Speech-to-Text (STT) Performance Deep Dive

Analyzing Google's STT Method: Key Insights & Trends

At a Glance: Performance Overview

The Google STT method shows varied performance across languages and conditions. While strong for standard US English in clean environments, significant challenges arise with other languages like Cantonese and Mandarin, especially when dealing with accents, specific terminology, and background noise.

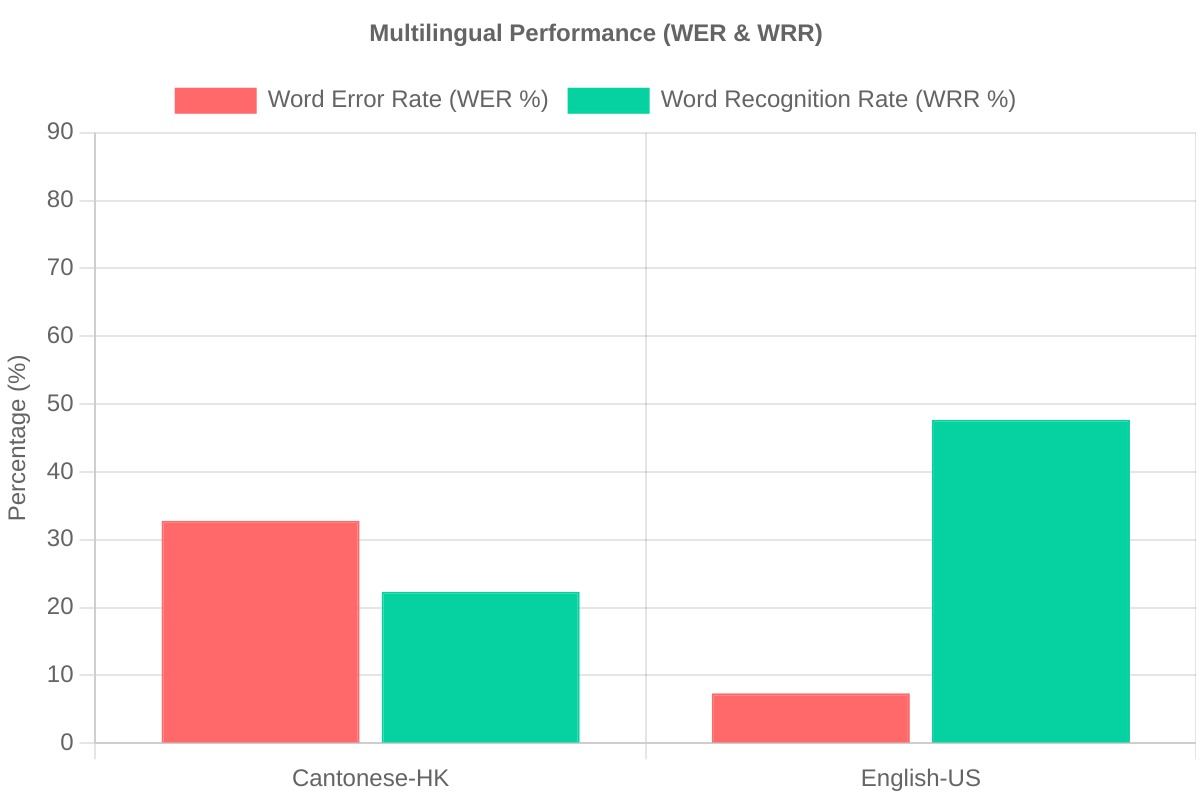
86.68%
WRR for English-US (Clean)
Strong baseline performance.

~60-66%
Avg. WER for Cantonese & Mandarin
Indicating significant transcription challenges.

0%
Punctuation Accuracy (Cantonese, Clean)
A major area for improvement.

TC-1: Multilingual Support

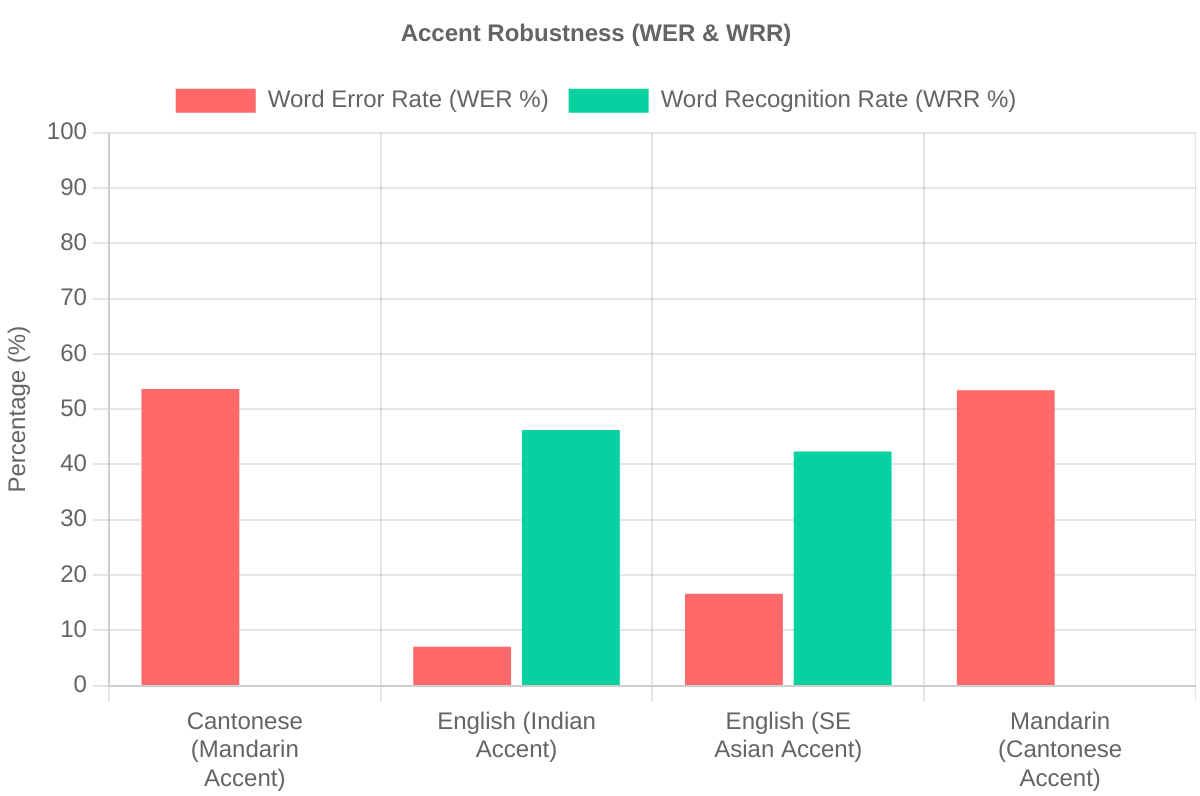
This test case evaluates the STT system's core transcription accuracy across different languages in relatively clean audio conditions. The focus is on Word Error Rate (WER) and Word Recognition Rate (WRR).



The chart clearly illustrates a significant performance disparity. English-US achieves a high Word Recognition Rate (WRR) and correspondingly low Word Error Rate (WER). In contrast, Cantonese-HK shows substantially lower WRR and higher WER, indicating greater difficulty in accurate transcription for this language by the Google STT method.

TC-2: Robustness Across Accents

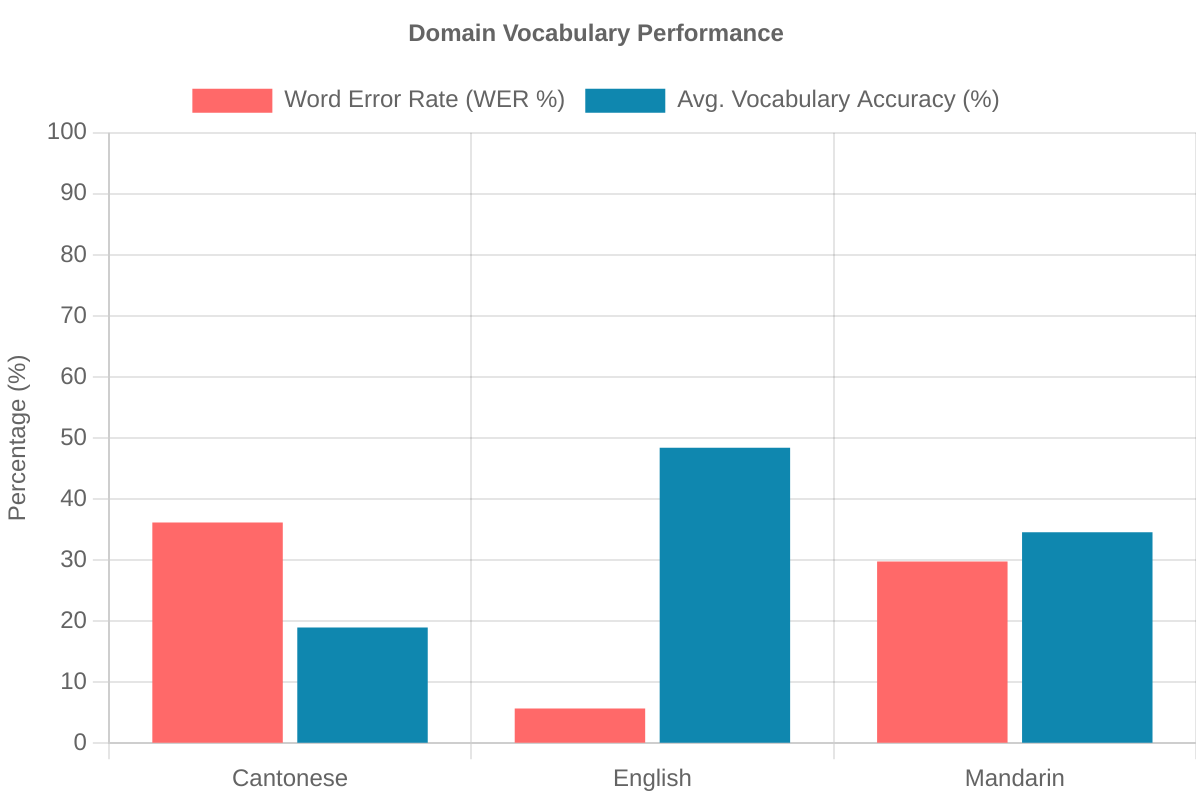
This section examines how the STT system performs when transcribing speech with various accents. High Word Error Rates (WER) indicate difficulty in understanding accented speech.



The Google STT method demonstrates strong performance for English with an Indian accent, achieving accuracy comparable to standard English-US. However, it struggles significantly with English (South East Asian accent), showing a higher WER. For Cantonese (Mandarin accent) and Mandarin (Cantonese accent), the system fails almost entirely, with WER reaching 100%, highlighting a critical weakness in handling these specific cross-language accent scenarios.

TC-3: Domain Vocabulary Support

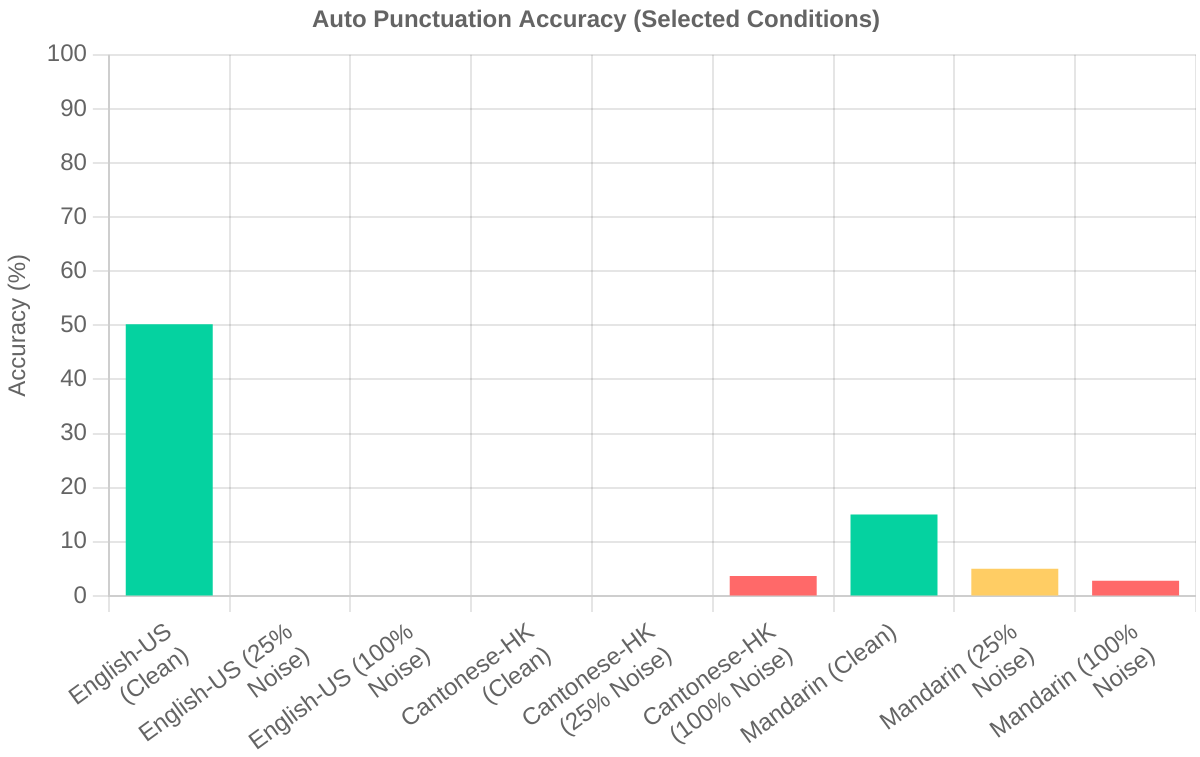
This test assesses the system's ability to recognize and transcribe domain-specific terminology, in this case, HSBC-related terms. Both Word Error Rate (WER) and Vocabulary Accuracy are key metrics.



The system shows excellent domain vocabulary support for English, with very high vocabulary accuracy and low WER. However, for Cantonese and Mandarin, the recognition of specific terms is significantly poorer, with vocabulary accuracy for Cantonese being particularly low. This suggests that while the general English model is robust for specialized terms, the non-English models require more adaptation for specific vocabularies.

TC-4: Auto Punctuation Accuracy

This section evaluates the STT system's capability to automatically insert correct punctuation. The chart displays Average Segmentation Accuracy for selected key language and noise conditions.

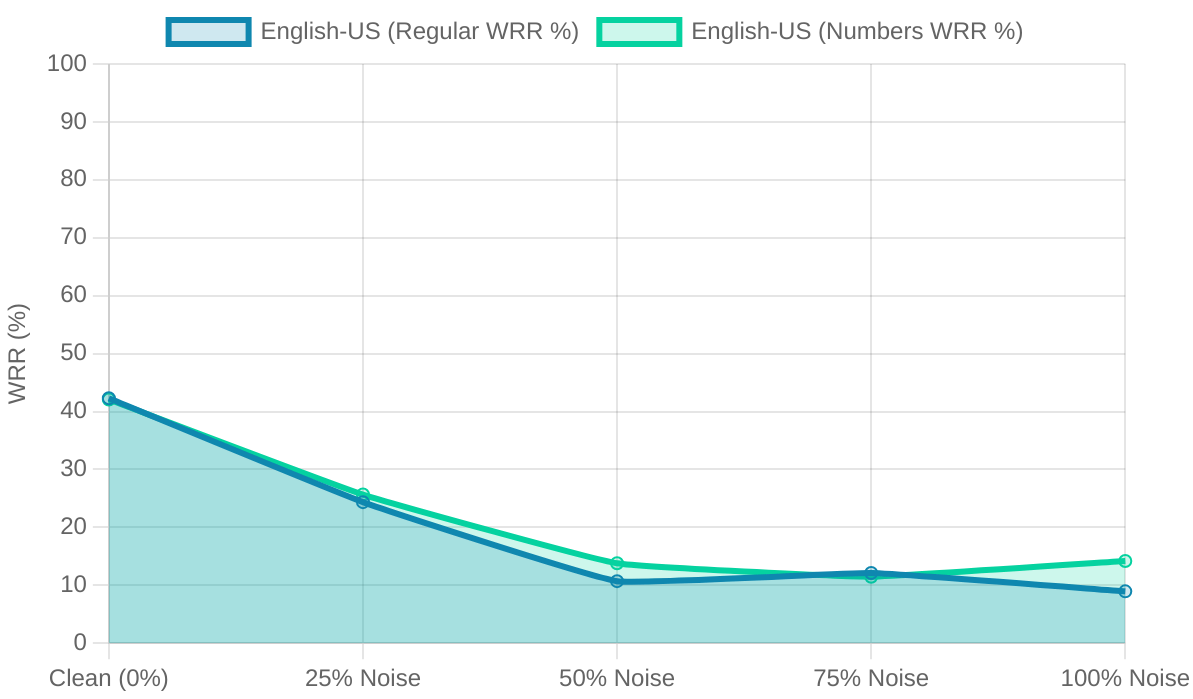


Auto-punctuation for English-US is perfect in clean audio but degrades significantly as noise increases. For Cantonese-HK and Mandarin, punctuation accuracy is extremely low even in clean conditions and worsens with noise. This indicates a major area for improvement, especially for non-English languages and noisy environments.

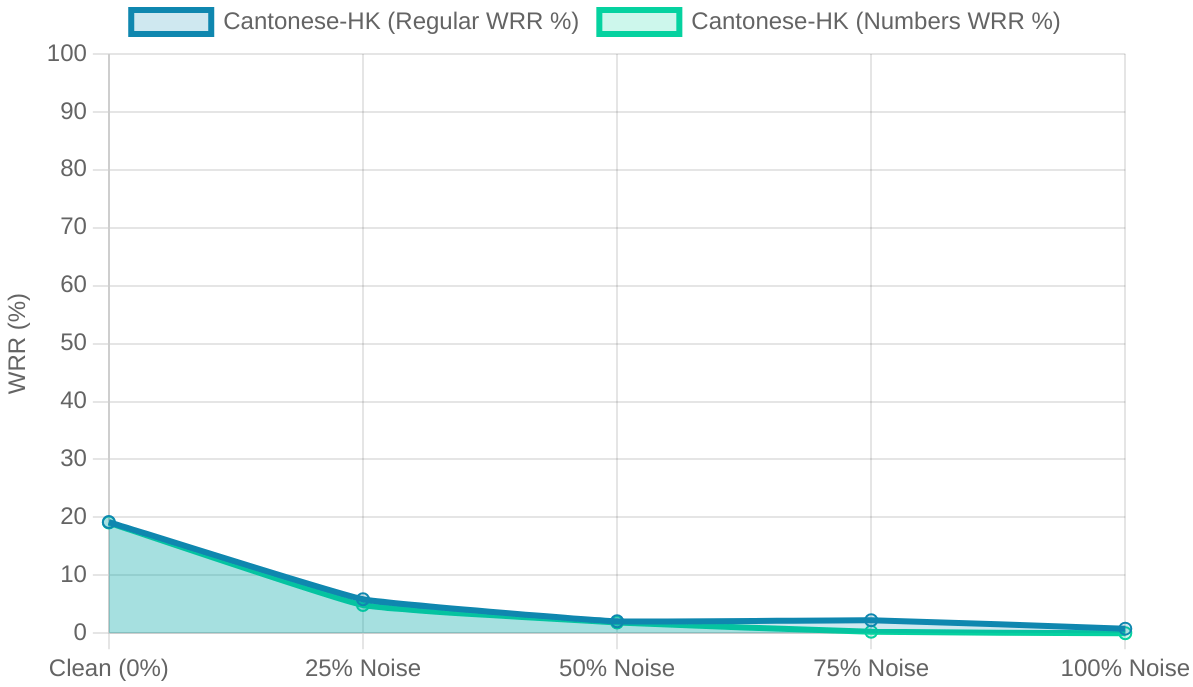
TC-7: Noise Robustness

This test measures the STT system's performance (Word Recognition Rate - WRR) when transcribing audio mixed with various levels of background noise. "Clean" represents 0% noise baseline from TC-1/TC-3 data.

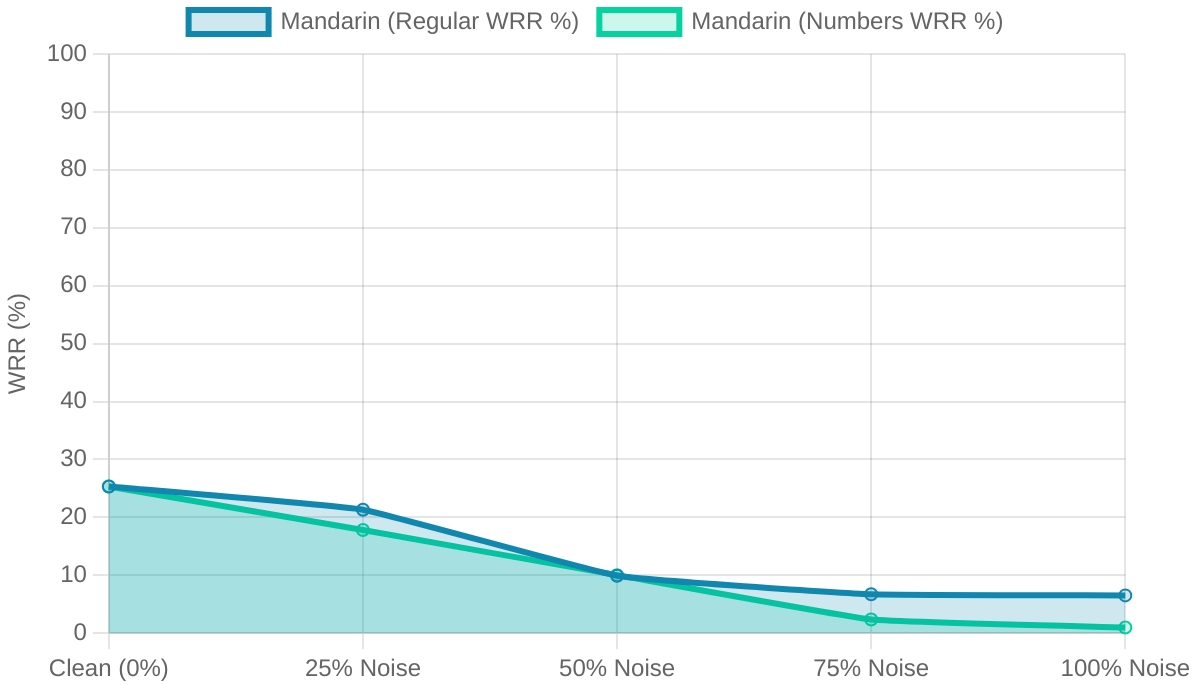
English-US Noise Robustness (WRR %)



Cantonese-HK Noise Robustness (WRR %)



Mandarin Noise Robustness (WRR %)



Noise significantly impacts transcription accuracy across all languages. English-US maintains better WRR at lower noise levels but still sees a sharp decline as noise increases. Cantonese-HK and Mandarin are highly susceptible to noise, with WRR dropping drastically even at 25% noise and becoming almost unusable at higher noise levels. The "Numbers" variants generally follow similar trends to regular speech within each language.

TC-6: Speed & Latency

This test measures the time taken for transcription.

1.930s

Average Actual Latency

For audio clips of 5-10 seconds. System-reported latency was not available.

An average latency of approximately 1.93 seconds for 5-10 second audio clips is generally acceptable for non-real-time transcription tasks. Individual processing times may vary.

TC-5: Profanity Filtering

Evaluation of the profanity filtering feature could not be completed for the Google STT method as there was no valid data to aggregate after filtering within the provided test set.

Conclusion & Key Recommendations

Strengths Identified:

- ✓ High accuracy for English-US in clean audio.
- ✓ Good recognition of English domain-specific vocabulary.
- ✓ Accurate English-US punctuation in clean conditions.
- ✓ Acceptable transcription latency for non-real-time tasks.

Areas for Improvement:

- ✗ Overall accuracy for Cantonese and Mandarin.
- ✗ Robustness to accents, especially for non-English languages.
- ✗ Performance in noisy environments across all languages.
- ✗ Punctuation accuracy for non-English languages and noisy English.
- ✗ Domain vocabulary support for Cantonese and Mandarin.

Recommendations:

- ✓ Invest in improving base models for Cantonese and Mandarin.
- ✓ Enhance training for accent robustness, particularly cross-language accents.
- ✓ Boost noise robustness through pre-processing and diverse training data.
- ✓ Refine punctuation models for non-English languages and noisy conditions.
- ✓ Facilitate easier domain adaptation (custom vocabularies) for all languages.