Speech-to-Text: A Comprehensive Analysis of State-of-the-Art Models, Challenges, and Future Directions

Himani(M23CSA516)[1], Ankit Kumar Chauhan(M23CSA509)[2]
[1]Email: m23csa516@iitj.ac.in
[2]Email: m23csa509@iitj.ac.in

January 31, 2025

1 Overview of the Task

The purpose of speech-to-text (STT) systems is to translate spoken words into written text. These systems are essential to many real-world applications, including closed captioning, transcription services, virtual assistants (like Siri and Alexa), and improving accessibility for those with impairments.

Real-World Importance: STT technology facilitates smooth human-machine interface, increases productivity, and expedites communication. In particular, it helps people with hearing impairments fill accessibility gaps.

2 State-of-the-Art (SOTA) Models and Methods

Language	**Model**
English	Wav2Vec2-Base-960h
English	SpeechT5-ASR
Punjabi	Wav2Vec2-Large-XLSR-Punjabi
Punjabi	Whisper

Table : Used various SOTA

3 Implementation and Analysis

Dataset Used: Datasets for High Resource Language: English and Low resource language: Punjabi , including original sentences and their predicted transcriptions.

Consider the first 500 records to evaluate the model's performance.

4 Evaluation Metrics

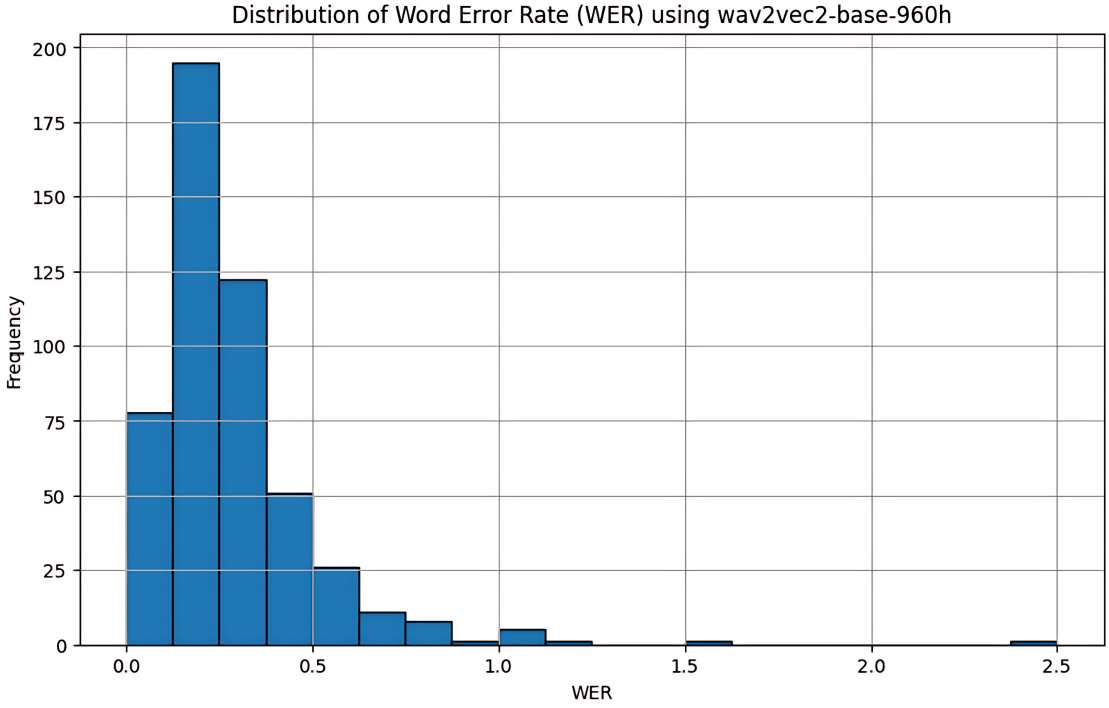
Metrics Used:

1. **Word Error Rate (WER):**

   - Formula:
     where = Substitutions, = Deletions, = Insertions, and = Total words in the reference.
   - **Strengths:** Provides a clear measure of transcription accuracy.
   - **Limitations:** Does not account for semantic understanding or context.

Analysis for High Resource language: English using Wav2Vec2-Base-960h:

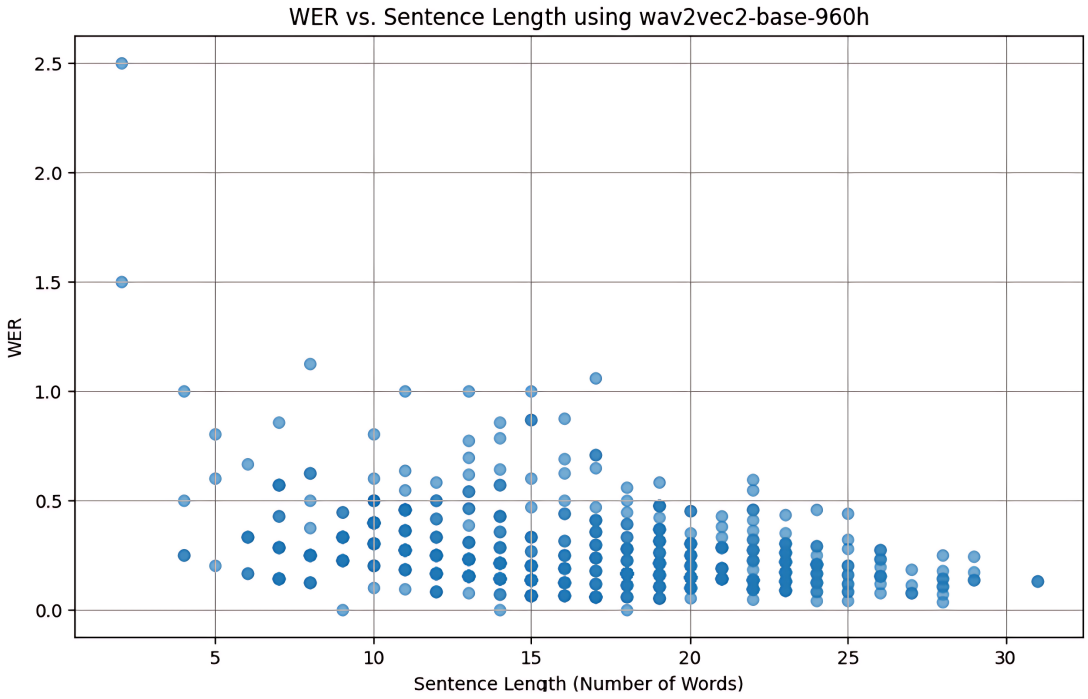
Results



(a)

Distribution of Word Error Rate (WER):

- For short or simple sentences, the majority of the sentences have a WER between 0.0 and 0.3, which indicates great transcription accuracy.
- A WER larger than 1.0 is seen in fewer instances, suggesting that complicated or loud sentences are difficult to accurately transcribe.
- As WER rises, the histogram displays a sharp drop in frequency, indicating that errors are rare and only happen under certain situations.



(b)

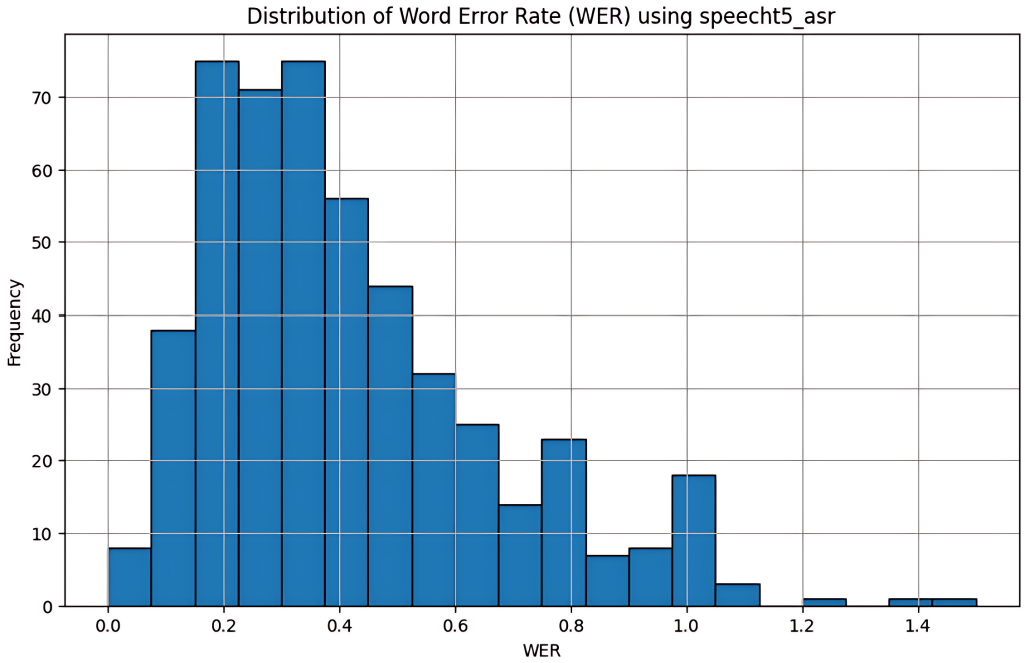
Fig (a)(b): WER Distribution and its Relationship with Sentence Length for Wav2Vec2-Base-960h

WER vs. Sentence Length:

- WER scores for short sentences are typically lower, demonstrating the model's ability to handle brief utterances.
- WER levels are marginally higher for longer words, with some outliers having WER values above 1.5.
- For most of the data, there is no significant link between sentence length and WER, although the scatter plot shows variation in performance with phrase complexity.

Analysis for High Resource language: English using speecht5_asr:

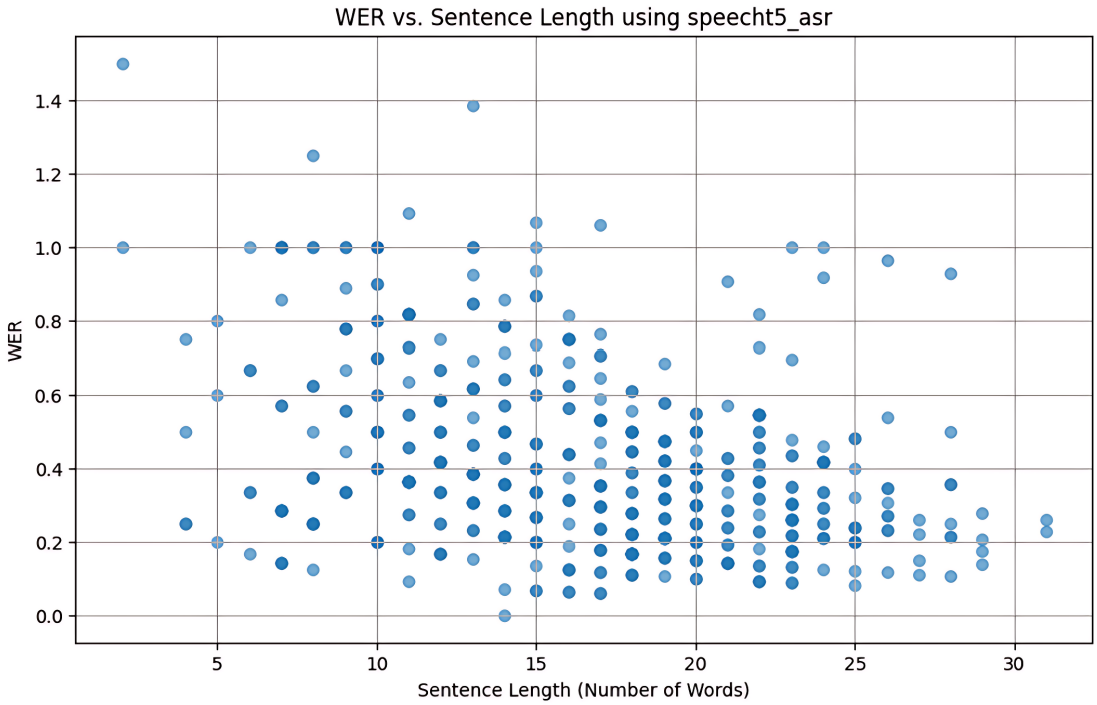
Results



(c)

Distribution of Word Error Rate (WER):

- For general sentences, the majority of WER values range from 0.2 to 0.6, which denotes moderate transcription accuracy.
- WER scores above 1.0 are less common, indicating sporadic difficulties with noisy or complex data.



(d)

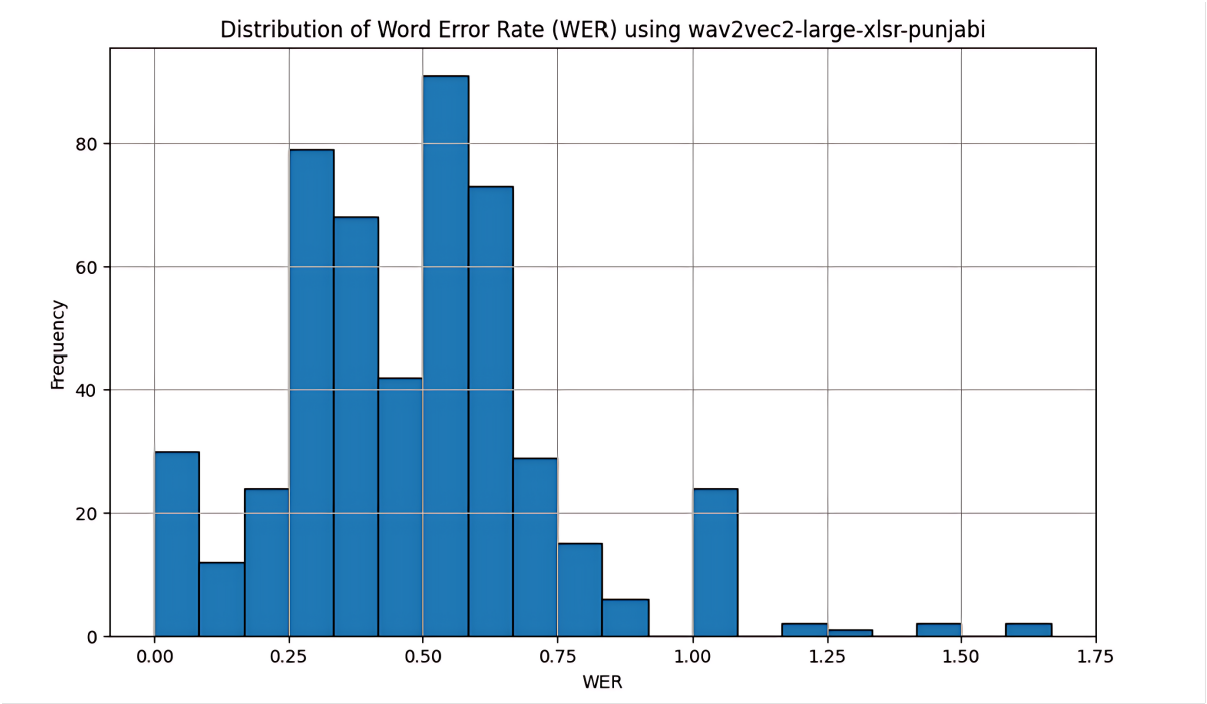
Fig (c)(d): WER Distribution and its Relationship with Sentence Length for speecht5_asr

WER vs. Sentence Length:

- Short sentences function well for brief inputs, as seen by their lower WER values.
- WER varies more in longer sentences, with errors clearly increasing for sentences longer than 15 words.
- High WER outliers (>1.0) indicate difficulties sustaining accuracy in particular situations, most commonly as a result of complicated sentences or different accents.

Analysis for low Resource language: Punjabi using wav2vec2-large-xlsr-punjabi:

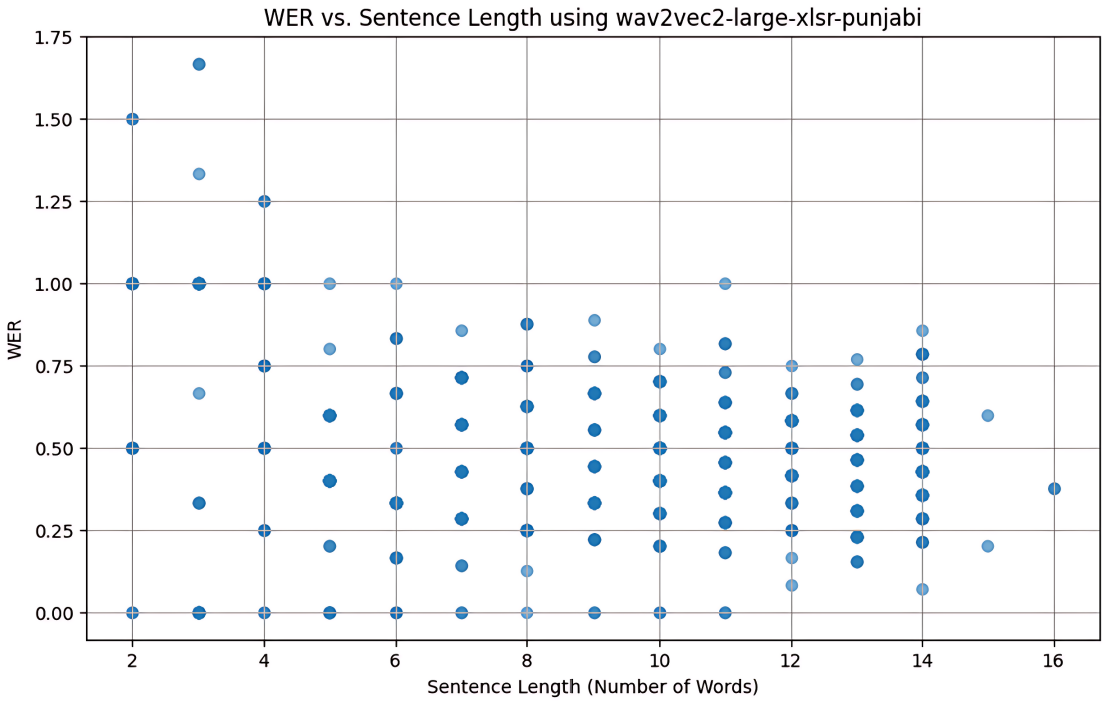
Results



(e)

Distribution of Word Error Rate (WER):

- For Punjabi data, the majority of WER values fall between 0.25 and 0.75, indicating a reasonable level of transcription accuracy.
- Fewer examples had WER near 0.0, indicating that it may be difficult to get high accuracy for simpler texts.
- Only a few sentences have a WER higher than 1.0, which may indicate problems with audio quality or complexity of language structures.



(f)

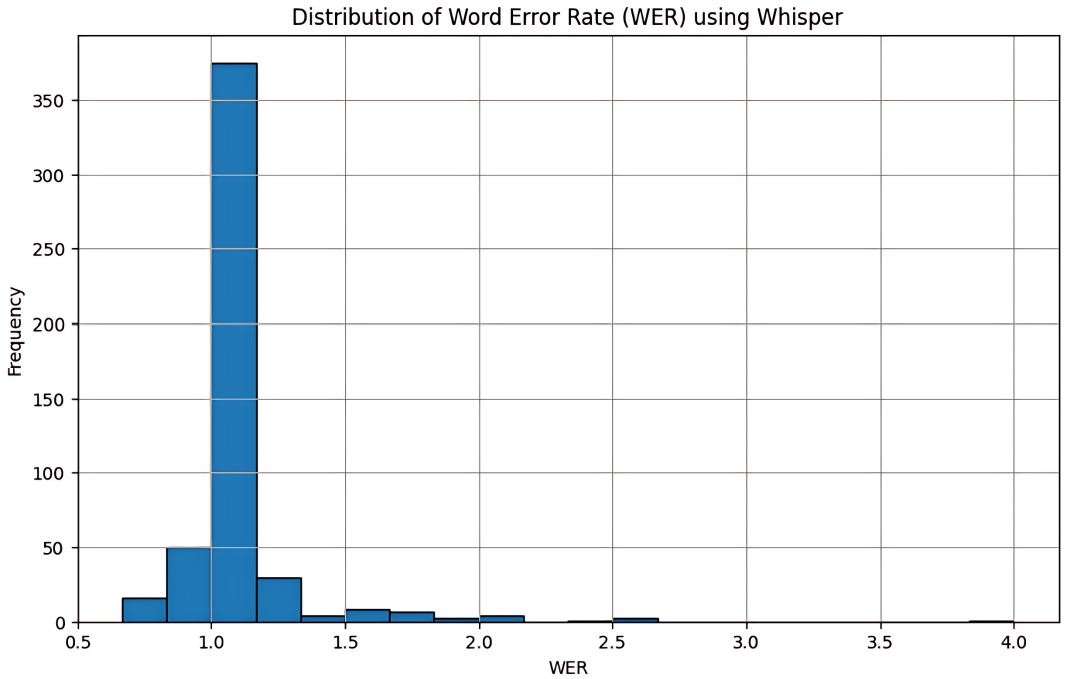
Fig (e)(f): WER Distribution and its Relationship with Sentence Length for wav2vec2-large-xlsr-punjabi

WER vs. Sentence Length:

- Although the errors are more uniformly spread throughout phrase lengths, short sentences typically have lower WER levels.
- WER varies for sentences longer than ten words, although there is no clear relationship between sentence length and transcription accuracy.
- Across a range of sentence lengths, there are outliers with WER greater than 1.0, which point to particular issue areas unrelated to sentence complexity.

Analysis for low Resource language: Punjabi using whisper:

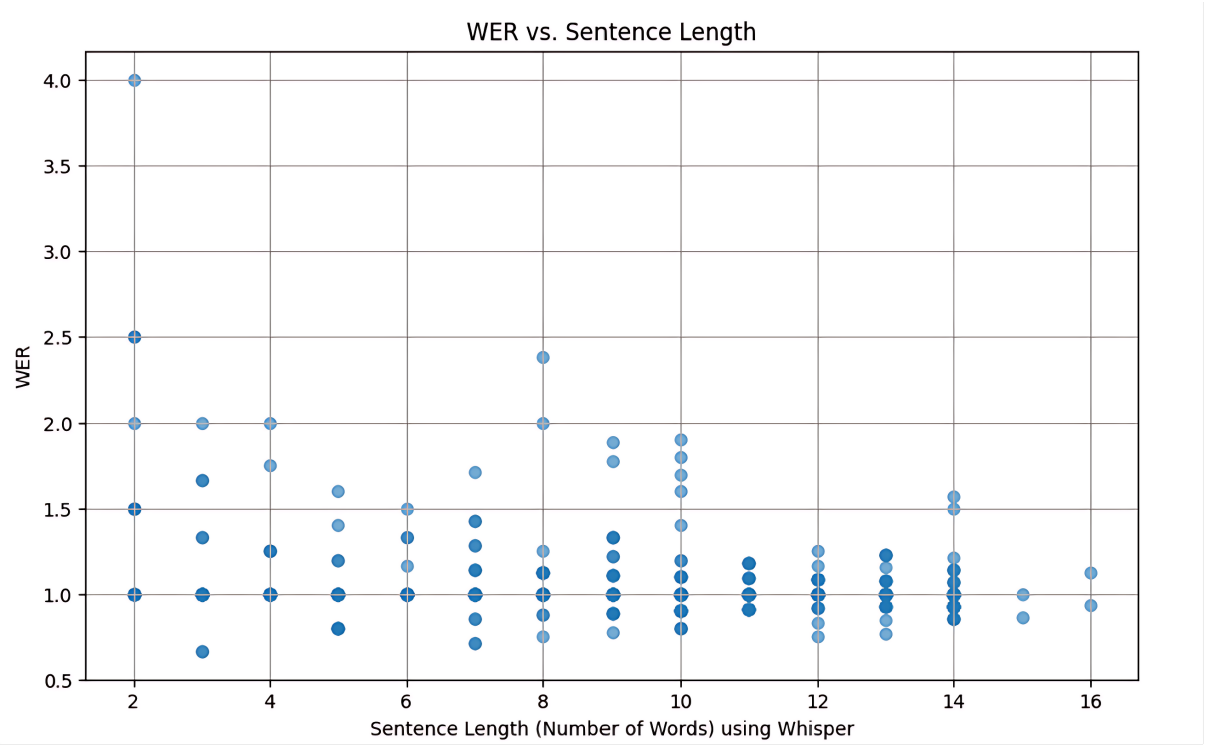
Results



(g)

Distribution of Word Error Rate (WER):

- For the bulk of sentences, moderate transcription errors are indicated by the concentration of WER values around 1.0.
- A small number of outliers have WER values higher than 2.0, indicating large errors in certain situations, possibly brought on by noisy data or complex linguistic patterns.
- Whisper's performance is constant, but skewed toward moderate error rates, according to the histogram.



(h)

Fig (g)(h): WER Distribution and its Relationship with Sentence Length for whisper

WER vs. Sentence Length:

- WER values for short sentences (less than five words) are often near 1.0.
- There is no discernible relationship between transcription mistakes and sentence length; WER values stay constant as sentence length increases.
- Across a range of phrase durations, outliers with WER values more than 2.0 occasionally surface, indicating that the model faces particular difficulties unrelated to sentence length.

Language	Model	Strengths	Limitations	Key Observations
English	Wav2Vec2-Base-960h	High accuracy for clean data	Struggles in noisy/multi-accent settings	Performs well for short/simple sentences; challenges with noise and long sentences
English	SpeechT5-ASR	Unified framework, strong accuracy	High computational cost, slower inference speed	Good for brief inputs but higher WER for long, complex sentences
Punjabi	Wav2Vec2-Large-XLSR-Punjabi	Effective in low-resource settings	Limited robustness due to small datasets	Handles short sentences well; errors spread uniformly across sentence lengths
Punjabi	Whisper	Robust, multilingual, noise-tolerant	High computational cost; slower inference speed. Need another library to convert the text output from English to Punjabi	Stable WER across sentence lengths but higher base error rates, also not generates the text in punjabi used another library for this text to text translation.

Table: Comparison of Speech-to-Text Models Across High-Resource and Low-Resource Languages

References

1. https://huggingface.co/facebook/wav2vec2-base-960h
2. https://huggingface.co/microsoft/speecht5_asr
3. https://huggingface.co/manandey/wav2vec2-large-xlsr-punjabi
4. https://huggingface.co/openai/whisper-base

[12pt,a4paper]article [utf8]inputenc graphicx amsmath booktabs float geometry margin=1in tabularx hyperref

5 Analysis and Experimenting with Spectrograms and Windowing Techniques

Himani(M23CSA516)[1]
[1]Email: m23csa516@iitj.ac.in

Task A

Spectrogram Comparison and Analysis: Windowing Methods

The signal is divided into overlapping frames using windowing during the Short-Time Fourier Transform (STFT). The trade-off between frequency resolution, amplitude accuracy, and spectral leakage is influenced by the window selection. An extensive comparison of spectrograms produced using Hann, Hamming, and Rectangular windows is given below, with an emphasis on the technical characteristics, visual distinctions, and accuracy of the windowing procedure. Figure 1 below displays three spectrograms for a single windowing approach.

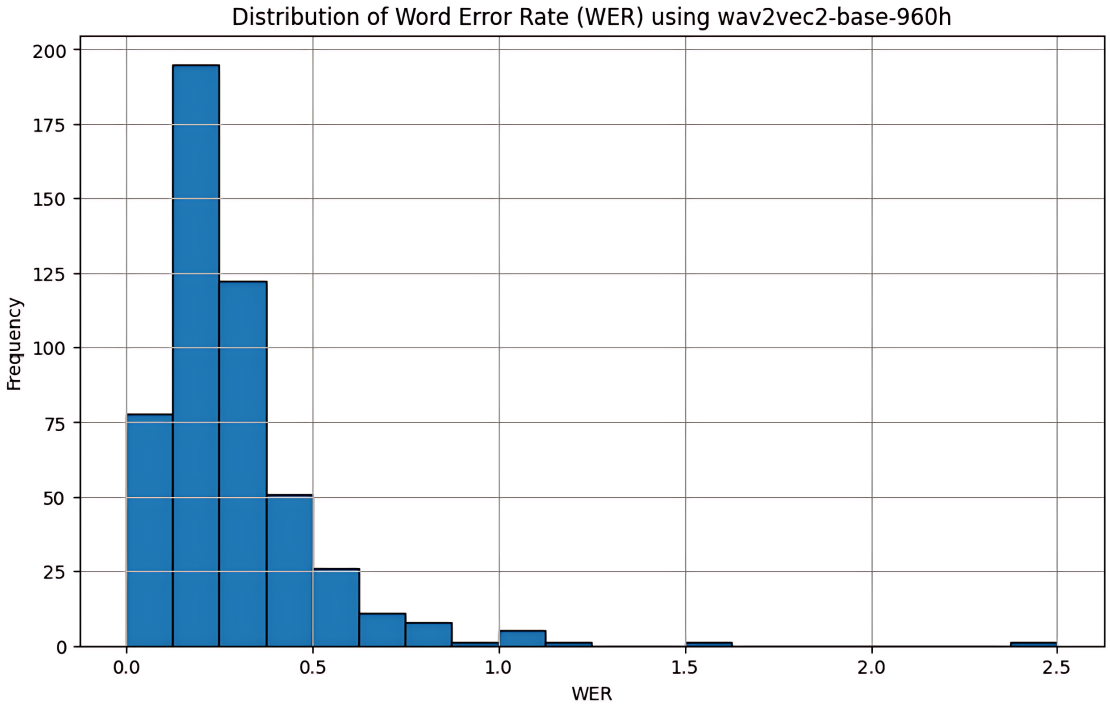


Figure 1: Spectrograms with different windowing for an audio file.

Hann Window Spectrogram:

Observation:

- Around each segment's edges, frequency components seem smoother and less noticeable.

- There is little spectral leakage into nearby frequencies, and the energy is well-concentrated.
- Amplitudes steadily decrease, particularly near the boundary.

Implication:

- The clear frequency bands show that this window effectively lowers spectral leakage.
- Due to the lack of significant differentiation between high-frequency content, there is a modest reduction in resolution.

Hamming Window Spectrogram:

Observation:

- A few more noticeable frequency bands than Hann, particularly near segment boundaries.
- There is slightly more leakage into neighboring frequencies, but there is better amplitude continuity across time. l leakage into nearby frequencies, and the energy is well-concentrated.
- Though slight, the change in leakage from the Hann window is discernible.

Implication:

- By maintaining improved resolution and reducing leakage, the Hamming window offers a well-rounded strategy.
- It works well for applications like voice or tonal signals because of its more seamless time segment transitions.

Rectangular Window Spectrogram:

Observation:

- The frequency bands exhibit distinct vertical boundaries between time segments and are well defined.
- It is evident that there is significant spectrum leakage, with energy from dominating frequencies leaking into nearby bands.
- Artifacts, or noise-like patterns, are caused by abrupt temporal segment boundaries.

Implication:

- Because it offers the highest resolution, this window is perfect for signals with isolated, basic frequencies.
- But because of its significant leakage, it is not appropriate for complicated, real-world signals like speech or audio.

Aspect	Hann Window	Hamming Window	Rectangular Window
Frequency Resolution	Moderate (smooth transitions)	Moderate to High (slightly sharper)	High (sharp transitions)
Spectral Leakage	Minimal (least leakage)	Low (slightly more leakage than Hann)	High (most leakage visible)
Edge Effects	Smooth tapering to zero at edges	Gradual tapering, but edges are slightly higher	No tapering; abrupt transitions at edges
Amplitude Continuity	Smooth and consistent	Smooth but slightly sharper transitions	Discontinuous; abrupt changes visible
Best Use Case	Signals needing smooth energy distribution (e.g., music, audio)	Signals with moderate continuity (e.g., speech)	Synthetic or periodic signals where resolution is key
Correctness	Correctly applied; ideal for real-world signals needing low leakage	Correctly applied; balances leakage and resolution	Correctly applied but suboptimal for real-world signals due to high leakage

Table: summarizing the **key differences** and the **correctness of windowing performed**

Comparative Summary:

- When smooth transitions and little spectrum leakage are essential, the **Hann window** is the ideal option.
- The **Hamming window** is adaptable for a variety of signals, including speech, since it successfully strikes a compromise between resolution and leakage.
- Although the **Rectangular window** is applied appropriately, it works best with regulated or synthetic signals where leaking is less of an issue. Its high leakage can impair performance for real-world signals.

Performance Evaluation of Spectrogram Features with Neural Network Classifier

A basic neural network classifier trained and evaluated using characteristics taken from spectrograms produced using the Hann, Hamming, and Rectangular windowing approaches. the features in order to explore the impact of windowing on classification performance.

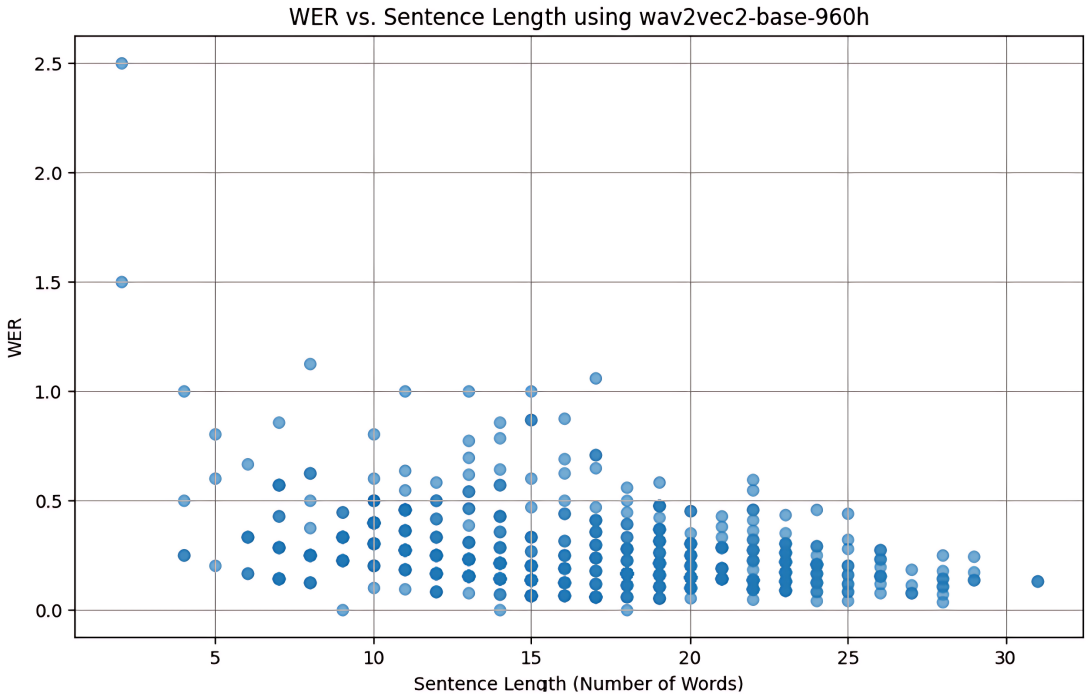


Figure 2: Accuracy and loss graphs of neural networks using different window type.

Key Observations from Graphs

Training and Validation Loss:

- Effective learning is demonstrated by the steady decrease in loss over epochs for all window types.
- Hann Window produces the smoothest loss curves, with the least amount of variation between validation and training losses. This suggests low overfitting and good generalization.
- Compared to Hann, Hamming Window exhibits a little quicker convergence and comparable loss behavior. This suggests modest generalization and good optimization.
- Because of the window's inadequate handling of leaks, the Rectangular Window shows larger variations in validation loss, which may indicate more overfitting.

Training and Validation Accuracy:

- Hamming Window achieves the highest validation accuracy across all epochs, closely followed by Hann. This aligns with its balance of resolution and leakage suppression.
- Hann Window achieves competitive accuracy but slightly underperforms compared to Hamming in this specific task.
- Rectangular Window shows the lowest accuracy, attributed to excessive spectral leakage and less smooth features.

Hann Window:

Strengths: Consistent performance throughout training and validation, seamless transitions, and less leakage.

Weakness: Accuracy and convergence are a little slower than with Hamming.

Use Case: Ideal for overlapping frequency components in real-world transmissions.

The Hamming Window:

Strengths: High accuracy and quick convergence due to the best balance between frequency resolution and leakage suppression.

Weakness: Slightly greater loss than Hann, indicating slight energy smoothing trade-offs.

Use Case: Perfect for activities like speech or tone signals that need for precise feature extraction.

Rectangular Window :

Strengths: Strong frequency resolution for hypothetical applications is one of its advantages.

Weakness: Low classification accuracy and poor generalization are caused by high spectral leakage.

Use Case: Restricted to synthetic or isolated signals with little overlap in frequency.

Conclusion:

The Hamming Window is the most appropriate option for the classifier since it offers the best overall performance in terms of accuracy and convergence. The Hann Window, which provides better stability and generalization, comes in close second. Because of its

high leakage, the Rectangular Window performs poorly, highlighting how crucial window selection is for feature extraction in spectrogram-based classification applications.

TASK B

Comparative Analysis of the Spectrograms

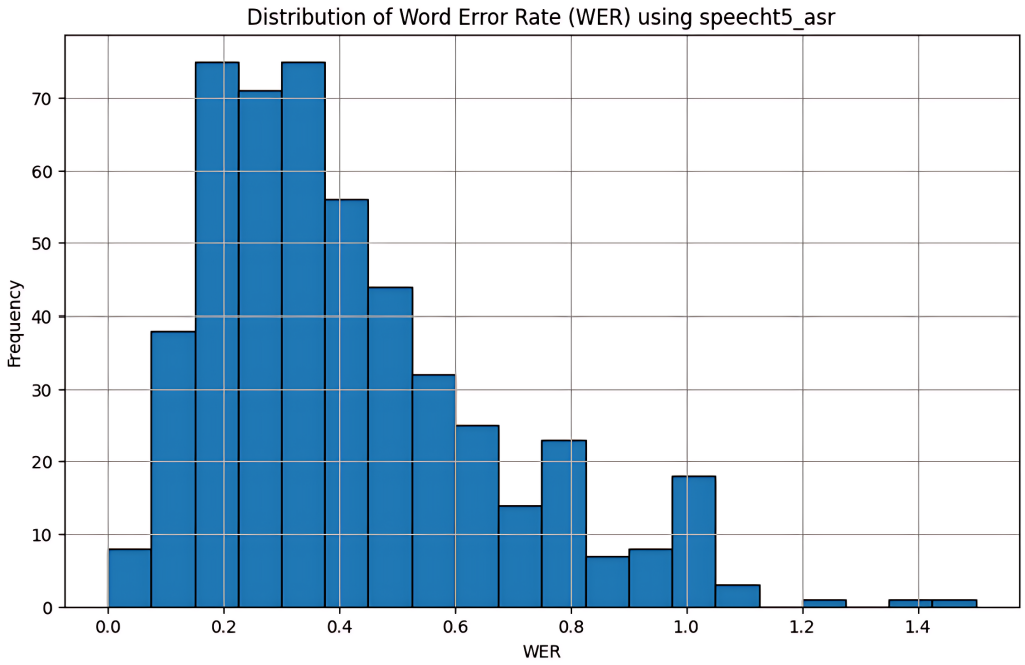


Figure 3: Spectrograms for songs of different genres

Romantic Song

Frequency Range:

- The frequency range of romantic songs is dominated by low-to-mid-frequency elements (0–3000 Hz).
- sparse high-frequency range activity.

Dynamics and Intensity:

- Amplitude changes that occur gradually over time, suggesting dynamics that are softer and more emotive.
- persistent low-amplitude harmonics, which are found in string instruments and romantic tunes.

Temporal Features:

- Even transitions devoid of abrupt stops.
- extended tones that probably indicate slow musical parts or vocals.

Pop Song

Frequency Range:

- Pop songs have a wide frequency range that includes both high and low frequencies (0–8000 Hz).
- bursts of great intensity in the mid- and high-frequency bands.

Dynamics and Intensity:

- recognizable recurring patterns that indicate rhythmic beats at both low and high frequencies.
- Pop music is known for its lively and vivacious amplitude changes.

Temporal Features:

- Frequent peaks in high frequencies, likely from percussion instruments (e.g., cymbals, hi-hats).
- Rhythm-driven energy dominating the spectrogram.

Classical Song

Frequency Range:

- The frequency range of a classical song is broad and evenly distributed across all frequencies.
- There is a lot of harmonic content, which is indicative of orchestral instruments like flutes, violins, and pianos.

Dynamics and Intensity:

- subtle changes in intensity that show how rich and intricate classical compositions are.
- gradual tonal changes that convey a deeper harmonic range and a slower tempo.

Temporal Features:

- Absence of recurring patterns that align with improvisational or through-composed structures.
- Frequencies that are continuous and smooth are a sign of richness in tone.

Party Song

Frequency Range:

- Energy spread throughout the entire spectrum (0–10,000 Hz), with a focus on low and high frequencies, is the party song's frequency range.
- powerful bass lines represented by dominant low-frequency components.

Dynamics and Intensity:

- Periodic, high-intensity bursts across frequencies.
- dynamic and quick changes, which are common in danceable music.

Temporal Features:

- Regular high-frequency sounds, perhaps produced by hi-hats and synthesizers.
- a consistent rhythmic pattern produced by structured beats.

Comparative Analysis of Spectrograms from Four Music Genres

6 Romantic Song

Frequency Range:

- Dominated by low to mid-frequency components (0–3000 Hz).
- Sparse activity in high-frequency ranges.

Intensity and Dynamics:

- Gradual changes in amplitude over time, indicating softer and more emotional dynamics.
- Sustained low-amplitude harmonics, characteristic of romantic melodies and string instruments.

Temporal Features:

- Smooth transitions without sharp bursts.
- Prolonged tones likely representing vocals or slow instrumental passages.

7 Pop Song

Frequency Range:

- Broad range covering both low and high frequencies (0–8000 Hz).
- High-intensity bursts in mid and high-frequency ranges.

Intensity and Dynamics:

- Noticeable repetitive patterns in both low and high frequencies, reflecting rhythmic beats.
- Quick transitions between amplitudes, typical of vibrant and energetic pop music.

Temporal Features:

- Frequent peaks in high frequencies, likely from percussion instruments (e.g., cymbals, hi-hats).
- Rhythm-driven energy dominating the spectrogram.

8 Classical Song

Frequency Range:

- Wide and balanced distribution of energy across all frequencies.
- High harmonic content present, representing orchestral instruments like violins, pianos, and flutes.

Intensity and Dynamics:

- Subtle variations in intensity, indicating the complexity and richness of classical compositions.
- Gradual transitions between tones, reflecting a slower pace and harmonic depth.

Temporal Features:

- Lack of repetitive patterns, consistent with through-composed or improvisational structures.
- Smooth and continuous frequency bands, indicative of tonal richness.

9 Party Song

Frequency Range:

- Energy distributed across the full spectrum (0–10,000 Hz), with emphasis on low and high frequencies.
- Dominant low-frequency components, representing strong bass lines.

Intensity and Dynamics:

- High-intensity, periodic bursts across frequencies.
- Fast-paced and energetic transitions, typical of danceable music.

Temporal Features:

- Consistent high-frequency activity, likely from synthesizers and hi-hats.
- Structured beats creating a regular rhythmic pattern.

Comparative Summary

Aspect	Romantic Song	Pop Song	Classical Song	Party Song
Frequency Range	Dominated by low-mid (0–3000 Hz)	Broad (0–8000 Hz)	Balanced across all frequencies	Full range (0–10,000 Hz), strong low/high
Energy Intensity	Gradual and smooth	Sharp bursts, rhythmic patterns	Subtle and harmonic	High-intensity bursts
Dynamics	Prolonged tones, emotional	Quick transitions, vibrant and energetic	Gradual transitions, tonal complexity	Fast-paced and consistent rhythm
Temporal Features	Smooth transitions, sustained tones	Frequent peaks in high frequencies	Continuous and rich harmonic structure	Periodic patterns, consistent beats

Conclusion:

The distinct acoustic traits of every genre are amply illustrated by the spectrograms:

- Low frequencies and slow transitions are common in romantic songs, which emphasize melody and fluid dynamics.

- Pop songs use repeating rhythms and high-frequency bursts to emphasize rhythm and vitality.
- Smooth harmonic transitions, balanced frequencies, and tonal richness and complexity are all features of classical songs.
- Party songs emphasize rhythm and energy, which are demonstrated by steady beats and sporadic bursts of great intensity.

References

1. https://librosa.org/doc/main/generated/librosa.stft.html
2. https://www.phon.ucl.ac.uk/courses/spsci/acoustics/week1-10.pdf
3. https://scikit-learn.org/1.6/modules/generated/sklearn.svm.SVC.html

Github Repositry Links:

1. hhttps://github.com/Ankit-IITJ/Speech$_U nderstanding_P A1.git$
2. https://github.com/m23csa516/speechunderstanding$_P A_1.git$