Speech-to-Text System Performance Insights

Visualizing the capabilities and challenges of the STT system based on comprehensive testing.

Overall Performance Snapshot

Best Language (General Speech)

English-US

WRR: 70.28%

Best Number Accuracy

English-US

96.88%

Average Latency

1.93s

For 5-10s clips

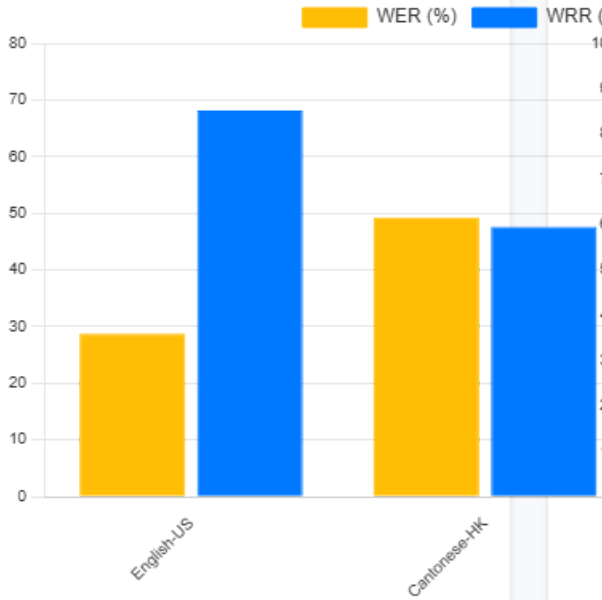
Key Metrics Explained

- **Word Error Rate (WER):** Percentage of words incorrectly predicted. Lower is better.
- **Word Recognition Rate (WRR):** Percentage of words correctly transcribed. Higher is better. (Often 100% - WER)
- **Number Sequence Accuracy:** Percentage of correctly identified numbers in a sequence.
- **Vocabulary Accuracy:** Percentage of correctly recognized domain-specific terms.
- **Segmentation Accuracy:** Correctness of auto-punctuation for sentence separation.

Multilingual Performance Deep Dive

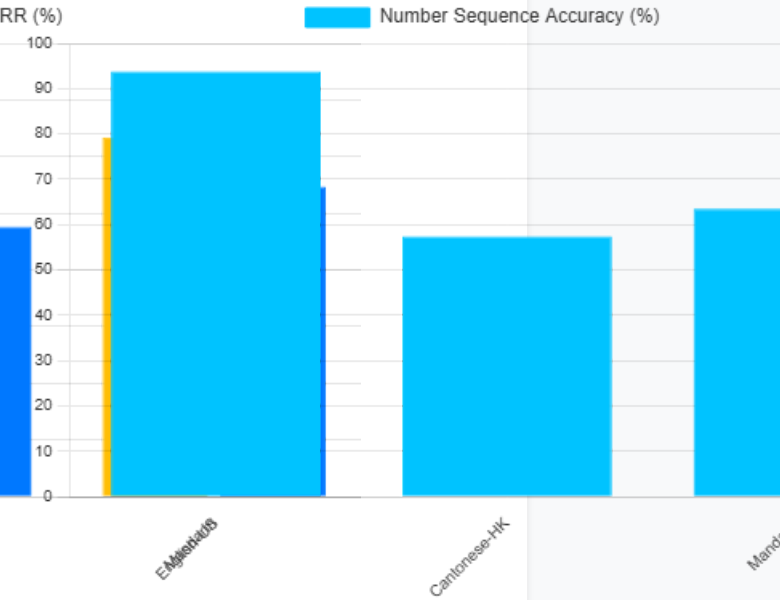
Comparing the STT system's effectiveness across English-US, Cantonese-HK, and Mandarin in clean audio conditions.

WER & WRR by Language (General Speech)



English-US shows the lowest Word Error Rate and highest Word Recognition Rate.

Number Sequence Accuracy by Language

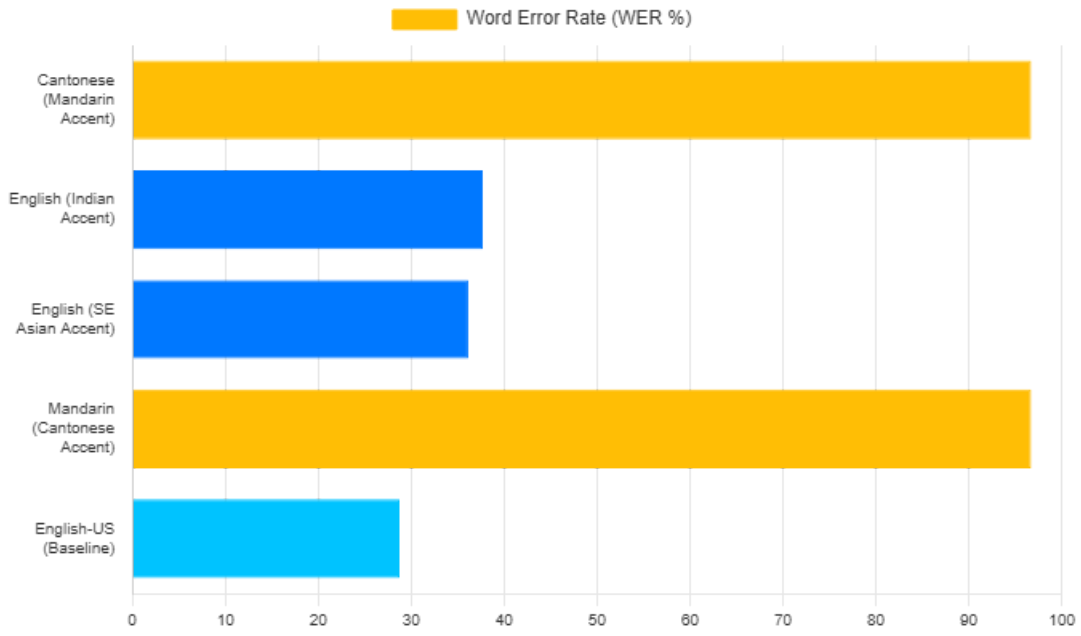


English-US demonstrates significantly higher accuracy in transcribing numerical sequences.

The Accent Challenge

Evaluating the system's ability to understand speech with various accents. Lower WER indicates better performance.

Word Error Rate (WER) Across Accents

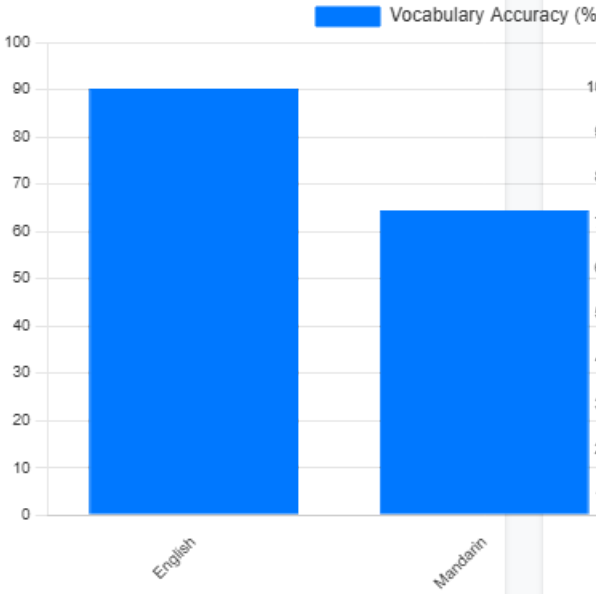


The system struggles significantly with Cantonese/Mandarin mixed accents, resulting in 100% WER. English with Indian and Southeast Asian accents also shows higher error rates compared to standard English-US.

Vocabulary and Punctuation Accuracy

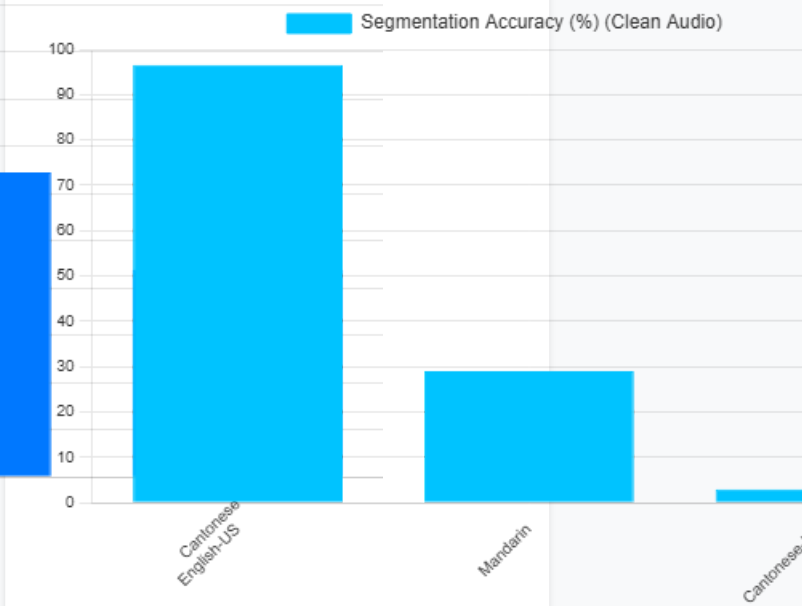
Assessing the recognition of specific domain terms and the correctness of automatic punctuation in clean audio.

Domain Vocabulary Accuracy



English leads in recognizing domain-specific vocabulary, while Cantonese shows the lowest accuracy.

Auto Punctuation Accuracy (Clean Audio)

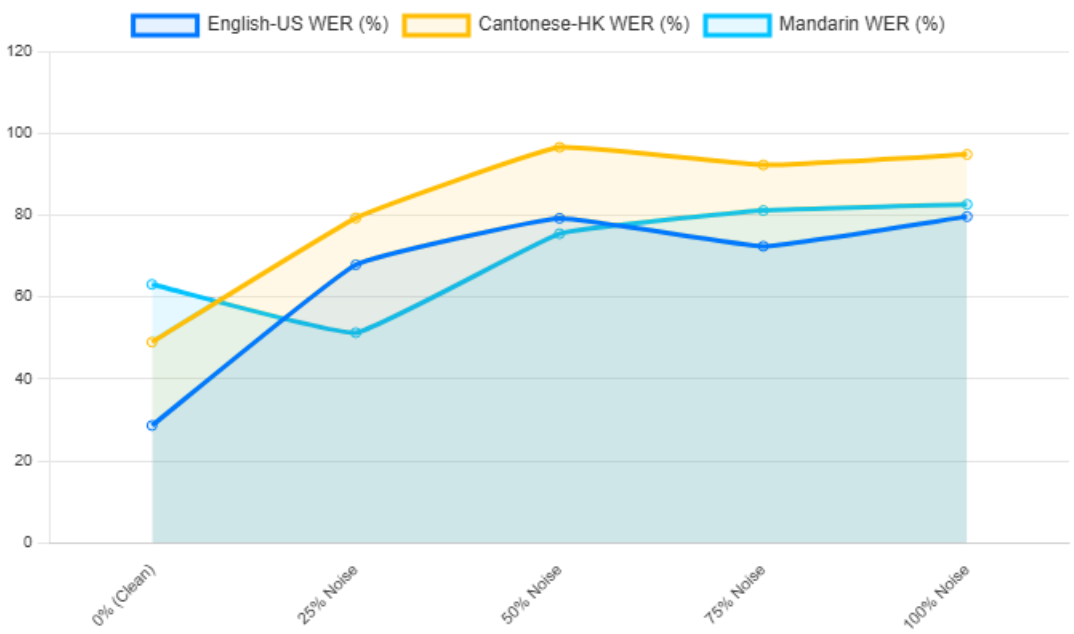


Auto-punctuation is highly effective for English-US, but performs poorly for Mandarin and is largely ineffective for Cantonese.

Impact of Noise on Performance

Visualizing how Word Error Rate (WER) for general speech increases with different levels of background noise. Lower WER is better.

WER vs. Noise Level (General Speech)



Performance degrades for all languages as noise increases. Cantonese-HK is most affected, while English-US shows greater resilience, though still impacted significantly at high noise levels.

Summary: Strengths & Weaknesses

Strengths ✔

- Strongest performance for **English-US** (general speech, numbers, vocabulary).
- Excellent auto-punctuation for English-US in clean conditions.
- Acceptable average latency (1.93s for 5-10s clips).

Weaknesses ❌

- Significantly lower performance for **Cantonese-HK and Mandarin**.
- Highly susceptible to **accents**, especially mixed Asian language accents.
- **Noise robustness** is a major concern across all languages.
- Auto-punctuation largely ineffective for Cantonese and poor for Mandarin.
- Weaker domain vocabulary recognition for Cantonese and Mandarin.
- No data available for profanity filtering assessment.

Conclusion & Recommendations

The STT system shows promise, particularly for English-US, but requires significant improvements in several key areas to be robustly effective across diverse languages, accents, and environments.

Key Recommendations:

- Invest in further model training and fine-tuning for Cantonese and Mandarin.
- Improve robustness against a wider range of accents, focusing on those with current high error rates.
- Enhance noise cancellation and reduction techniques.
- Develop and improve auto-punctuation features for Cantonese and Mandarin.
- Expand training data for domain-specific vocabularies in non-English languages.