deviations. To the best of our knowledge, we are the first to explore such a joint modeling for singing quality assessment, leading to a high correlation of objective and qualitative metrics with human experts.

4. References

[1] C. Gupta, H. Li, and Y. Wang, "Automatic evaluation of singing quality without a reference," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, pp. 990–997.

[2] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International*

Conference on Acoustics, Speech and Signal Processing. IEEE, 2021, pp. 391–395.

[3] H. Wu, "Vocal performance evaluation based on bidirectional gated recurrent units and caps net," in *2024 International Conference on Data Science and Network Security (ICDSNS)*, 2024, pp. 1–5.

[4] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, Y. Liu, J. Huang, Z. Tian, B. Deng *et al.*, "Marble: Music audio representation benchmark for universal evaluation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39 626–39 647, 2023.

[5] M. Kang, S. Park, and K. Choi, "Hclas-x: Hierarchical and cascaded lyrics alignment system using multimodal cross-correlation," *arXiv preprint arXiv:2307.04377*, 2023.

[6] P.-C. Hsieh, Y.-L. Shen, N.-S. Tran, and T.-S. Chi, "Tonality-based accompaniment-guided automatic singing evaluation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech Communication Association, 2025, pp. 3085–3089.

[7] S. Wu, J. He, R. Yuan, H. Wei, X. Wei, C. Lin, J. Xu, and J. Lin, "Songtrans: An unified song transcription and alignment method for lyrics and notes," *arXiv preprint arXiv:2409.14619*, 2024.

[8] Y. Tang, L. Liu, W. Feng, Y. Zhao, J. Han, Y. Yu, J. Shi, and Q. Jin, "Singmos-pro: An comprehensive benchmark for singing quality assessment," *arXiv preprint arXiv:2510.01812*, 2025.

[9] J. Narang, N. C. Tamer, V. De La Vega, and X. Serra, "Automatic estimation of singing voice musical dynamics," *arXiv preprint arXiv:2410.20540*, 2024.