STT System Performance Insights

An Analysis of the 'Google' Method Based on Recent Evaluations

Understanding the Landscape

This infographic visualizes the performance of the 'Google' Speech-to-Text (STT) system across various critical benchmarks, including multilingual support, accent robustness, handling of domain-specific vocabulary, and performance under noisy conditions. The insights are drawn from the "STT System Performance Analysis Report (Google Method)" to provide a clear view of the system's capabilities and areas for improvement.

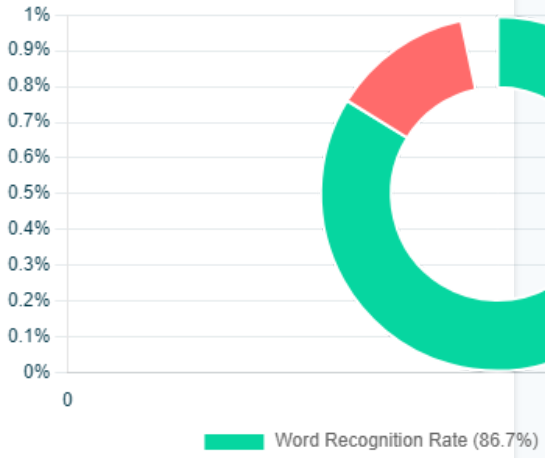
Overall Performance Snapshot

A high-level comparison reveals significant performance disparities across languages. English-US shows strong results, while Cantonese-HK and Mandarin face greater challenges.

English-US: Strong Performance

86.7% Avg. Word Recognition Rate (WRR)

13.3% Avg. Word Error Rate (WER)

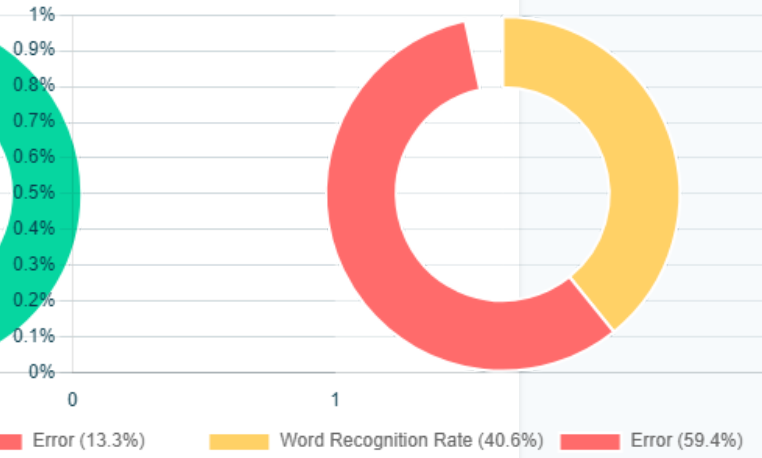


English-US demonstrates high accuracy in clean audio conditions, making it reliable for standard transcription tasks.

Cantonese-HK: Significant Challenges

40.6% Avg. Word Recognition Rate (WRR)

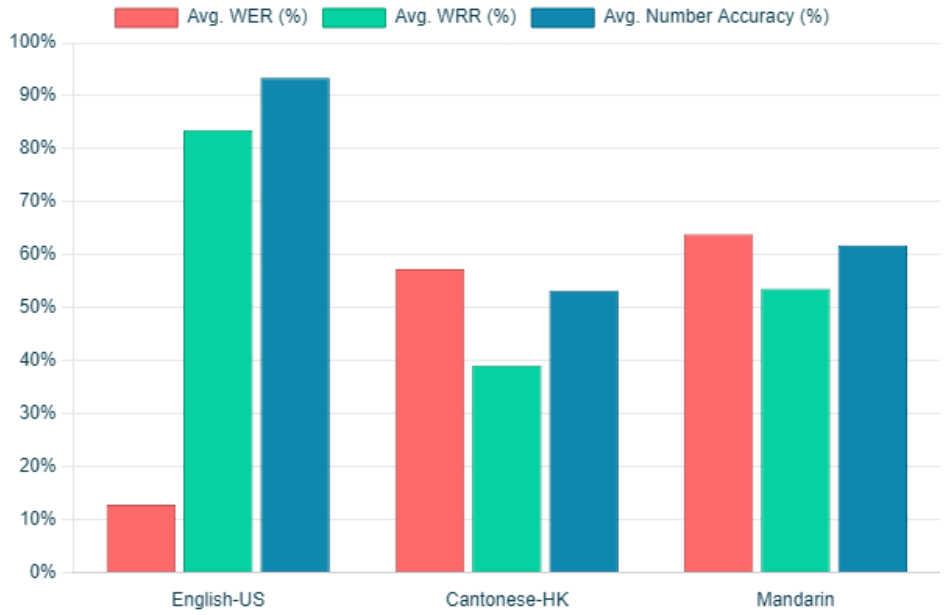
59.4% Avg. Word Error Rate (WER)



Cantonese-HK transcriptions exhibit a considerably higher error rate, indicating a need for model improvement for this language.

Language Performance Deep Dive

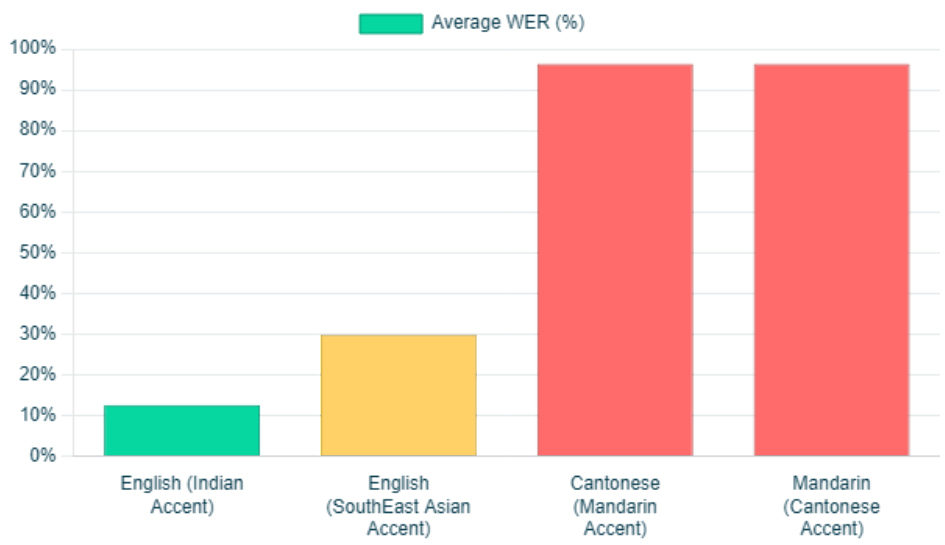
Examining Word Error Rate (WER), Word Recognition Rate (WRR), and Number Sequence Accuracy reveals varying levels of proficiency across the primary languages tested. English-US consistently outperforms Cantonese-HK and Mandarin.



English-US: Shows low WER (13.32%), high WRR (86.68%), and excellent Number Accuracy (96.88%).
Cantonese-HK: Faces challenges with a high WER (59.42%), lower WRR (40.58%), and moderate Number Accuracy (55.21%). For example, in log `SbduiuP_gek_noisy_0_SbduiuP_gek.whisper.auto.gt_audio_140.vtt`, the Cantonese prediction was entirely incorrect (100% WER).
Mandarin: Also struggles with a high WER (66.25%), moderate WRR (55.62%), and moderate Number Accuracy (64.06%). Log `7bLFdBennvI_noisy_0_7bLFdBennvI.whisper.auto.gt_audio_153.vtt` showed a complete failure ("No speech recognized").

Accent Robustness

The system's ability to handle different accents varies significantly. While it performs well with Indian English, cross-language accents (e.g., Cantonese with a Mandarin accent) pose a major hurdle, often resulting in complete transcription failure (100% WER).

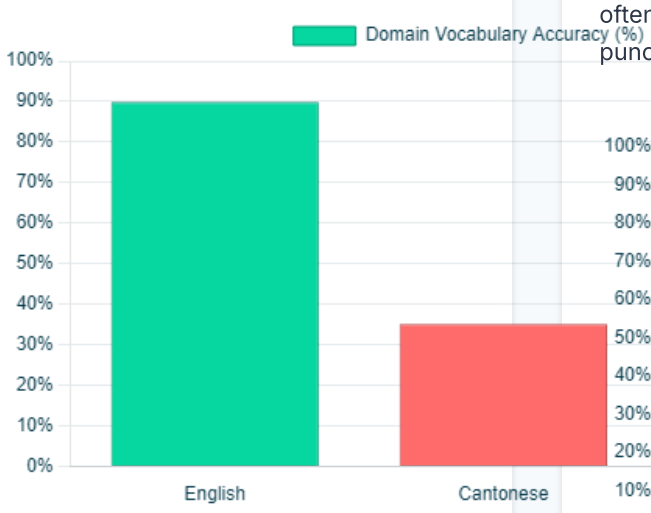


English (Indian Accent): Performs well with a low WER of 13.13%.
English (South East Asian Accent): Shows moderate performance with a WER of 31.05%. Some colloquialisms, like in log `1AnPurpl81c_noisy_0_1AnPurpl81c.whisper.auto.gt_audio_15.vtt`, were poorly transcribed.
Cantonese (Mandarin Accent) & Mandarin (Cantonese Accent): Both resulted in 100% WER, indicating a critical lack of robustness for these specific accent combinations.

Specialized Capabilities

Domain Vocabulary Accuracy

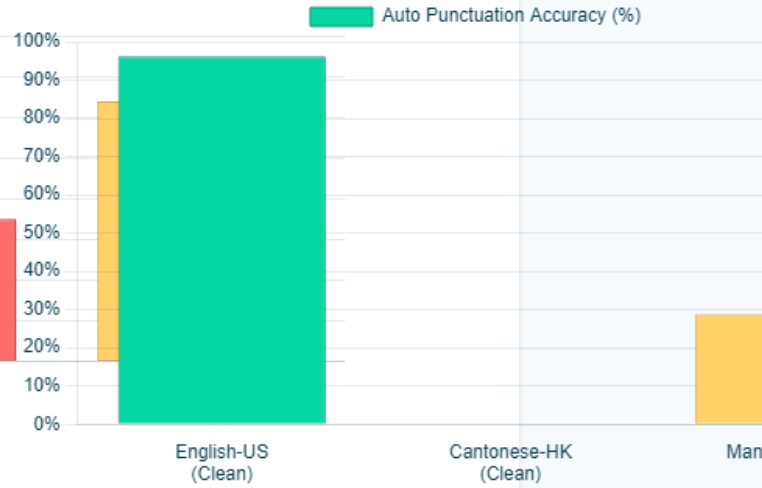
Recognition of HSBC-specific terms is strong for English (93.33%) but significantly lower for Cantonese (36.53%) and moderate for Mandarin (66.67%).



For instance, while "HSBC" was correctly identified in English (log `RwImeANAQFQ_noisy_0_RwImeANAQFQ.whisper.
Cantonese missed terms like "八達通" (log `FmpXfFwzvP8_noisy_0_FmpXfFwzvP8.whisper.au

Auto Punctuation Accuracy (Clean Audio)

Auto-punctuation is highly effective for English-US (100%) in clean conditions. However, it performs poorly for Cantonese (0%) and Mandarin (30%), often failing to insert necessary punctuation.

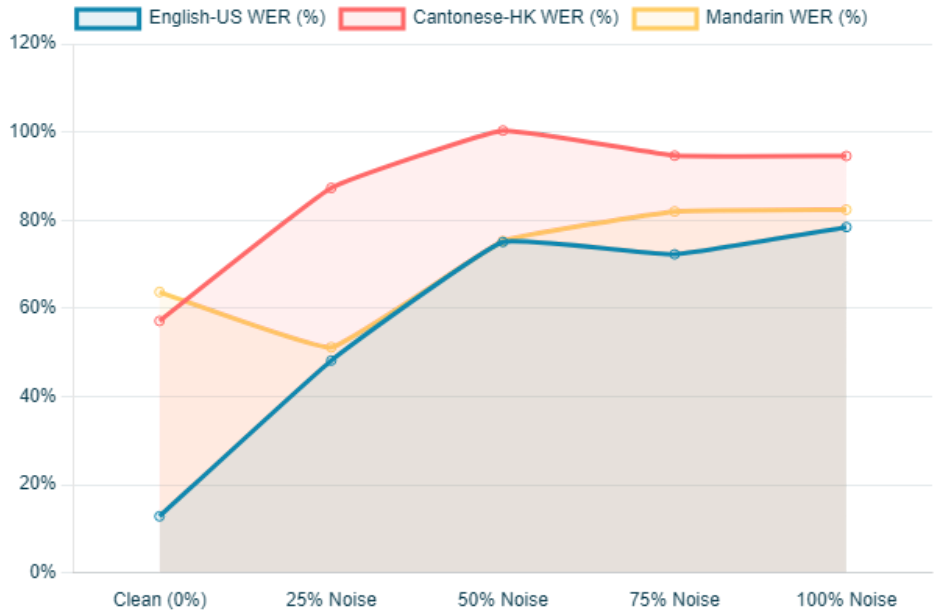


English benefits from accurate sentence segmentation, whereas Cantonese transcriptions often lack any punctuation

(e.g., log `TC-4,TC-1,Cantonese-HK,,,4FEN95kaPlo_noisy_0_4FEN95kaPlo.whisper.auto.gt_audio_17.v

The Impact of Noise on Performance

Background noise severely degrades transcription accuracy across all languages. The charts below illustrate the sharp increase in Word Error Rate (WER) as noise levels rise from 0% (clean) to 100%.



Even at 25% noise, English-US WER jumps to 50.08%. Cantonese-HK and Mandarin fare much worse, with WERs exceeding 90% and 50% respectively. At 100% noise, accurate transcription becomes nearly impossible for all languages, often resulting in "No speech recognized" (e.g., English-US log `TC-7,,English-US\noisy_100,NoiseTest,100%,1AnPurpl81c_noisy_100_1AnPurpl81c.whisper.auto.gt_audio_2.vtt`).

Transcription Speed & Latency

1.93s

Average Actual Latency for 5-10s Audio Clips

The average processing time is generally acceptable for non-real-time applications, being roughly 0.2x to 0.4x the audio's real time length for the tested clips.

Key Takeaways & Recommendations

Strengths

- ✓ Excellent English-US performance in clean audio.
- ✓ Good robustness to English (Indian Accent).
- ✓ Acceptable latency for non-real-time tasks.
- ✓ Strong domain vocabulary recognition for English.
- ✓ Perfect auto-punctuation for clean English-US audio.

Areas for Improvement

- ✗ Poor performance for Cantonese-HK and Mandarin.
- ✗ Critical failure with cross-language accents.
- ✗ Severe degradation in noisy conditions for all languages.
- ✗ Unreliable auto-punctuation for non-English languages and noisy audio.
- ✗ Lower domain vocabulary accuracy for non-English languages.

Recommendations

- Enhance acoustic and language models for Cantonese-HK and Mandarin.
- Improve robustness to diverse accents, especially cross-language ones.
- Strengthen noise reduction and speech enhancement techniques.
- Refine auto-punctuation models for non-English languages and noisy audio.
- Investigate and clarify any metric calculation discrepancies.
- Ensure test data availability for features like profanity filtering if required.