Speech-to-Text

A Comprehensive Analysis of State-of-the-Art Models, Challenges, and Future Directions

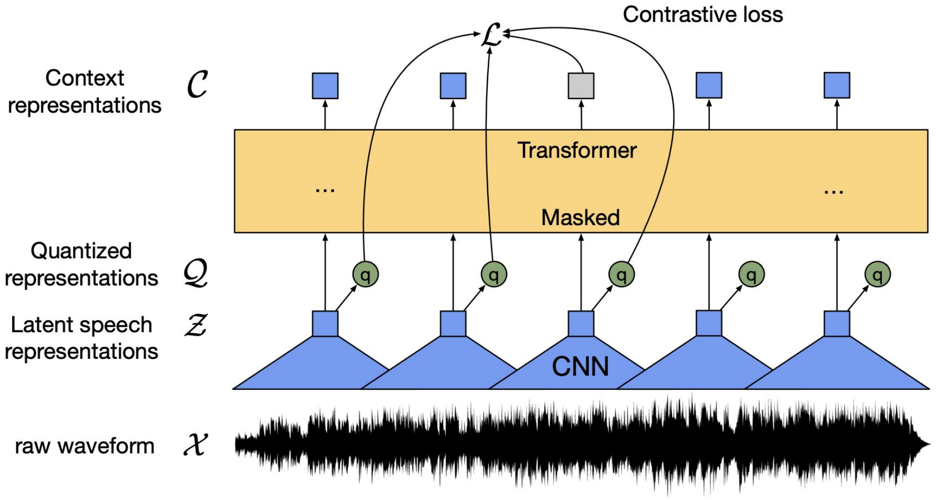
Presented By-
Himani (M23CSA516)
Ankit Kumar Chauhan (M23CSA509)

State-of-the-Art (SOTA) Models

Language	Model	Dataset
English	Wav2Vec2-Base-960h	LJSpeech sr16k
English	SpeechT5-ASR	
Punjabi	Wav2Vec2-Large-XLSR-Punjabi	Punjabi Speech
Punjabi	Whisper	

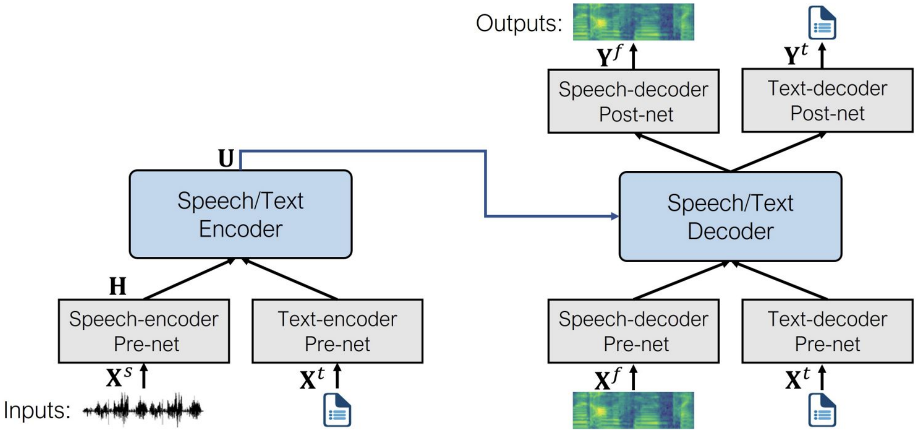
Speech-to-Text: How wav2vec 2.0 Works? 🎙️ ➡️📜

- **Step 1: Input - Raw Audio**
- **Step 2: Feature Extraction (CNN)**
- **Step 3: Self-Supervised Learning (Masked Audio Model)**
- **Step 4: Decoding (Speech-to-Text Conversion)**
- **Step 5: Language Model Correction**



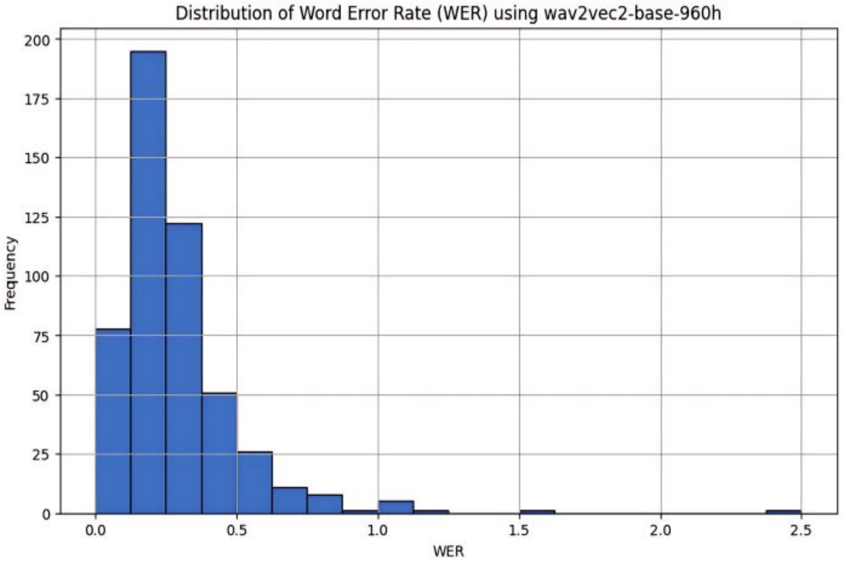
Paper Link: https://arxiv.org/abs/2006.11477

- **Step 1: Input - Raw Speech Signal**
- **Step 2: Feature Extraction (Speech Pre-Net)**
- **Step 3: Encoding (Transformer Encoder)**
- **Step 4: Speech-to-Text Mapping (Cross-Modal Alignment)**
- **Step 5: Decoding (Transformer Decoder)**

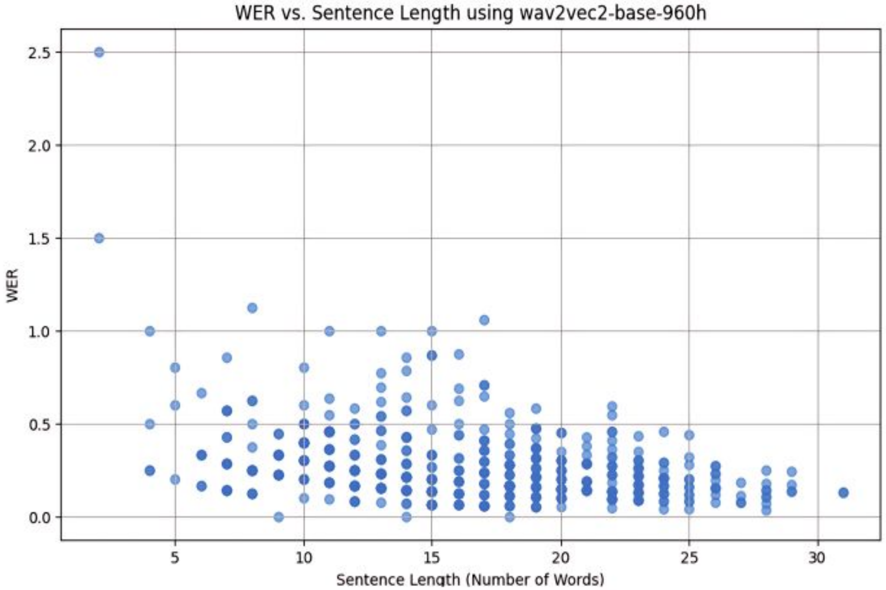


Paper Link: https://arxiv.org/pdf/2110.07205

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



(a)



(b)

Metrics Used:

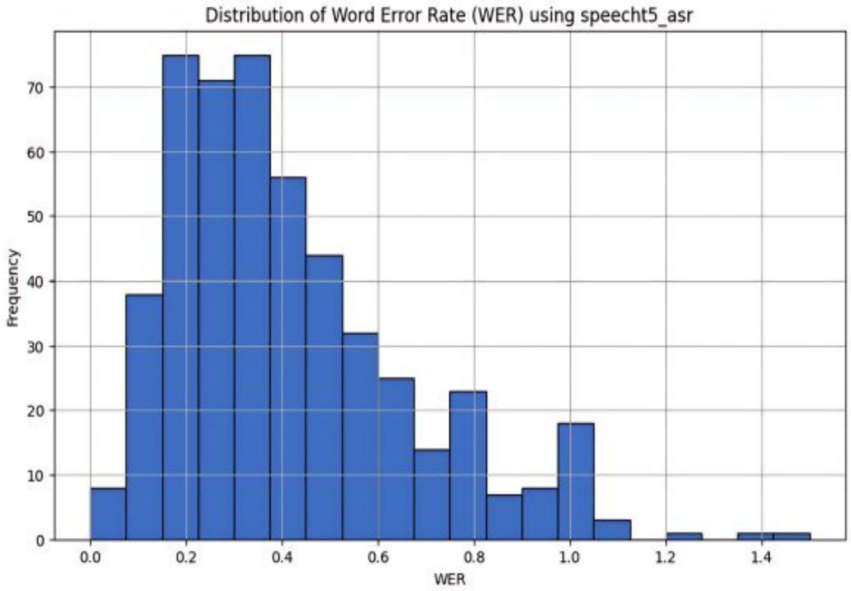
1. **Word Error Rate (WER):**

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

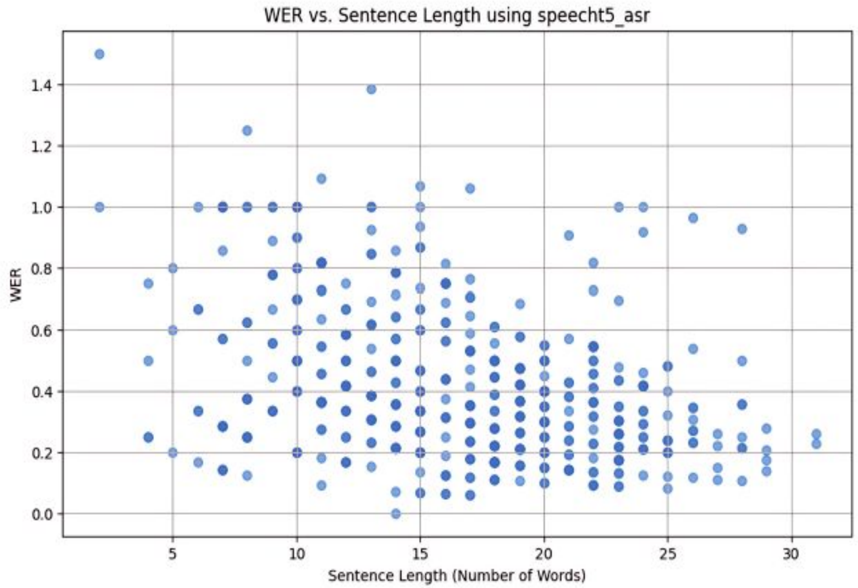
where:

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

Analysis for High Resource language: English using speecht5 asr:

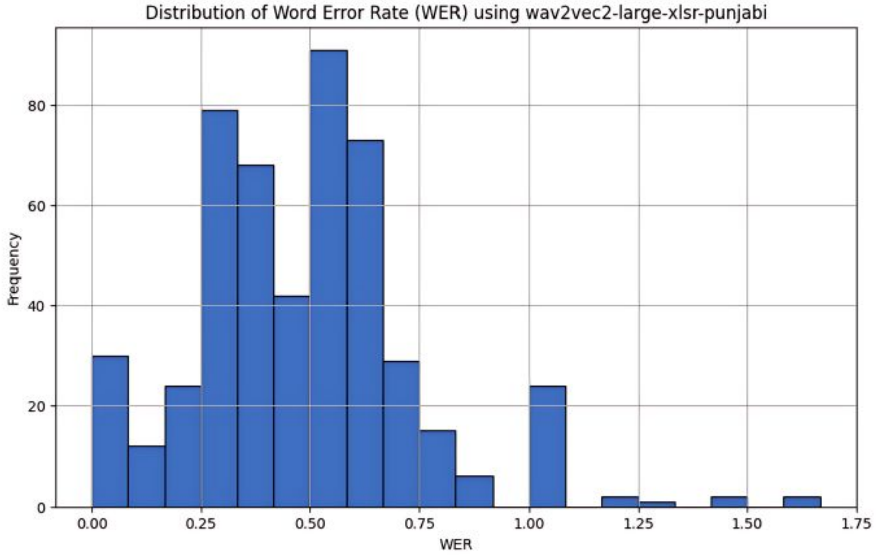


(c)

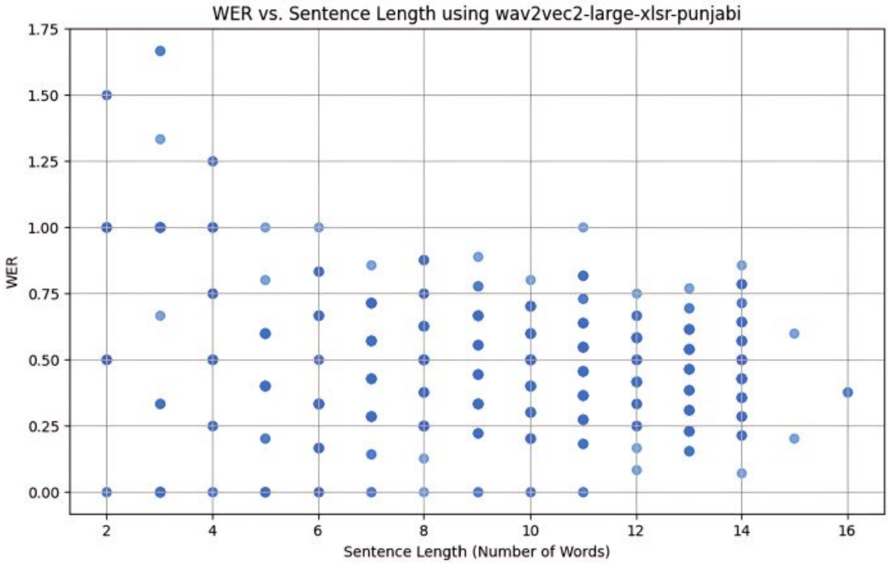


(d)

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

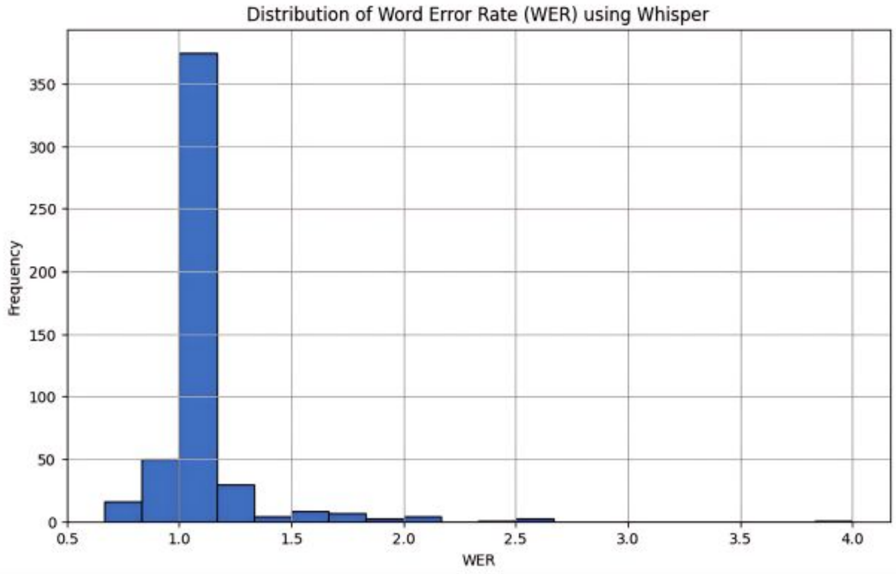


(e)

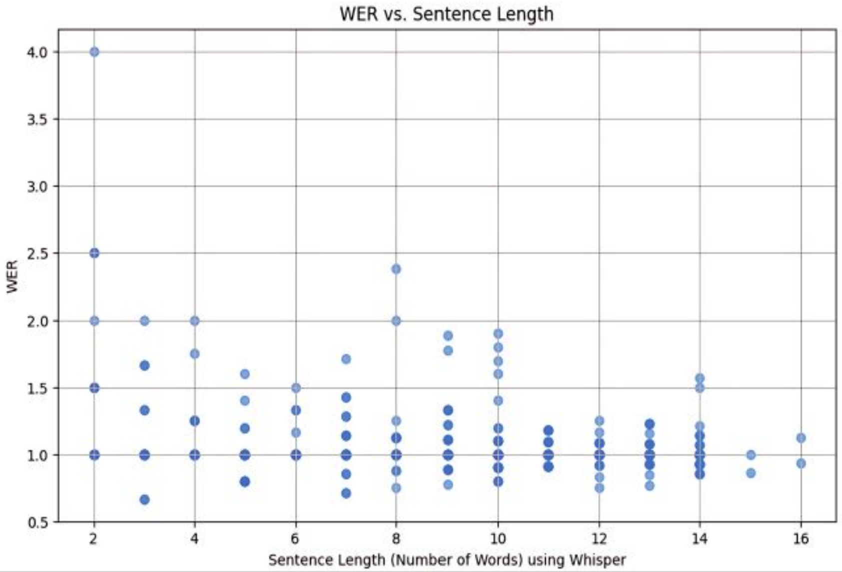


(f)

Analysis for low Resource language: Punjabi using whisper:



(g)



(h)

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.

Conclusion & Future Directions

Summary

- Speech-to-Text technology is **essential for automation and accessibility**.
- **High-resource** languages like English have **better models, but require significant computing power**.
- **Low-resource languages** like Punjabi still face **challenges**, particularly in **dataset availability and model robustness**.

Future Improvements

- **Enhancing noise tolerance** to make models more robust.
- **Expanding multilingual capabilities** to support a broader range of languages

Thank You