Exploratory Data Analysis for Lyric Intelligibility Prediction: ICASSP 2026 Cadenza CLIP1 Challenge

Muhammad Musaab Ul Haq (501739)
SEECS, NUST
Islamabad, Pakistan

Usman Amjad (516261)
SEECS, NUST
Lahore, Pakistan

Ahmed Hassan Raza (511263)
SEECS, NUST
Lahore, Pakistan

Abdul Mueed Habib Raja (501166)
SEECS, NUST
Chasma, Pakistan

***Abstract*—This report presents an exploratory data analysis (EDA) of the Cadenza CLIP1 dataset. The goal of this challenge is to predict how well people with hearing impairments can understand song lyrics when using different hearing aids. Our analysis looks at four main areas: (1) the general distribution of intelligibility scores, (2) how word-level features affect understanding, (3) the role of audio properties, and (4) how listener mistakes relate to the scores. By exploring the entire training dataset, we engineered features to find what factors influence lyric intelligibility. These findings provide a starting point for building a model to help improve hearing aids for listening to music.**

***Index Terms*—Hearing aids, lyric intelligibility, audio signal processing, exploratory data analysis, assistive technology.**

I. Introduction and Motivation

Listening to music is often difficult for hearing aid users. While modern hearing aids are good at making speech clearer in noisy places, the complicated sounds in music, especially song lyrics, are a major challenge. The ICASSP 2026 Cadenza CLIP1 challenge aims to solve this problem by asking participants to predict how well a person with hearing loss can understand lyrics. The task can be done through two methods. One method would be directly calculate the score through the use of audio and perform operations on the audio and evaluate the score from that. Another approach given to us, is to use the whisper model and convert to text first and then perform an evaluation criteria from the text such that it results in the correction score.

The main objective is to build a model that can predict a listener's understanding of lyrics based on their hearing aid settings. This is important for a few reasons:

- It can help create better hearing aids for music.
- It allows for personalized settings for each user.
- It can improve the quality of life for music lovers with hearing loss.

II. Dataset Overview

The Cadenza CLIP1 dataset is large and contains thousands of audio clips and corresponding listener data. It gives us everything we need to analyze this problem and includes a few key types of data (modalities):

- **Audio Signals**: Music clips are provided in two versions: the original, unprocessed audio and the version processed by a simulated hearing aid.
- **Listener Transcripts**: Text of what listeners thought they heard.
- **Ground Truth Lyrics**: The actual lyrics from the song.
- **Intelligibility Scores**: A score from 0 to 1 that tells us how much of the lyrics the listener understood correctly.
- **Hearing Profiles**: Information on the type of hearing loss for each listener (e.g., Mild, Moderate).

An example entry includes a processed audio file, the listener's attempt at transcribing the lyrics (e.g., "I see a red door"), the true lyric ("I see a red door and I want it painted black"), and a correctness score.

III. Exploratory Data Analysis (EDA)

A. Distribution of Intelligibility Scores

First, we looked at the distribution of the "correctness" scores, which is our target variable. As shown in Figure 1, the scores form two main groups. There is a large peak of very high scores (above 0.8) and another large peak of very low scores (below 0.2). This suggests that listeners either understood the lyrics very well or not at all, with fewer people in the middle. This observation is important because it might mean we can classify listeners into "high understanding" and "low understanding" groups. **For each of the analysis following from here onward, there would be a corelation score at the top right of the plot, to allow the reader to comprehend the corelation between the two variables in a much more clearer manner (as some plots often have too many scattered points).**

The corelation score is the variable 'r'.

B. Signal Comparison

Figure 2 compares an original audio clip with its processed version. The waveforms and spectrograms show that

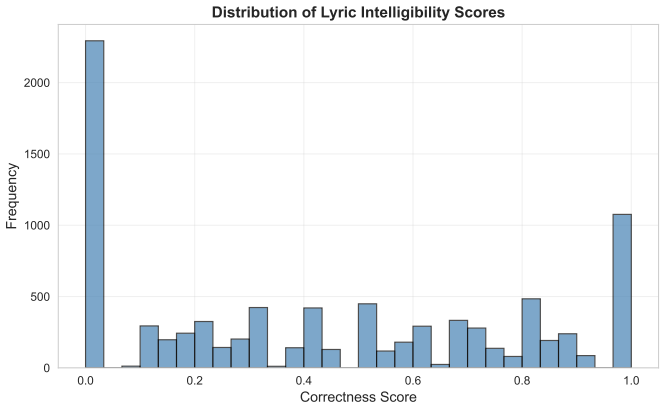


Fig. 1: Distribution of lyric intelligibility scores, showing two distinct groups of listeners.

the processing changes the audio's amplitude and frequency content by quite a big difference. It dampens alot of the larger amplitudes.

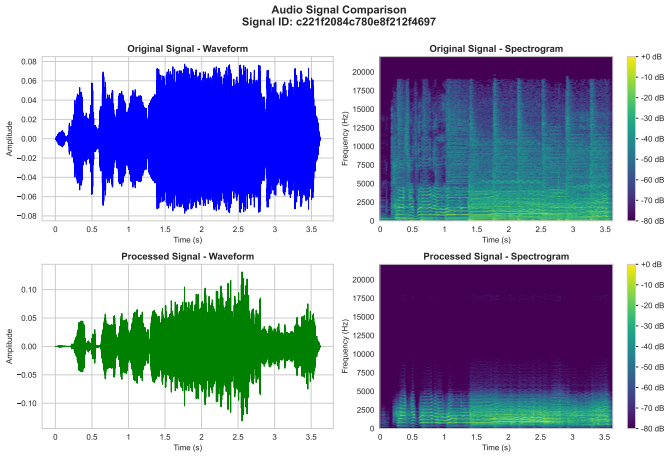


Fig. 2: A comparison of original and processed audio signals.

C. Linguistic Features

We explored if the words themselves made understanding harder.

- **Word Length**: Figure 3 shows that there is a slight negative trend (albeit small enough to be neglegible, r=-0.098) between the average word length and the intelligibility score. Our hypothesis was that lyrics with longer words might be a bit harder to understand but that seems to be untrue to a large extent.
- **Word Frequency**: We also checked if common words are easier to hear. In Figure 4, we see that the hypothesis 'words that are moderately common tend to have the highest intelligibility scores' is hardly true, given the corelation score of only -0.020.

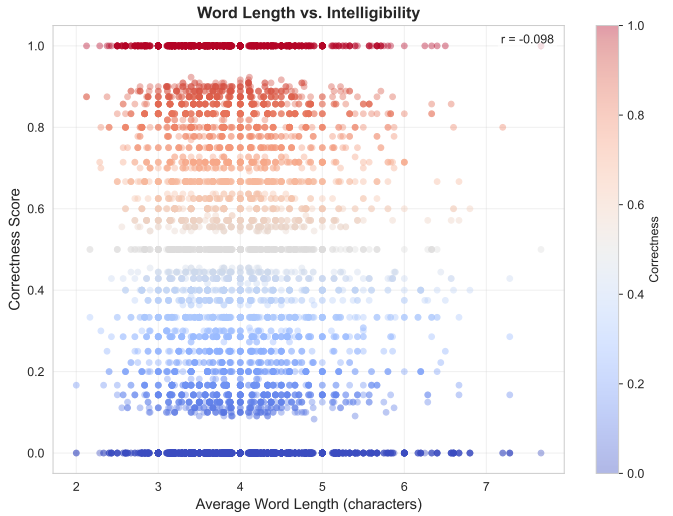


Fig. 3: Longer words show a slight trend towards lower intelligibility.

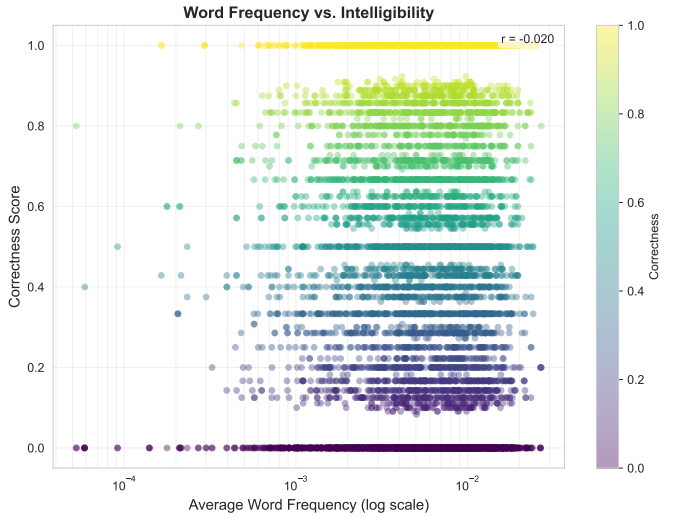


Fig. 4: Moderately common words appear to be the easiest to understand.

D. Acoustic Analysis

We analyzed the audio's "brightness" using a feature called the spectral centroid. Figure 5 (figure at next page) shows that intelligibility is somewhat correlated to spectral centroid. This suggests that an optimal frequency range does affect lyric clarity.

E. Listener Error Analysis

To confirm that the correctness score is a good measure of understanding, we compared it to the number of mistakes in the listeners' transcriptions. We used the normalized Levenshtein edit distance to count these mistakes. Figure 6 (figure at next page) shows a strong inverse relationship: as the number of mistakes (edit distance) goes up, the correctness score goes down. This confirms the edit distance can be a reliable metric

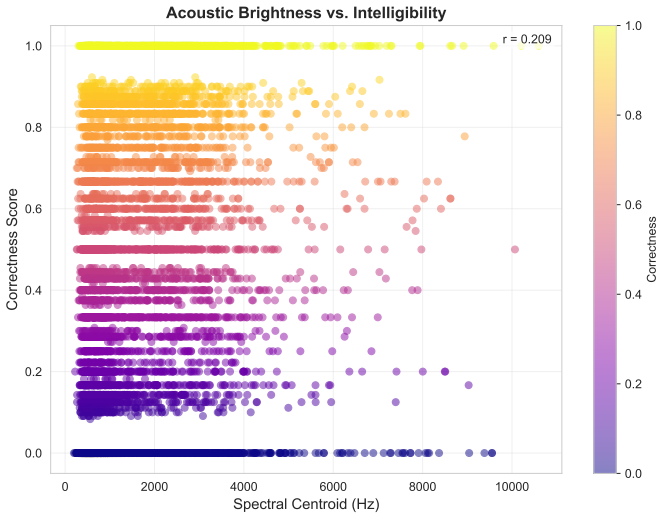


Fig. 5: An optimal ”brightness” range seems to exist for the best lyric understanding.

for identifying correct lyrics. Unfortunately, edit score requires text and cant be directly applied to audio, hence if we would be approaching the task from approach 2, only then can we apply this metric in some form. If confused what I mean by approach 2, then please refer to **Introduction and Motivation** section.

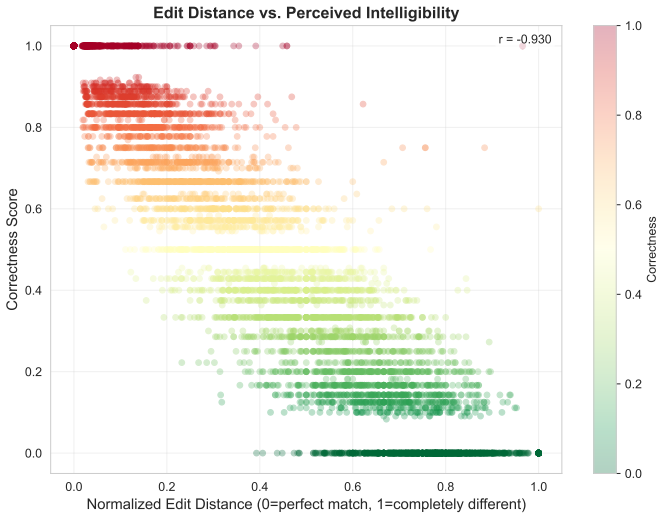


Fig. 6: Fewer transcription mistakes (lower edit distance) strongly correlates with higher intelligibility scores.

F. Analysis Across Hearing Loss Groups

Finally, we wanted to see if these trends were different for people with different levels of hearing loss. Figure 7 (figure attached at the end) splits the word length analysis by hearing loss category. It shows that the negative effect of longer words is stronger for listeners with more severe hearing loss, if we find the ratio of corelation among all the loss category. This is a weak insight, as it suggests that the best prediction models will need to take a listener's individual hearing profile into account.

IV. Problem Formulation and Metrics

A. Problem Formulation

The main task is to predict the lyric intelligibility for a given listener and audio clip. This is a regression problem where the goal is to predict a continuous value.

- **Input**: Processed audio signal, listener's hearing loss profile, and the original lyrics.
- **Output**: A predicted ”correctness” score between 0 and 1.

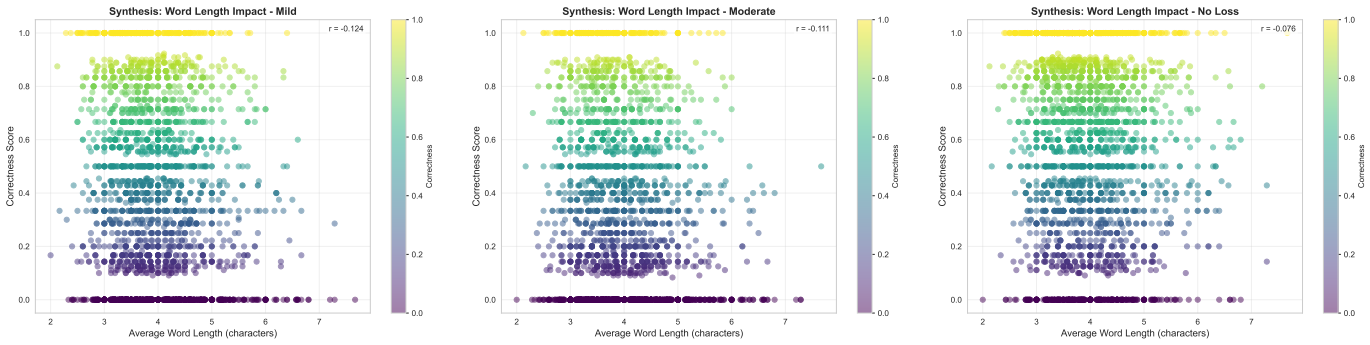
B. Evaluation Metrics

The primary metric for evaluating our model will be how close our predicted scores are to the true scores. A common metric for this is the Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values. A lower MSE means a better model. Additionally, because of the bimodal distribution we observed, we might also explore classification metrics like accuracy to see how well a model can classify a song as having ”high” or ”low” intelligibility.

V. Task Division

The work for this exploratory data analysis was divided among the team members as follows:

- **Muhammad Musaab Ul Haq**: Created Figures 1, 2, and 3, focusing on the foundational analysis of the target variable, signals, and word length.
- **Usman Amjad**: Created Figures 4 and 5, which involved analyzing linguistic frequency and acoustic properties like the spectral centroid.
- **Abdul Mueed Habib Raja**: Created Figures 6 and 7, focusing on perceptual error analysis and the synthesis of features across different hearing loss groups.
- **Ahmed Hassan Raza**: Was responsible for compiling the findings and creating the final LaTeX report.



(a) Mild Hearing Loss (b) Moderate Hearing Loss (c) No Hearing Loss

Fig. 7: The impact of word length on intelligibility for different hearing loss groups. The negative trend is stronger for those with more severe hearing loss.