ASR Evaluation Report

1. Introduction

This document presents a performance assessment of two Automatic Speech Recognition (ASR) system outputs generated from a single spoken audio recording. The goal of this evaluation is to compare both systems using standard accuracy measures such as Word Error Rate (WER) and Character Error Rate (CER). In addition, the acoustic properties of the speech signal are examined using waveform, spectrogram, and RMS energy visualizations.

2. Dataset Description

The dataset used for this analysis consists of:

One audio file:

single_s1_22032_11908.wav

Two ASR transcription files:

single_s1_22032_11908_asr1.tsv

single_s1_22032_11908_asr2.tsv

Each TSV file includes the following four columns:

Start time of the segment

End time of the segment

ASR-generated transcription

Ground truth reference transcription

3. Evaluation Approach

The evaluation process involved computing both sentence-level and corpus-level accuracy scores using standard ASR performance metrics. The following steps were implemented using Python:

Loading both TSV transcription files

Performing text normalization

Calculating sentence-wise WER and CER

Computing overall corpus-level WER and CER

Saving the processed results into CSV output files

(asr1_results.csv and asr2_results.csv)

The jiwer Python library was utilized to perform the WER and CER calculations.

4. Evaluation Results

Results for single_s1_22032_11908_asr1.tsv

Output File: asr1_results.csv

Corpus WER: 0.271712158808933

Corpus CER: 0.13094034378159758

Results for single_s1_22032_11908_asr2.tsv

Output File: asr2_results.csv

Corpus WER: 0.4739454094292804

Corpus CER: 0.2206774519716886

Performance Comparison

From the corpus-level error rates, it is clearly observed that ASR1 outperforms ASR2. ASR1 records approximately 27% word-level error, whereas ASR2 shows about 47%, indicating that ASR1 produces significantly more accurate transcriptions for this audio sample.

5. Error Analysis

A close examination of the sentence-level errors reveals that ASR2 introduces a larger number of substitutions and deletions compared to ASR1. Frequent issues include incorrect recognition of acoustically similar words, omission of function words, and insertion of unrelated phrases. In contrast, ASR1 remains more consistent with the reference transcript, demonstrating improved alignment and decoding stability.

Example Comparisons

Timestamp: (46.5 - 59.01)

ASR1 Output:

Yes, I have got an idea about the topic and what has to be spoken. Definitely, I believe individuals at...

ASR2 Output: ON YIS I HAVE GOT A IDEAN AOGI WITH THE GOPIC HAND THE WORT HAS TO BE SPOKEN AND DEFINITELY I BELIEVE ON INDOVIJUELS AT THE EARL

Timestamp: (502.8 - 517.2)

ASR1 Output:

Yes, I already said you this only that I have also the same issue. But I think better is to stick to the task because even if it is in English, because the rule is...

ASR2 Output:

YES I O I ALRETE AT YOU THIS ONLY THAT I HAVE ALSO THE SAME ICIU BUT ELL I THINK BETTERLY STRU AND STICK TO THE TASK BECAUSE EH EVEN IF IT ISN'T ENGLISH BECAUSE THE RULE IS

These examples highlight the stronger transcription consistency achieved by ASR1.

6. Audio Signal Analysis

To study how speech characteristics influence ASR performance, the audio signal (single_s1_22032_11908.wav) was examined using:

Waveform visualization

Time–frequency spectrogram

RMS energy plot

These representations assist in understanding speech structure, pauses, and energy variations that may impact recognition accuracy. The figures were generated using Python libraries such as librosa and matplotlib.

6.1 Waveform Visualization

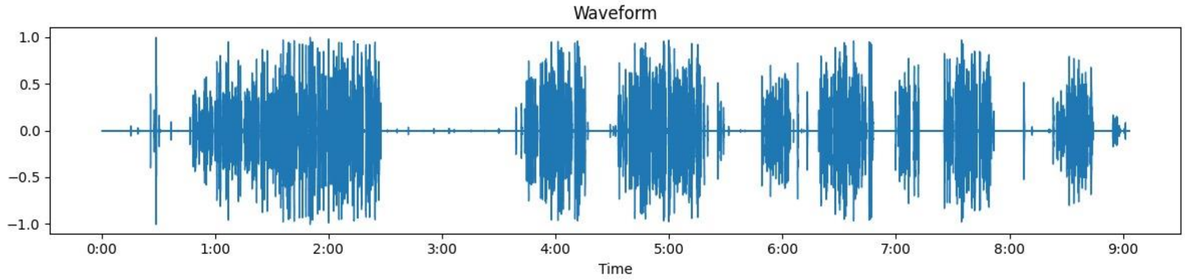

Figure 6.1: The waveform plot clearly outlines regions of active speech separated by noticeable silent gaps, suggesting a well-structured speaking pattern.

The waveform displays amplitude variations across time. Higher peaks correspond to spoken content, while flatter regions indicate silence. This plot helps relate speech intensity to possible ASR difficulty.

6.2 Spectrogram (Time–Frequency Analysis)

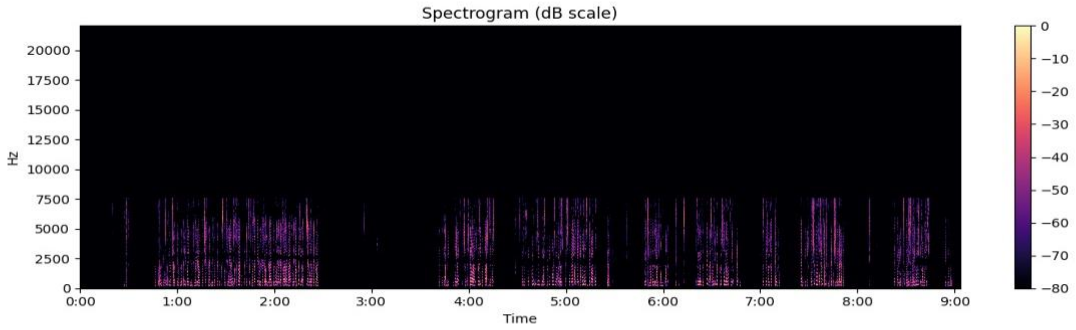

Figure 6.2: The spectrogram indicates that most speech energy is concentrated below 6 kHz, highlighting voiced speech regions, articulation details, and silence periods.

Bright frequency bands represent high-energy speech components, whereas darker areas correspond to weaker or silent segments. This visualization aids in understanding pronunciation clarity and potential causes of recognition errors.

6.3 RMS Energy Plot

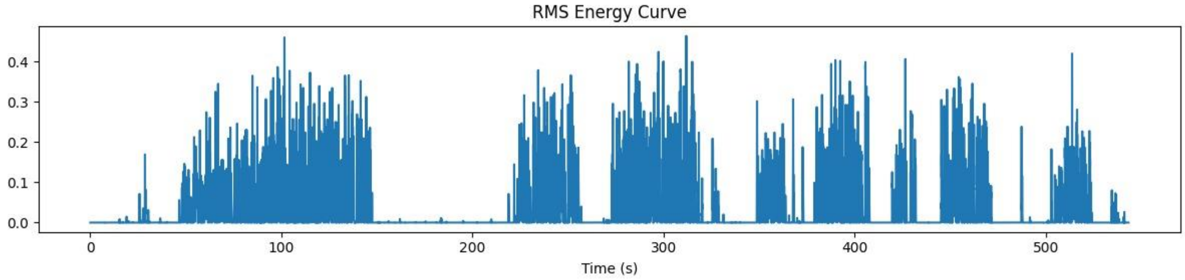

Figure 6.3: The RMS energy curve reflects frame-wise energy variations, closely matching the waveform and revealing speaking rhythm and emphasis.

Higher RMS peaks correspond to strongly articulated speech, while lower values suggest pauses or silence. RMS analysis is effective for detecting speech boundaries.

Overall, the visual analysis confirms a clear segmentation of voiced and silent intervals within the speech signal. These acoustic properties are consistent with the segmentation in the ASR files. Energy and frequency variations may explain why ASR2 exhibits a higher error rate in certain regions.

7. Conclusion

This study demonstrates how ASR system performance can be quantitatively evaluated using WER and CER, and how acoustic visualizations support the interpretation of transcription challenges. The results clearly show that ASR1 achieves significantly lower error rates than

ASR2, making it the more reliable system for this audio sample. The combined use of numerical metrics and signal analysis offers a robust framework for assessing ASR behavior.

8. References

Dataset provided by IISc SPIRE Lab

Python Libraries Used: pandas, jiwer, librosa, matplotlib

Spectrogram obtained through Short-Time Fourier Transform (STFT)