CADENZA CLIP1 Lyric Intelligibility Challenge: Problem and Data Understanding

Team Invictus: Maham Mansoor, Rudaina Aamir, Wamiq ul Islam, Zaina Zia
School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST), Islamabad, Pakistan

Abstract

This study investigates the CADENZA CLIP Lyric Intelligibility Challenge, an official ICASSP 2025 task focused on predicting how understandable sung lyrics are under varying acoustic and hearing conditions. Through comprehensive exploratory data analysis (EDA), this work examines the dataset's structure, modalities, and relationships to develop an informed foundation for intelligibility modeling. Results reveal strong polarization in intelligibility scores and notable effects of hearing loss on lyric comprehension, offering valuable insights for subsequent feature engineering and model design.

I Introduction & Motivation

This project is based on the CADENZA CLIP Lyric Intelligibility Challenge, an official ICASSP 2025 task focused on predicting how understandable song lyrics are under different acoustic and hearing conditions. Lyric intelligibility refers to how clearly listeners can perceive and recognize words in sung audio. It plays a vital role in audio accessibility, hearing aid optimization, and music production quality. The motivation behind this study is to understand how various factors such as signal clarity, hearing loss, and audio features affect lyric intelligibility. Through detailed exploratory data analysis (EDA), we aim to uncover meaningful relationships in the dataset that can guide future model development. This assignment focuses on problem and data understanding, ensuring a deep grasp of the dataset structure, its modalities, and patterns in the intelligibility scores before any modeling begins.

II Dataset Overview

The dataset provided in the challenge includes 8,802 audio samples, each paired with metadata describing intelligibility scores, word-level correctness, and hearing loss categories. Every sample represents a short sung lyric clip that has been evaluated for how accurately listeners could recognize its words.

Main components:

- Audio modality: .flac files representing sung lyrics.
- Text modality: Metadata fields such as n_words, words_correct, and correctness.
- Hearing condition: Categorical variable with three balanced groups — Mild (2,935), Moderate (2,934), and No Loss (2,933).
- Sampling rate: Uniform at 44.1 kHz for all clips.
- Duration: Ranges from 1.1 to 22.9 seconds, averaging 4.48 seconds.

No missing files were found, ensuring data completeness and reliability. The dataset is well-balanced across hearing loss categories, which reduces potential bias in downstream model training.

III Exploratory Data Analysis Results & Insights

The EDA explored statistical properties, signal characteristics, and relationships across modalities to understand how they relate to lyric intelligibility. Nine visualizations were produced to summarize key findings.

III-A Distribution of Intelligibility Scores

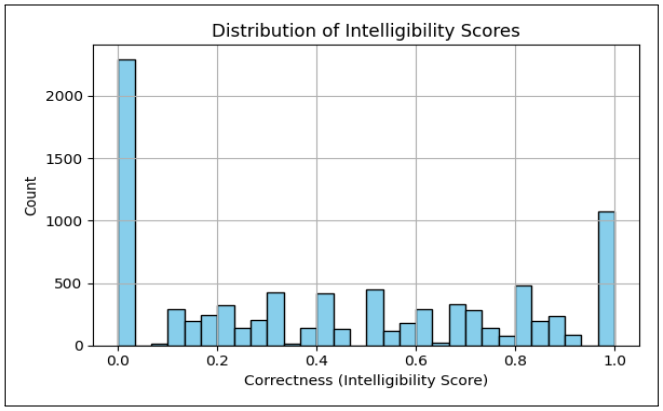

Fig. 1: Histogram of correctness metric.

A histogram of the correctness metric reveals a bimodal distribution, with two prominent peaks near 0.0 and 1.0. This means that many lyric clips are either completely unintelligible or fully intelligible, while fewer samples fall in the intermediate range. **Insight:** The target variable is polarized rather than continuous, suggesting that the dataset contains two dominant types of samples — very clear and very unclear clips. Predictive models must therefore handle both extremes effectively, as standard regression approaches might struggle with this sharp bimodality.

III-B Intelligibility by Hearing Loss

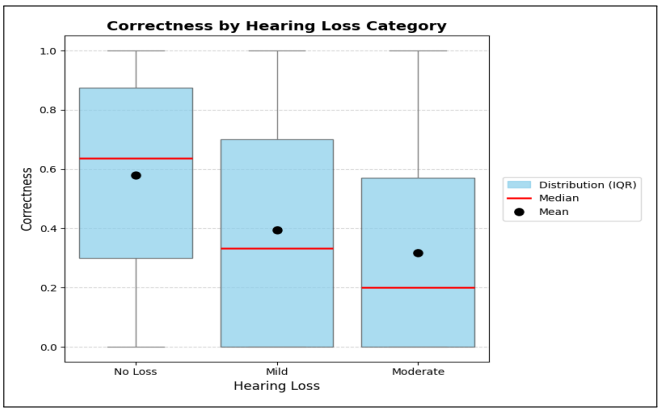

Fig. 2: Boxplot comparing correctness by hearing loss.

A boxplot comparing correctness across Mild, Moderate, and No Loss hearing loss categories reveals clear median differences. Listeners with no hearing loss exhibit the highest median intelligibility (around 0.65), while mild and moderate hearing loss groups show progressively lower medians (0.35 and 0.20, respectively). All three groups display a broad range of scores, indicating that intelligibility varies within each condition. **Insight:** Hearing loss appears to have a negative influence on lyric intelligibility, particularly for moderate impairment. However, the overlapping score distributions suggest that other acoustic and linguistic factors also contribute to intelligibility beyond hearing condition alone.

III-C Distribution of Audio Durations

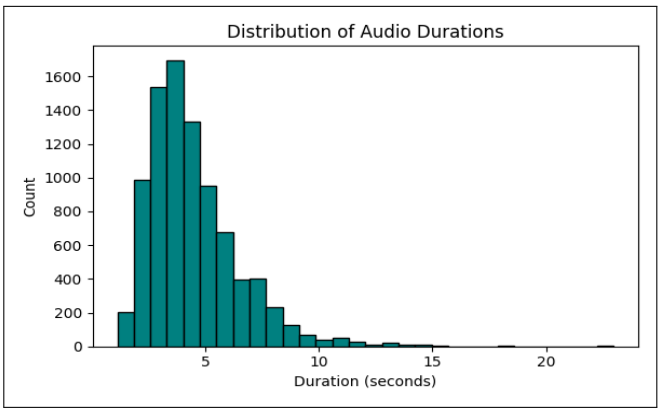

Fig. 3: Histogram of audio durations.

Audio durations follow a right-skewed distribution centered around 4.5 seconds. Most clips are short lyric excerpts, with a few extending beyond 20 seconds. **Insight:** For modeling, input length should be standardized (e.g., trimmed or padded) to ensure consistent feature extraction and reduce bias from long clips.

III-D Duration vs. Sample Rate

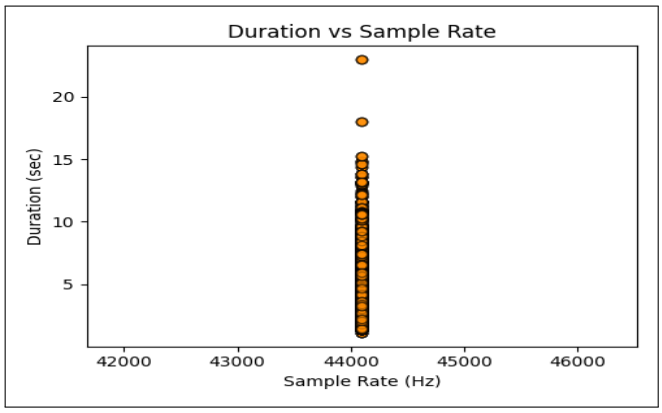

Fig. 4: Scatter plot showing duration vs. sample rate.

A scatter plot confirms that all audio files share the same sample rate (44.1 kHz), as all points align horizontally. **Insight:** The uniform sampling rate ensures acoustic consistency and eliminates the need for resampling. The variation in duration therefore reflects content diversity rather than recording inconsistencies, confirming technical uniformity in data acquisition.

III-E Frame Count Differences (Processed vs. Unprocessed Signals)

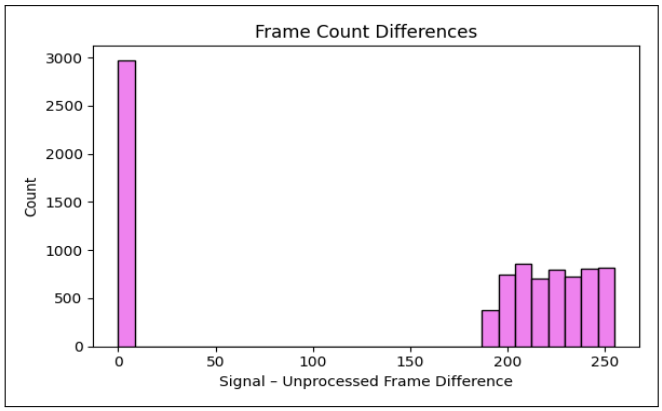

Fig. 5: Comparison of frame counts between processed and unprocessed signals.

The histogram comparing frame counts between processed and unprocessed audio clips shows that most clips remained unchanged after preprocessing, indicated by a strong peak around zero difference. However, a smaller subset of clips displays noticeable frame count differences, suggesting that certain samples experienced alterations during cleaning or noise reduction. **Insight:** Preprocessing largely preserved the linguistic and temporal structure of the audio data, though some clips were affected more significantly. This observation highlights the importance of verifying preprocessing consistency to ensure that downstream intelligibility patterns are not distorted for specific subsets of the data.

III-F Waveform Example

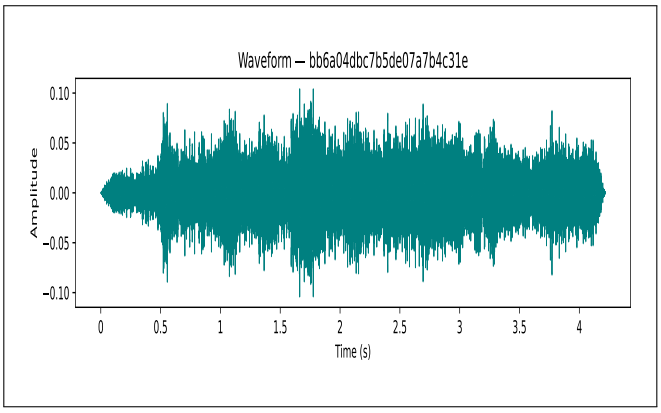

Fig. 6: Waveform visualization.

A randomly selected waveform visualization displays clear amplitude variation with minimal background noise or clipping. **Insight:** The dataset maintains high-quality audio, suitable for feature extraction and intelligibility modeling. The observed amplitude dynamics correspond to lyrical articulation patterns.

III-G Mel Spectrogram Example

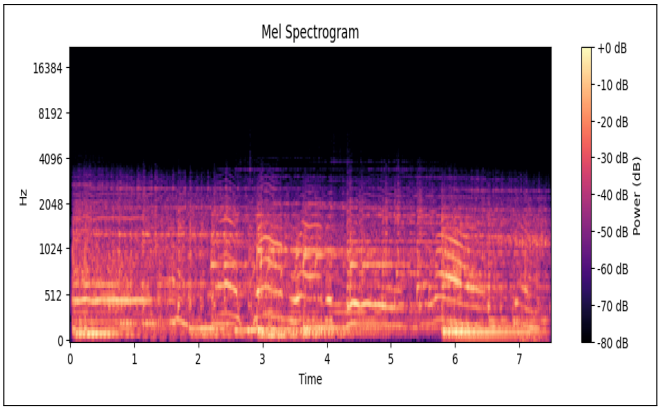

Fig. 7: Mel spectrogram example.

The Mel spectrogram, which represents frequency energy over time on a perceptual scale, shows distinct vocal frequency bands (primarily between 1–4 kHz). **Insight:** Energy concentration in mid-frequency ranges confirms that human vocal components are prominent and well-captured — a critical factor for assessing lyric clarity.

III-H MFCC Features Example

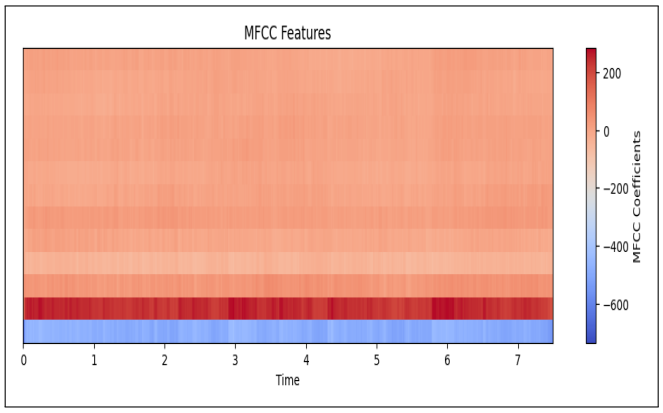

Fig. 8: MFCC feature visualization.

Mel-Frequency Cepstral Coefficients (MFCCs) were extracted as perceptual features that summarize how humans perceive sound timbre. The MFCC plot shows smooth temporal variation, validating correct extraction and the dataset's stability. **Insight:** MFCCs effectively encode vocal articulation and are strong candidate features for predicting intelligibility levels.

III-I Correlation Matrix

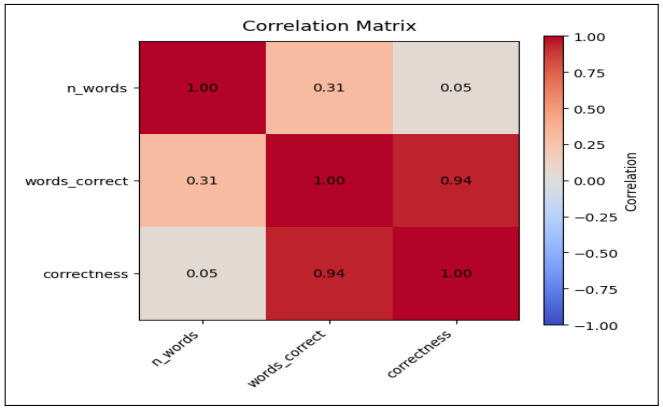

Fig. 9: Feature Correlation heatmap.

A correlation heatmap between n_words, words_correct, and correctness reveals:

• Strong correlation between words_correct and correctness (r = 0.94)

• Weak correlation between n_words and correctness (r = 0.05)

Insight: Lyric length has little effect on intelligibility, while correctness strongly depends on the number of correctly recognized words. Redundant variables should be avoided in model inputs.

IV Problem Formulation and Metrics

The CADENZA CLIP challenge defines the intelligibility prediction task as a regression problem — predicting a continuous correctness score between 0 and 1 for each lyric clip.

$$\text{Correctness} = \frac{\text{Number of Correct Words}}{\text{Total Words}} \tag{1}$$

The official evaluation metric, as per the challenge guidelines, is the Word Correctness Score, measuring how accurately a system estimates perceived intelligibility. Additional metrics such as Mean Absolute Error (MAE) or Pearson correlation may be used during local validation to quantify prediction accuracy.

V Challenges and Observations

TABLE I: Observed Issues and Recommendations

Category	**Observed Issue**	**Implication / Recommendation**
Target imbalance	Most samples have high correctness scores (> 0.7).	Consider weighted loss or balanced sampling.
Duration variability	Some clips are significantly longer (20s).	Standardize input lengths.
Acoustic noise	Minor differences in preprocessing.	Noise handling may improve robustness.
Correlation redundancy	High correlation between variables.	Avoid redundant features.
Hearing loss factor	Minor variation across groups.	Include as categorical input.

VI Task Division Table

TABLE II: Team Member Task Description

Team Member	**Task Description**
Maham Mansoor	Audio duration and sample rate analysis, frame difference histogram.
Rudaina Aamir	Intelligibility distribution and hearing loss comparison plots.
Wamiq ul Islam	Waveform, Mel Spectrogram, and MFCC visualization.
Zaina Zia	Correlation matrix analysis and report compilation.

VII Conclusion

The exploratory data analysis of the CADENZA CLIP Lyric Intelligibility Challenge dataset provides a comprehensive understanding of its structure, quality, and predictive challenges. The dataset is clean, balanced, and acoustically consistent, making it well-suited for downstream modeling. EDA results confirm that intelligibility is influenced primarily by audio clarity and word-level accuracy rather than clip length or hearing loss group. These insights will guide the next phases — feature engineering and model design — to develop systems capable of accurately predicting lyric intelligibility in real-world listening scenarios.

References

[1] ICASSP 2025 CADENZA CLIP1 Lyric Intelligibility Challenge. https://cadenzachallenge.org/docs/clip1/intro