Parameter-Efficient LoRA Adaptation for Lyric Intelligibility Prediction in the Cadenza 2026 Challenge

Qiao Jiayi
National University of Singapore
Singapore
E1583084@u.nus.edu

Reiner Anggriawan Jasin
National University of Singapore
Singapore
E1503344@u.nus.edu

Ram Gopalakrishnan
National University of Singapore
Singapore
E1503341@u.nus.edu

Ye Guoquan
National University of Singapore
Singapore
E0324531@u.nus.edu

Abstract

This paper presents the submission of Team T094 to the ICASSP 2026 Cadenza Challenge on lyric intelligibility prediction under simulated hearing loss [2]. We adopt a parameter-efficient strategy based on Low-Rank Adaptation (LoRA) [3] to fine-tune a pretrained encoder–decoder speech model while updating fewer than 1% of its parameters. In accordance with challenge requirements, our system uses only the reference text and the hearing-loss–processed mixture audio, without relying on the unprocessed clean signal or any perceptual metrics such as STOI or PESQ. The official evaluation reports a validation RMSE of 32.75 (correlation 0.58) and an evaluation RMSE of 33.25 (correlation 0.56). These outcomes highlight that LoRA provides an effective and computationally lightweight mechanism for adapting pretrained models to the singing-intelligibility domain and can preserve stable generalization under limited task-specific data.

1 Introduction

Predicting lyric intelligibility in musical mixtures is challenging due to several acoustic phenomena characteristic of singing. Compared with conversational speech, singing contains wide pitch excursions, long sustained vowels, artistic articulation choices, and expressive timing variations. These traits alter the spectral envelope and temporal structure that speech models typically rely on. Furthermore, musical accompaniment introduces dense harmonic and percussive masking that can obscure consonant transitions. Under simulated hearing loss, high-frequency cues essential for distinguishing fricatives or plosives are attenuated even further, making intelligibility estimation substantially more difficult. The Cadenza Lyric Intelligibility Prediction (CLIP) dataset [1] and the ICASSP 2026 Cadenza Challenge [2] provide a controlled benchmark for studying these effects.

Large pretrained speech models encode rich linguistic structure, but the computational cost of fully fine-tuning them often exceeds the constraints typical of singing-focused applications, especially when only limited in-domain data are available. Additionally, the CLIP dataset offers a relatively small number of task-specific examples compared with generic speech corpora, so fully updating hundreds of millions of parameters may lead to overfitting.

Low-Rank Adaptation (LoRA) [3] provides an appealing alternative: by injecting compact low-rank matrices into the backbone, it enables efficient domain specialization while maintaining most layers in a frozen state. This design aligns well with the challenge objective of building a simple, transparent, and resource-efficient system.

Our approach intentionally restricts itself to a single LoRA-enhanced model. No auxiliary models, handcrafted features, or perceptual intelligibility metrics are used, allowing us to isolate the benefits of parameter-efficient tuning within a controlled setting.

2 Methodology

Our system adapts a pretrained encoder–decoder transformer using LoRA [3]. In accordance with challenge rules, the system uses only the **reference lyric text** and the **hearing-loss–processed mixture audio**. The unprocessed clean signal is not used, and no perceptual metrics (e.g., STOI, ESTOI, PESQ) are incorporated at any stage. All results reported in this paper are computed using the official CLIP metadata v1.2 labels released by the organizers [1].

2.1 Dataset and Preprocessing

We use the officially released CLIP labels (metadata v1.2) for both validation and evaluation [1, 2]. The hearing-loss–processed mixture audio is downmixed to mono and resampled to 16 kHz. We compute 80-bin log-Mel spectrograms using the feature extractor associated with the pretrained backbone. Text transcripts are lowercased and stripped of punctuation. Dynamic padding is applied to allow efficient batching of variable-length inputs. These steps follow standard practice for adapting pretrained encoder–decoder speech models to new downstream tasks while minimizing distribution shift.

2.2 Low-Rank Adaptation (LoRA)

LoRA [3] modifies the weight update of a pretrained matrix $W \in \mathbb{R}^{d\times k}$ by introducing a low-rank decomposition:

$$\Delta W = BA^{\top}, \quad A \in \mathbb{R}^{k\times r},\ B \in \mathbb{R}^{d\times r},$$

with $r \ll \min(d, k)$. Only the parameters of A and B are trained, while W remains frozen. This design provides a lightweight mechanism for domain adaptation without modifying the majority of the model.

Conceptually, LoRA restricts task-specific updates to a small subspace while retaining the general-purpose knowledge encoded in the original weights. This is particularly suitable for lyric intelligibility prediction: the model must adapt to stylistic aspects of singing and hearing-loss processing while preserving the broad linguistic and acoustic priors acquired during large-scale pretraining.

2.3 Model Configuration

LoRA is applied to the query and value projection matrices across all encoder and decoder attention layers. These projections play a central role in shaping how the model weighs contextual frames and how strongly it attends to internal dependencies between lyric tokens and acoustic cues.

Our final configuration uses:

- Rank $r = 16$ and scaling factor $\alpha = 32$,
- LoRA dropout of 0.05 to encourage robustness,
- A total of 1.77M trainable parameters ($< 1\%$ of the model).

This parameter budget allows the model to incorporate meaningful domain-specific structure while maintaining stability across training runs. Higher ranks increase model capacity but exhibit mild overfitting, whereas lower ranks tend to underrepresent the acoustic variability of singing.

2.4 Training

We fine-tune the model using cross-entropy loss over lyric tokens, with padding masked to prevent loss distortion. The Adam optimizer with a linear warmup schedule ensures smooth early-stage convergence. Because only LoRA parameters are trainable [3], each optimization step is computationally inexpensive and benefits from reduced variance in gradient updates.

We further monitor validation loss to trigger early stopping, preventing over-adaptation to the limited dataset. No external text, audio data, or augmentation is used; all experiments rely strictly on the official CLIP dataset, preserving the controlled evaluation setting expected in the challenge [1].

3 Official Results

Table 1 reports the official challenge results released by the organizers [2]. Our submission (Team T094) performs competitively relative to the provided baselines.

Table 1: Official Cadenza 2026 Challenge results.

System	Val RMSE	Val Corr	Eval RMSE	Eval Corr
Baseline STOI	36.11	0.14	34.89	0.21
Baseline Whisper	29.32	0.59	29.08	0.58
T094 (ours)	**32.75**	**0.58**	**33.25**	**0.56**

Our system matches the baseline Whisper correlation on validation and remains close on evaluation, demonstrating that LoRA preserves modeling power despite updating only a small subset of parameters. The competitive performance suggests that the intelligibility regression task benefits from the linguistic structure encoded in the pretrained model and that lightweight adaptation can compensate for domain mismatches introduced by singing and hearing-loss processing.

4 Ablation Study

We investigate the influence of LoRA rank on training behavior, following the parameter-efficient tuning perspective in [3]. Smaller ranks ($r = 4$ and 8) provide insufficient capacity to model the expanded acoustic variability present in singing, resulting in underfitting. In contrast, larger ranks such as $r = 32$ increase the number of tunable parameters and introduce mild overfitting. The intermediate configuration of $r = 16$ offers a balanced trade-off between flexibility and stability.

Table 2: Effect of LoRA rank.

Rank	Trainable Params	Observation
4	0.44M	Underfitting
8	0.88M	Improved stability
16	**1.77M**	**Best performance**
32	3.54M	Mild overfitting

5 Conclusion

We presented a parameter-efficient lyric intelligibility prediction system using LoRA-enhanced adaptation of a pretrained encoder–decoder speech model. Updating less than 1% of parameters allows the model to specialize to the singing-intelligibility domain while retaining the robustness of the pretrained backbone. The competitive results obtained in the Cadenza 2026 evaluation [2] highlight the utility of LoRA [3] for music-related intelligibility modeling and suggest that lightweight adaptation techniques are viable solutions when domain-specific data are limited [1].

References

[1] G. Roa-Dabike et al. The Cadenza Lyric Intelligibility Prediction (CLIP) Dataset. *Data in Brief*, 2025.

[2] G. Roa-Dabike et al. Overview of the ICASSP 2026 Cadenza Challenge: Predicting Lyric Intelligibility. In *Proc. IEEE ICASSP*, 2026. To appear.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. ICLR*, 2022.