ASR Evaluation Report

1. Introduction

This report presents the evaluation of two automatic speech recognition (ASR) outputs corresponding to a single spoken audio sample. The objective is to compare two ASR systems using accuracy metrics such as Word Error Rate (WER) and Character Error Rate (CER), and to interpret acoustic characteristics through audio signal visualization including waveform, spectrogram, and RMS energy.

2. Dataset Information

The dataset contains:

- One speech audio file: *single_s1_22032_11908.wav*
- Two aligned ASR transcription files:
    - *single_s1_22032_11908_asr1.tsv*
    - *single_s1_22032_11908_asr2.tsv*

Each TSV file contains four columns:

- Start timestamp
- End timestamp
- ASR Output (system-generated text)
- Reference transcript (human transcription)

3. Evaluation Methodology

The ASR evaluation was conducted using corpus-level and sentence-level accuracy metrics. Word Error Rate (WER) and Character Error Rate (CER) were computed using standard definitions based on substitutions, deletions, and insertions relative to the reference transcript.

A Python script was developed to:

- Read each TSV file
- Normalize text
- Compute sentence-wise WER and CER
- Calculate overall WER and CER for the entire audio
- Export evaluated results into CSV files (*asr1_results.csv* and *asr2_results.csv*)

The library jiwer was used to compute WER/CER.

4. Results

==== Results for single_s1_22032_11908_asr1.tsv ====

Output CSV: asr1_results.csv

Corpus WER: 0.271712158808933

Corpus CER: 0.13094034378159758

==== Results for single_s1_22032_11908_asr2.tsv ====

Output CSV: asr2_results.csv

Corpus WER: 0.4739454094292804

Corpus CER: 0.2206774519716886

Based on corpus-level WER and CER, ASR1 performs better than ASR2.
ASR1 has ~27% word-level error versus ASR2's ~47%, showing ASR1 produces significantly more accurate transcriptions.

5. Error Analysis

Inspection of sentence-level errors shows ASR2 tends to make more substitutions and deletions than ASR1. Common issues include misrecognition of similar-sounding words,

dropped function words, and insertion of irrelevant phrases. ASR1 generally follows the reference transcript more closely, indicating stronger alignment and more stable decoding.

For Reference:

At 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

At 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

6. Audio Signal Analysis

To understand how speech properties affect ASR system behaviour, the audio file (*single_s1_22032_11908.wav*) was analysed using waveform visualization, time–frequency (spectrogram) representation, and Root Mean Square (RMS) energy analysis.

These visualizations help interpret speech structure, pauses, and intensity variations which may relate to recognition accuracy. Python libraries such as *librosa* and *matplotlib* were used to generate the figures.

6.1 Waveform Plot

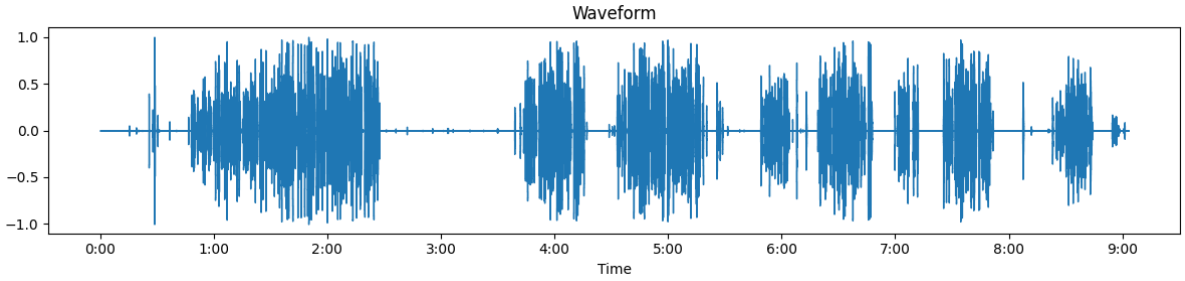

Fig 6.1: The waveform identifies active speech segments separated by clear pauses, indicating structured spoken delivery.

The waveform visualization shows the amplitude variation over time. Peaks indicate spoken segments, whereas flatter regions indicate silence or pauses. This helps correlate transcription difficulty with speech energy changes.

6.2 Spectrogram (Time–Frequency Representation)

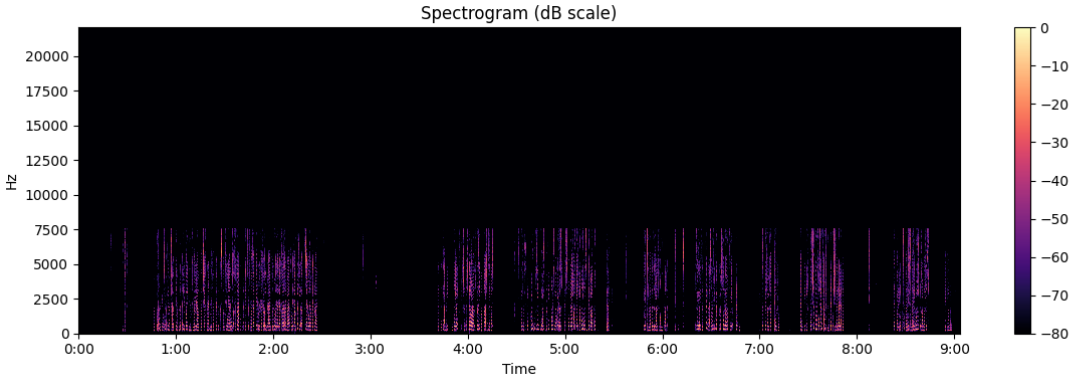

Fig 6.2: The spectrogram shows speech energy concentrated below 6 kHz, revealing voiced regions, articulation, and silence intervals.

The spectrogram illustrates how different frequency components evolve over time. Bright regions represent higher energy frequencies associated with voiced phonemes, while darker regions correspond to silence or weak articulation. Observing the spectrogram helps understand pronunciation clarity and possible sources of ASR confusion.

6.3 RMS Energy Plot

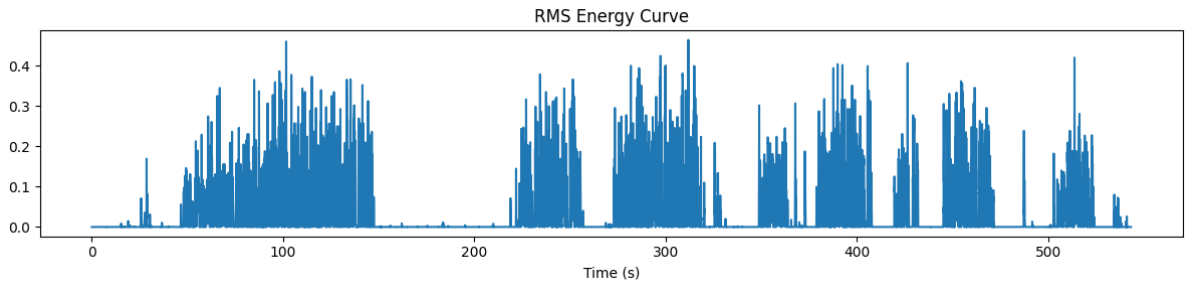

Fig 6.3: RMS energy curve highlights frame-wise energy variation, matching the waveform and reflecting speaking rhythm and emphasis.

The RMS energy curve highlights the dynamic intensity of speech. Higher peaks relate to strong articulation or stressed syllables, while minima correspond to silence or pauses. RMS analysis is useful for segment boundary visualization and detecting speech/non-speech separation.

Overall, the visual inspection confirms clear speech structure with identifiable voiced and silent regions. These acoustic properties align with the grammatical segmentation used in the ASR files. Variations in amplitude and frequency richness may indicate regions where ASR performance differs, particularly for ASR2 which shows higher error rates.

7. Conclusion

This evaluation task demonstrated how ASR system performance can be quantified using WER and CER, and how acoustic visualization assists in interpreting transcription difficulties. ASR1 exhibited significantly lower error rates than ASR2, making it the preferred system for this sample. The combination of quantitative metrics and audio analysis provides a structured approach for assessing ASR behaviour.

8. References

- Dataset provided by IISc SPIRE Lab
- Python libraries: pandas, jiwer, librosa, matplotlib
- Spectrogram visualization based on STFT analysis