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):**

$$WER = \frac{S + D + I}{N} \tag{1}$$

where:

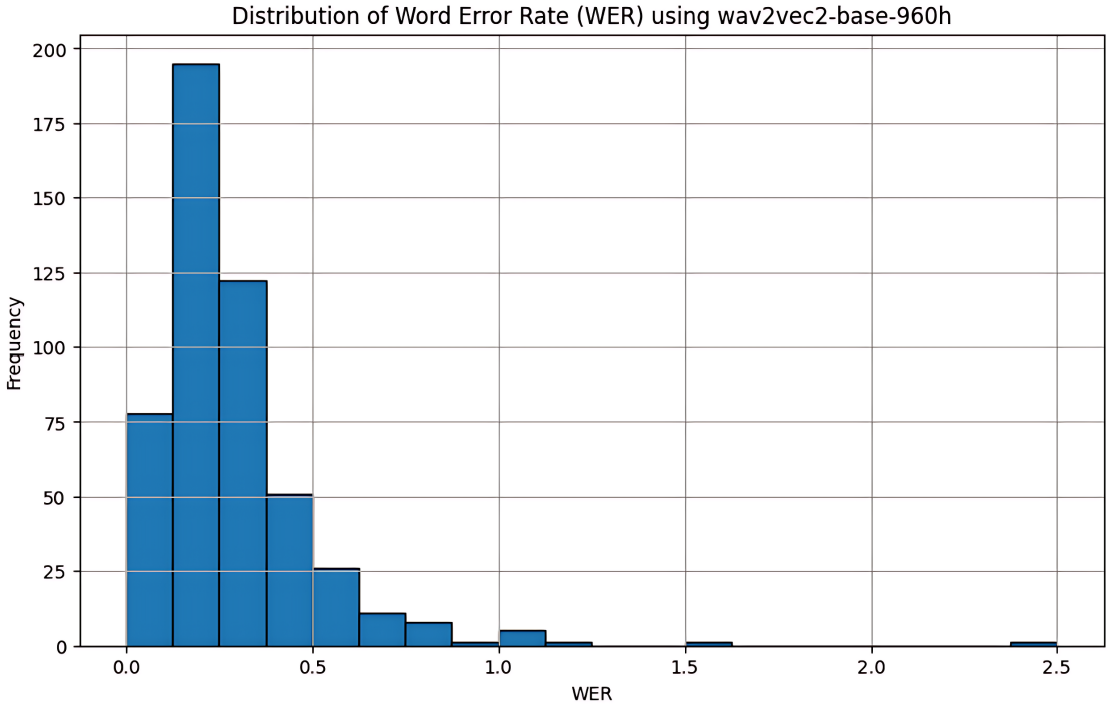
- S = Number of Substitutions
- D = Number of Deletions
- I = Number of Insertions
- N = Total words in the reference

2. **Strengths:** Provides a clear measure of transcription accuracy.

3. **Limitations:** Does not account for semantic understanding or context.

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

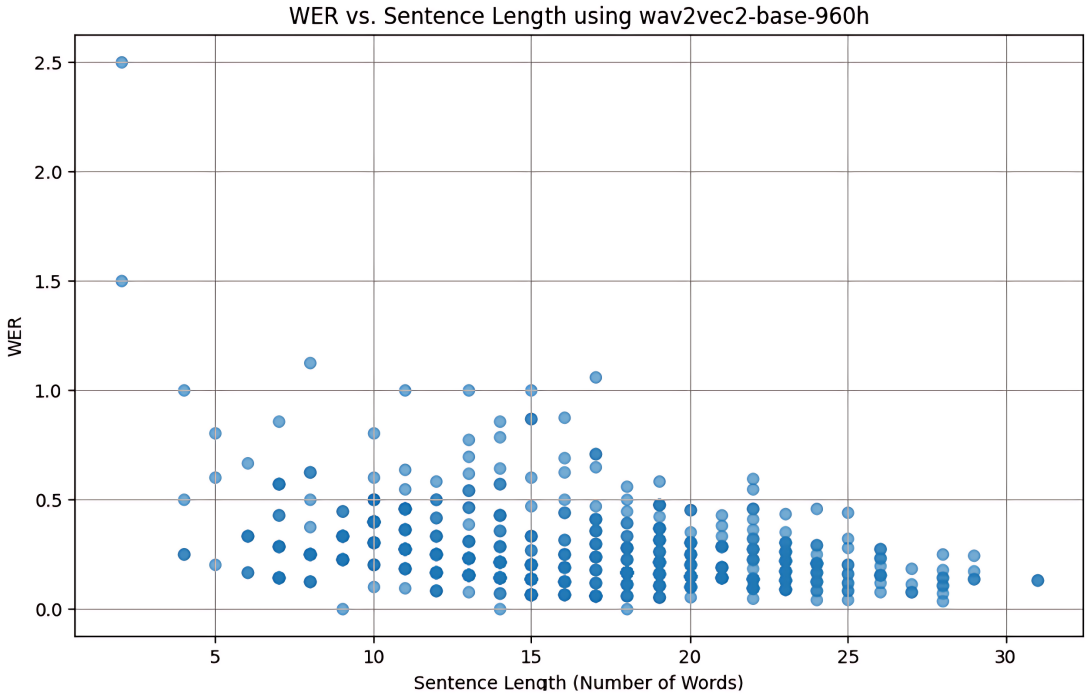
Results



(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)

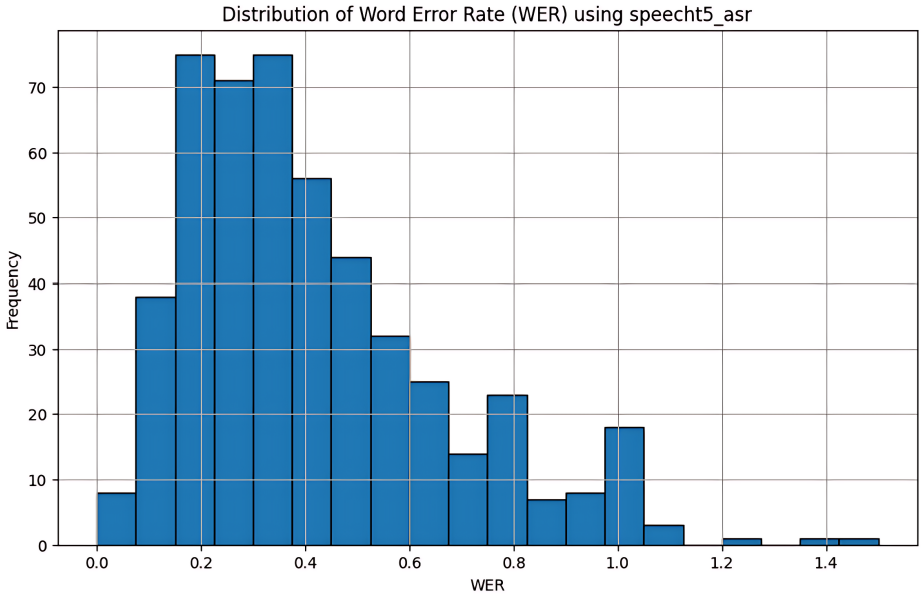
Fig (a)(b): WER Distribution and its Relationship with Sentence Length for Wav2Vec2-Base-960h

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:

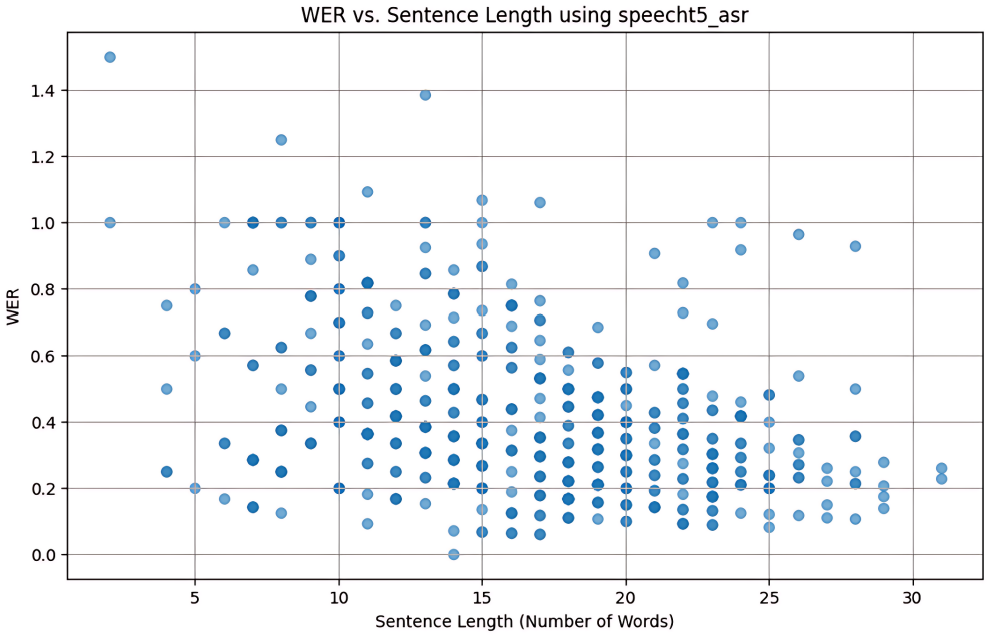
Results



(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)

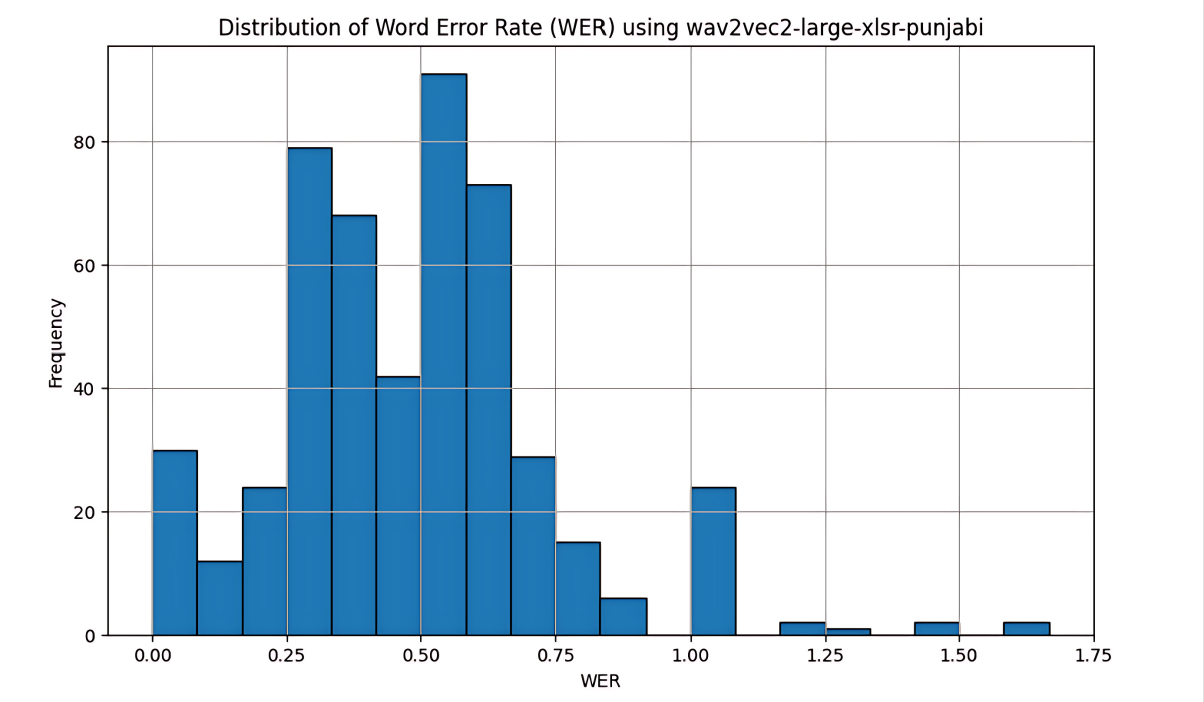
Fig (c)(d): WER Distribution and its Relationship with Sentence Length for speecht5_asr

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:

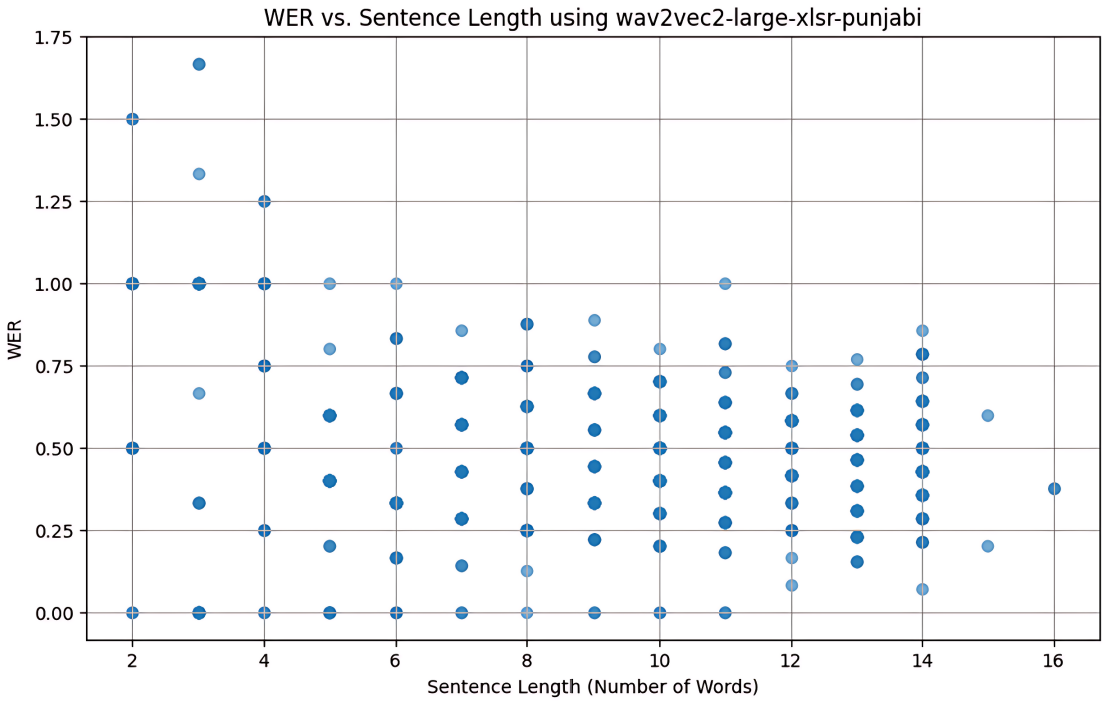
Results



(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)

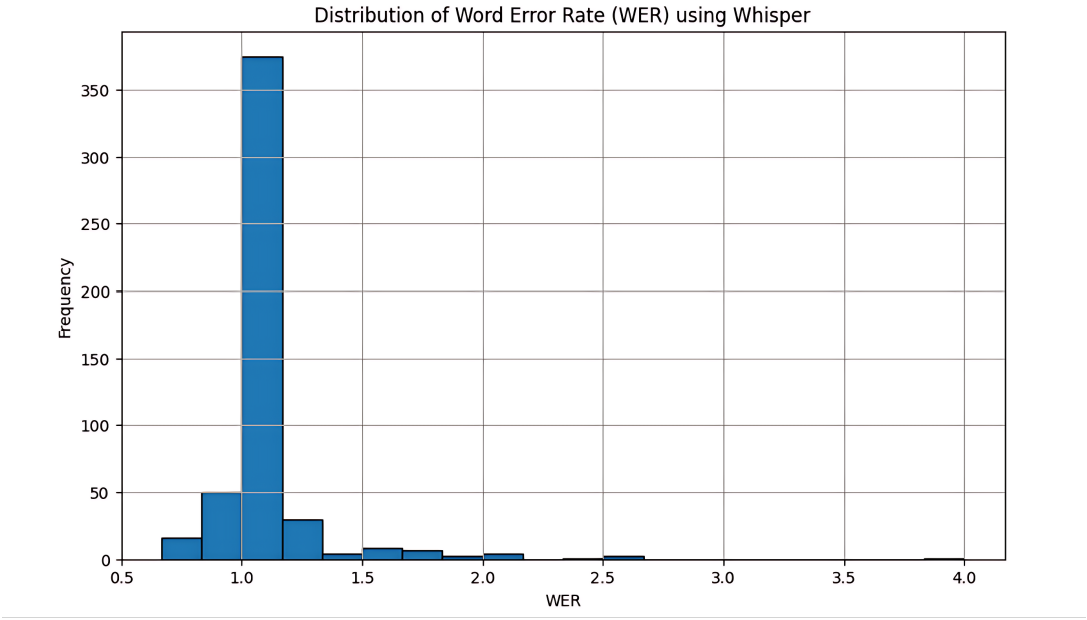
Fig (e)(f): WER Distribution and its Relationship with Sentence Length for wav2vec2-large-xlsr-punjabi

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:

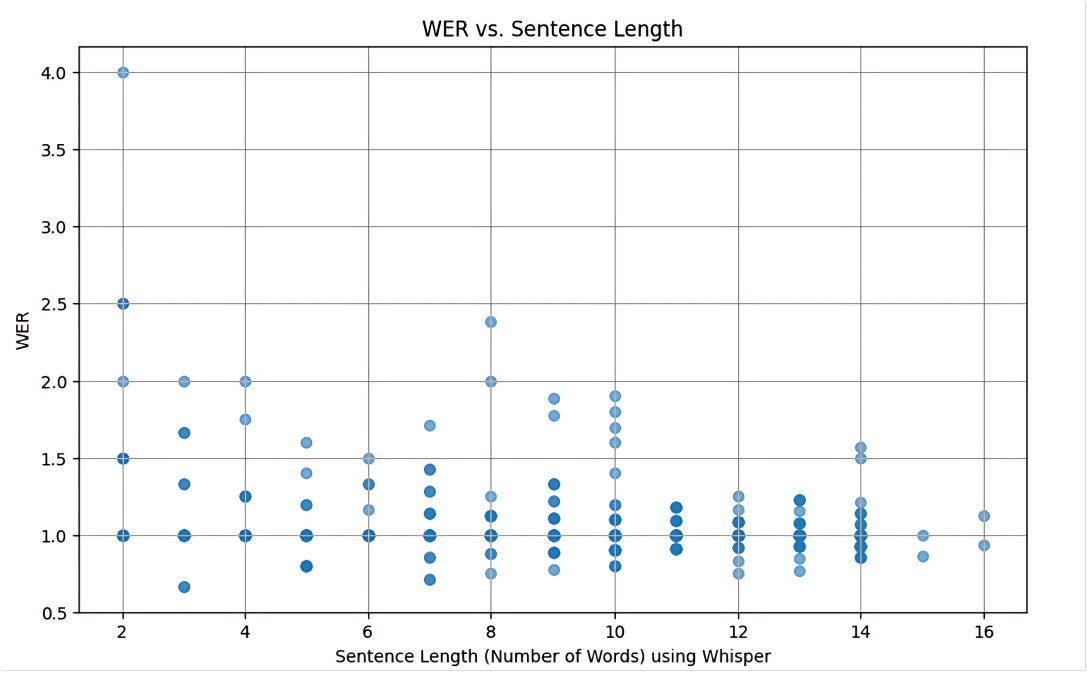
Results



(g)

Distribution of Word Error Rate (WER):

- Most sentences show moderate transcription errors, with WER values clustering around 1.0, while a few outliers exceed 2.0 due to noisy data or complex linguistic patterns.
- The histogram indicates that Whisper maintains consistent performance, though errors are skewed toward moderate levels.



(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

Github Repositry Links:

1. https://github.com/m23csa516/speechunderstanding$_PA_1.git$
2. hhttps://github.com/Ankit-IITJ/Speech$_Understanding_PA1.git$