

# FLIP: A FROZEN LYRIC INTELLIGIBILITY PREDICTOR FOR MUSIC AUDIO UNDERSTANDING

Muhammad Musaab Ul Haq, Huma Ameer, and Mehwish Fatima

School of Electrical Engineering and Computer Science (SEECS),  
National University of Sciences and Technology (NUST)  
Islamabad, Pakistan

{mhaq.bscs24seecs, hameer.msds20seecs, mehwish.fatima}@seecs.edu.pk

## ABSTRACT

The goal of the ICASSP 2026 Cadenza CLIP1 Challenge is to predict “Lyric Intelligibility”, i.e., whether listeners correctly perceive sung lyrics. This paper describes our pipeline from exploratory data analysis (EDA) to FLIP, a Frozen Lyric Intelligibility Predictor. Our analysis shows that the dataset is strongly bimodal and that linguistic features (e.g., word length, word frequency) exhibit negligible correlation with intelligibility. In contrast, acoustic brightness (spectral centroid) shows a meaningful positive correlation. We propose FLIP, which uses a frozen OpenAI Whisper *large-v3* encoder as an acoustic feature extractor, combined with a learnable embedding layer for hearing-loss profiles and a custom regression head. FLIP achieves an RMSE of 0.2638 and an NCC of 0.666 while outperforming the provided baselines.

**Index Terms**— Lyric intelligibility, Whisper, audio embeddings, frozen encoder, EDA

## 1. INTRODUCTION

Understanding lyrics is fundamental to music enjoyment, yet listeners with hearing loss often struggle to perceive sung words clearly and effortlessly [3, 4]. According to the World Health Organization, approximately 1.5 billion people worldwide currently live with some degree of hearing loss, and this number continues to grow due to aging populations and increased exposure to loud sound<sup>1</sup>. While speech intelligibility prediction is extensively studied with metrics such as Short-Time Objective Intelligibility (STOI) and Hearing Aid Speech Perception Index (HASPI), virtually no equivalent work targets sung lyrics.

The ICASSP 2026 Cadenza Lyric Intelligibility Prediction Challenge (CLIP1) [5] addresses this gap by proposing a task to predict the fraction of words a listener correctly identifies from short musical excerpts. The CLIP1 dataset comprises thousands of Western popular music segments from the FMA database [2], paired with listener transcriptions from perceptual experiments. Each audio clip appears in three conditions: unprocessed, or processed through mild or moderate hearing loss simulation using the Cambridge auditory model. The target variable is a continuous “correctness” score ranging from 0 (no words identified) to 1 (all words correct).

This challenge introduces difficulties beyond speech intelligibility. Sung language differs from spoken language in rhythm, intonation, and phonetic realization [3]. Studio production techniques,

including reverb, dynamic compression, and vocal layering, further alter acoustic characteristics in ways that standard speech metrics do not capture [7]. These factors motivate an approach that leverages acoustic representations rather than text-based ASR outputs.

Our key contributions are as follows:

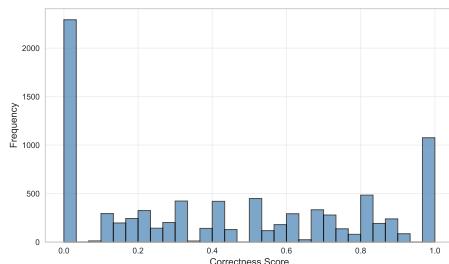
- We conduct an exploratory data analysis showing that acoustic features correlate with intelligibility while simple linguistic features do not.
- We adapt the frozen speech foundation model approach from Cuervo & Marixer [1] to lyric intelligibility, proposing FLIP (Frozen Lyric Intelligibility Predictor) that combines Whisper *large-v3* audio embeddings with learnable hearing-loss embeddings.

## 2. DATASET

We conduct exploratory data analysis (EDA) on the CLIP1 training set [6] to understand factors that influence lyric intelligibility.

### 2.1. Dataset Overview

The training set contains 8,802 audio-response pairs. Each sample includes a short musical excerpt, ground-truth lyrics, the listener’s transcription, and a correctness score. The hearing loss conditions are balanced: No Loss ( $n = 2933$ ), Mild ( $n = 2935$ ), and Moderate ( $n = 2934$ ). Each audio clip is presented in three conditions: unprocessed, or processed through mild or moderate hearing loss simulation using the Cambridge auditory model. The target variable is a continuous “correctness” score ranging from 0 (no words identified) to 1 (all words correct).



**Fig. 1.** Distribution of lyric intelligibility scores showing strong bimodality at 0.0 and 1.0.

Corresponding author: mehwish.fatima@seecs.edu.pk

<sup>1</sup>WHO Link

**Table 1.** Feature correlations with intelligibility score.

Feature	Pearson $r$
Average Word Length	-0.098
Average Word Frequency	-0.020
Spectral Centroid	+0.209

## 2.2. Target Distribution

As shown in Fig. 1, the correctness scores exhibit a strongly bimodal distribution, with prominent peaks at 0.0 and 1.0. This indicates that listeners either understood nearly all words or virtually none, a challenging pattern for regression models trained with Mean Squared Error (MSE), which tend to predict conservative middle values to minimize loss.

## 2.3. Linguistic vs. Acoustic Features

We hypothesize that linguistic complexity affects intelligibility, e.g., longer or rarer words are harder to perceive. However, correlation analysis indicates negligible relationships between these linguistic features and correctness scores.

In contrast, the spectral centroid, a measure of acoustic “brightness”, shows a meaningful positive correlation with correctness ( $r = 0.209$ ). This finding, summarized in Table 1, supports the view that **acoustic signal quality, not linguistic content, drives intelligibility**. We also visualize spectrograms and observe that hearing-loss simulation attenuates high frequencies, which aligns with the observed spectral-centroid correlation.

This raises a critical concern: if we rely on ASR transcripts (e.g., Whisper decoding), a robust ASR model may recover plausible lyrics even from degraded audio because it is trained to be noise-robust. A human listener with hearing loss cannot do that. We therefore design a model that “listens” to signal quality rather than one that “reads” text.

**Table 2.** Summary statistics by hearing loss category.

Condition	Mean Corr.	Std Dev.	Centroid (Hz)
No Loss	0.579	0.339	2,482
Mild	0.395	0.348	1,218
Moderate	0.316	0.334	977

## 2.4. Hearing Loss Effects

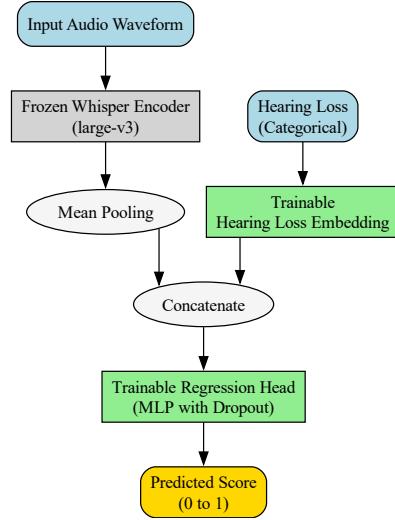
Table 2 presents summary statistics by hearing-loss category. As expected, intelligibility degrades with increasing hearing-loss severity. Notably, the mean spectral centroid drops from 2,482~ Hz (No Loss) to 977~ Hz (Moderate), reflecting the high-frequency attenuation introduced by the hearing-loss simulation.

These findings inform FLIP design: rather than relying on ASR transcription accuracy, we focus on extracting acoustic representations that capture signal degradation patterns.

## 3. FROZEN LYRIC INTELLIGIBILITY PREDICTOR

Fig. 2 shows the FLIP (Frozen Lyric Intelligibility Predictor) pipeline. FLIP consists of three main components: a backbone (Whisper Large V3), a hearing-loss embedding, and a regression head.

**Backbone (Whisper Large V3):** We select *whisper-large-v3* after experimentation with other variants. The encoder weights remain *frozen* to reduce overfitting; training a model of this size on a limited dataset can degrade generalization. We use the encoder as a



**Fig. 2.** FLIP pipeline combining Whisper features, hearing-loss embeddings, and an MLP regression head.

feature extractor. The encoder outputs a sequence of hidden states, and we simplify the approach in [1] by taking the *mean* across the time dimension to produce a single vector representation for each clip.

**Hearing Loss Embedding:** The dataset provides listener-condition metadata (No Loss, Mild, Moderate). Without this signal, the model lacks explicit awareness of the hearing condition. We train a small learnable embedding layer that maps each category to a 32-dimensional vector.

**Regression Head:** We concatenate the audio vector and hearing-loss vector and pass them through a multi-layer perceptron (MLP). The MLP uses Linear layers, GELU activations, Layer-Norm, and Dropout ( $p = 0.5$ ) to regularize.

**Output:** A single sigmoid node constrains the prediction to the  $[0, 1]$  range.

## 4. EXPERIMENTAL SETUP

We train FLIP using MSELoss. We use bfloat16 (mixed precision) and set batch\_size=8 with gradient\_accumulation\_steps=8, yielding an effective batch size of 64 when using *whisper-large-v3*. All experiments run in the PyTorch framework.

### 4.1. Input Specification

FLIP uses only the processed audio signal (Audio 1) and the hearing loss severity metadata: **Processed Signal:** The stereo audio heard during listener tests, which may include hearing loss simulation. **Hearing Loss Metadata:** The categorical severity label (No Loss, Mild, or Moderate).

## 5. RESULTS AND DISCUSSION

FLIP significantly outperforms the organizer-provided baselines, as illustrated in Table 3. The bimodal target distribution identified in our EDA affects regression behavior: FLIP remains conservative,

**Table 3.** Model Performance on the Internal Validation Set.

Model	RMSE	NCC
STOI Baseline	0.36	0.14
Whisper ‘base.en’ Baseline	0.29	0.59
<b>FLIP (our model)</b>	<b>0.28</b>	<b>0.63</b>

rarely predicting values near 0.0 or 1.0, and instead concentrating predictions in the 0.2–0.8 range. This behavior aligns with MSE training on bimodal data, where the model reduces squared error by avoiding extreme predictions. Despite this compression, the high NCC indicates that FLIP correctly ranks clips by intelligibility even when absolute values shift toward the mean.

**Table 4.** Official Challenge Scores (Team T063).

Set	RMSE	Correlation
Validation	0.2837	0.66
Evaluation	0.2881	0.63

### 5.1. Official Challenge Results

Table 4 reports FLIP’s performance on the official challenge test sets. Despite a higher RMSE relative to internal validation (0.2638), FLIP continues to outperform the baseline systems. The increase reflects distributional differences between the development split and the organizers’ held-out evaluation data, indicating a limited and expected domain shift.

**Table 5.** Ablation study results. All experiments use *whisper-small* except where noted. HL = Hearing Loss embedding, L = MLP layers, Dr = Dropout rate.

Group	Variation	HL	L	RMSE	NCC
<i>Pooling Strategy (Dr=0.5)</i>					
	Attention	Used	2	<b>0.2815</b>	<b>0.632</b>
	Max	Used	2	0.2861	0.623
	Mean (Baseline)	Used	2	0.2913	0.598
<i>Hearing Loss Embedding</i>					
	Without HL Embed	Not Used	2	0.2912	0.606
<i>Regression Head Depth</i>					
	Shallow (1 layer)	Used	1	0.2898	0.609
	Deep (3 layers)	Used	3	0.2961	0.591
<i>Regularization</i>					
	Lower Dropout (0.3)	Used	2	0.2902	0.607
<i>Backbone Model</i>					
	Whisper-base	Used	2	0.3101	0.529

### 5.2. Ablation Study

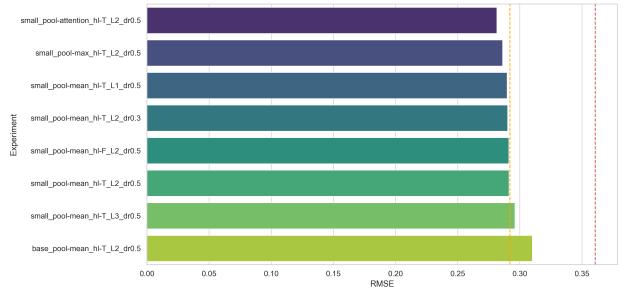
To understand the contribution of each FLIP component, we run 8 ablation experiments. Table 5 presents the configuration matrix and results.

Fig. 3 visualizes the RMSE performance across all configurations. Key findings include:

1. **Pooling strategy** strongly influences performance. Attention pooling achieves the best RMSE (0.2815), a 3.4% improvement over mean pooling.
2. **Backbone capacity** matters. *whisper-small* outperforms *whisper-base* by 6.1% RMSE, suggesting that larger frozen encoders provide richer acoustic representations.

3. **Hearing-loss embeddings** provide marginal benefit (0.2912 vs. 0.2913), suggesting that the encoder already captures degradation patterns implicitly.

4. **Head depth** shows diminishing returns. Two layers perform best; three layers overfit while one layer underfits slightly.
5. **Dropout rate** has minimal effect (0.2902 at 0.3 vs. 0.2913 at 0.5).

**Fig. 3.** RMSE comparison across ablation configurations. Orange dashed line indicates Whisper baseline; red dashed line indicates STOI baseline.

These ablations support our FLIP design: *whisper-large-v3* as the backbone (consistent with the small vs. base trend), mean pooling for simplicity with competitive performance, hearing-loss embeddings for interpretability despite marginal gains, and a 2-layer regression head with dropout 0.5.

## 6. CONCLUSION

We propose FLIP, a regression model for lyric intelligibility that “listens” rather than “reads.” By identifying that acoustic brightness correlates with intelligibility while linguistic features do not, we utilize a frozen *whisper-large-v3* encoder as a feature extractor. Hence, FLIP achieves an RMSE of 0.2638. Future work will explore alternative loss functions and task reformulations to address the conservative prediction bias.

## 7. REFERENCES

- [1] Santiago Cuervo and Ricard Marixer. Speech foundation models on intelligibility prediction for hearing-impaired listeners. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1421–1425. IEEE, 2024.
- [2] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.
- [3] Philip A Fine and Jane Ginsborg. Making myself understood: perceived factors affecting the intelligibility of sung text. *Frontiers in Psychology*, 5:809, 2014.
- [4] Alinka Greasley, Harriet Crook, and Robert Fulford. Music listening and hearing aids: perspectives from audiologists and their patients. *International Journal of Audiology*, 59(9):694–706, 2020.
- [5] Gerardo Roa-Dabike, Jon P. Barker, Trevor J. Cox, Michael A. Akeroyd, Scott Bannister, Bruno Fazenda, Jennifer Firth,

- Simone Graetzer, Alinka Greasley, Rebecca R. Vos, and William M. Whitmer. Overview of the icassp 2026 cadenza challenge: Predicting lyric intelligibility. In *Proc. IEEE ICASSP*, 2026. To appear.
- [6] Gerardo Roa-Dabike, Trevor J. Cox, Jon P. Barker, Bruno M. Fazenda, Simone Graetzer, Rebecca R. Vos, Michael A. Akeroyd, Jennifer Firth, William M. Whitmer, Scott Bannister, and Alinka Greasley. The cadenza lyric intelligibility prediction (clip) dataset. *Data in Brief*, 2025.
- [7] Bidisha Sharma and Ye Wang. Automatic evaluation of song intelligibility using singing adapted stoi and vocal-specific features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:319–331, 2019.