STT Benchmark Analysis Report

1. Introduction

This report presents an analysis of Speech-to-Text (STT) performance based on the provided benchmark logs. The analysis covers various test cases designed to evaluate multilingual support, robustness across accents, domain-specific vocabulary recognition, auto-punctuation, profanity filtering, transcription speed, and noise robustness. All tests were conducted using the "google" STT method, as indicated in the logs.

2. Executive Summary

The STT service demonstrated strong performance for standard English (US), particularly in ideal conditions, with high Word Recognition Rates (WRR) and Number Accuracy. Performance for Cantonese and Mandarin varied, generally showing higher Word Error Rates (WER) compared to English. Accented speech significantly impacted recognition, especially for Cantonese with a Mandarin accent and Mandarin with a Cantonese accent, which resulted in complete recognition failure in the provided samples. Domain vocabulary recognition was strong for English but weaker for Cantonese and Mandarin. Auto-punctuation accuracy was generally low across most languages and conditions, with some exceptions in English and specific Mandarin number cases. Noise robustness tests indicated a significant degradation in performance as noise levels increased, with English showing more resilience at lower noise levels compared to Cantonese and Mandarin. Transcription latency was generally within a few seconds for most successful transcriptions. Data for profanity filtering (TC-5) was limited in the provided logs, relying on vocabulary accuracy metrics from other test sets.

3. Detailed Analysis per Test Case

TC-1: Multilingual Support

Objective: To evaluate STT performance across different languages (English-US, Cantonese-HK, Mandarin) for general speech and speech containing numbers.

Metrics:

- Word Error Rate (WER)
- Word Recognition Rate (WRR)
- Number Accuracy (%)

Findings:

A. General Speech (Regular FolderType):

Language	Average WER (%)	Average WRR (%)	Notes
Cantonese-HK	59.84	40.16	High variability; some samples >80% WER
English-US	13.02	86.98	Generally good performance
Mandarin	64.52	41.11	One sample: WER 450% (No speech recognized)

Note: English-UK and English-HK (non-accented, non-number) samples were not distinctly available under TC-1 "Regular" in the provided logs.

B. Speech with Numbers (Numbers FolderType):

Language	Average Number Accuracy (%)	Notes
Cantonese-HK-Numbers	57.89	Wide range (0% to 100%)
English-US-Numbers	97.22	Consistently high, one sample at 50%
Mandarin-Numbers	60.42	Wide range (0% to 100%)

TC-2: Robustness Across Accents

Objective: To evaluate STT performance with various accented speech inputs.

Metrics:

- Word Error Rate (WER)
- Word Recognition Rate (WRR)

Findings:

Accent	Average WER (%)	Average WRR (%)	Notes

Cantonese (Mandarin Accent)	100.00	0.00	Complete recognition failure across all 18 samples.
English (Indian Accent)	13.71	86.29	Relatively good performance.
English (South East Asian Accent)	29.55	70.35	Moderate performance. One sample had an API error (WER 650%).
Mandarin (Cantonese Accent)	100.00	0.00	Complete recognition failure across all 9 samples.

Note: Number accuracy for accented speech was not available in the NumberAccuracy_Percent column for TC-2 samples in the logs.

TC-3: Domain Vocabulary Support (HSBC Specific Terms)

Objective: To evaluate STT performance in recognizing domain-specific vocabulary (e.g., HSBC terms).

Metrics:

- Word Error Rate (WER)
- Vocabulary Accuracy (%)

Findings:

Language	**Average WER (%)**	**Average Vocabulary Accuracy (%)**	**Notes**
Cantonese	66.30	35.53	High WER. Vocab accuracy varied (many 0%, some 33-100%).
English	10.97	90.63	Good performance, high vocabulary recognition.
Mandarin	53.46	50.00	Based on two samples; one with 100% vocab

			accuracy, one with 0%.

Note: The logs included Mandarin samples for TC-3, though the initial notes only specified English and Cantonese-HK.

TC-4: Auto Punctuation Feature

Objective: To evaluate the accuracy of the auto-punctuation feature.

Metric: Proportion of correct punctuation placements (represented by SegmentationAccuracy_Percent in logs).

Findings:
Segmentation accuracy for punctuation was generally low across most languages and conditions. Many samples showed 0% accuracy.

- **Cantonese-HK (Standard & Domain):** Average Segmentation Accuracy: ~0.00%. Most samples showed 0% accuracy.
- **English-US (Standard & Domain):** Average Segmentation Accuracy: ~22.73% (excluding N/A). Highly variable, with some samples achieving 100% but many at 0% or N/A.
- **Mandarin (Standard & Domain):** Average Segmentation Accuracy: ~20.83% (excluding N/A). Variable; some number-specific samples showed higher accuracy (e.g., 100%), while general speech was often 0%.
- **Accented Speech (TC-2 linked):**
  - English (Indian Accent): Average ~18.18% (excluding N/A).
  - English (South East Asian Accent): Average ~30.79% (excluding N/A, one API error).
  - Accented Cantonese/Mandarin: Predominantly 0% or N/A.
- **Noisy Speech (TC-7 linked):**
  - Punctuation accuracy was severely impacted by noise, with most samples across all languages showing 0% or very low accuracy. For instance, Cantonese-HK under various noise levels mostly had 0% segmentation accuracy. English-US showed some resilience at 25% noise but degraded quickly. Mandarin also showed poor punctuation in noisy conditions.

Overall, the auto-punctuation feature appears to be unreliable in its current state as per the logs, especially for non-English languages and in adverse conditions.

TC-5: Profanity Filtering

Objective: To evaluate the STT's ability to identify and filter profanity.

Metric: Rate of profanity vocabulary identified (represented by VocabularyAccuracy_Percent from re-purposed TC-3 logs, as specific TC-5 logs with this metric were not distinct or populated differently in the snippet).

Findings:
The provided log snippet does not contain direct TestCase == "TC-5" entries with populated VocabularyAccuracy_Percent that would clearly represent profanity identification rates. The analysis below is based on interpreting VocabularyAccuracy_Percent from the TC-5 labeled entries in the log, which appear to be duplicates or re-categorizations of TC-3 (Domain Vocabulary) data for the purpose of this report structure. This metric in TC-3 measures recognition of specific (HSBC) terms, not profanity.
Assuming these TC-3 vocabulary entries are used as a proxy for TC-5:

Language	**Average "Profanity" (Domain Vocab) Accuracy (%)**	**Notes**
Cantonese	35.53	Based on TC-3 data; indicates recognition of HSBC terms, not profanity.
English	90.63	Based on TC-3 data; indicates recognition of HSBC terms, not profanity.
Mandarin	50.00	Based on TC-3 data; indicates recognition of HSBC terms, not profanity.

Conclusion for TC-5: The provided logs do not allow for a direct assessment of the profanity filtering feature as described in the test case notes. The VocabularyAccuracy_Percent for the available TC-5 entries (which mirror TC-3) reflects domain term recognition, not profanity.

TC-6: Transcription Speed and Latency

Objective: To measure transcription speed and latency.

Metric: Actual latency in seconds (ResponseSpeed_s). System-reported latency was not available in the logs.

Findings:

The average ResponseSpeed_s across all analyzable samples in the log was approximately 1.85 seconds.
Average Latency by Language Group (approximate):

- **Cantonese-HK (Standard & Numbers):** ~1.81 s
- **English-US (Standard & Numbers):** ~1.88 s
- **Mandarin (Standard & Numbers):** ~1.58 s
- **Accented English:** ~2.02 s
- **Accented Cantonese/Mandarin (where transcription was attempted):** ~1.70 s (though most failed)
- **Noisy Samples (across languages):** Varied, but generally within 1-2.5s for samples that produced output. Many "No speech recognized" entries also had latency recorded.

The latency does not show a strong correlation with language type for successful transcriptions in this dataset. The specific audio clip lengths (8s, 30s, 60s) mentioned in the test case notes could not be directly correlated with latency from the provided log format.

TC-7: Noise Robustness

Objective: To evaluate STT performance in noisy environments at different Signal-to-Noise Ratio (SNR) levels.

Metrics:

- Word Error Rate (WER)
- Word Recognition Rate (WRR)
- Number Accuracy (%) for number-specific noisy samples.

Findings:
Performance degraded significantly as noise levels increased for all languages. "No speech recognized" became common at higher noise levels.

A. Cantonese-HK:

- **25% Noise:** Avg WER 74.58%, Avg WRR 25.42%
- **50% Noise:** Avg WER 98.21%, Avg WRR 3.93% (Many "No speech recognized")
- **75% Noise:** Avg WER 98.25%, Avg WRR 1.75% (Mostly "No speech recognized")
- **100% Noise:** Avg WER 98.53%, Avg WRR 1.47% (Mostly "No speech recognized")
- **Numbers in Noise (Cantonese-HK-Numbers):**
    - 25% Noise: Avg Number Accuracy ~0.00%
    - 50% Noise: Avg Number Accuracy ~0.00%
    - 75% Noise: Avg Number Accuracy ~0.00%

    - 100% Noise: Avg Number Accuracy ~0.00%

B. English-US:

- **25% Noise:** Avg WER 25.64%, Avg WRR 74.36%
- **50% Noise:** Avg WER 60.42%, Avg WRR 39.58% (One "No speech recognized")
- **75% Noise:** Avg WER 83.33%, Avg WRR 16.67% (Many "No speech recognized")
- **100% Noise:** Avg WER 88.54%, Avg WRR 11.46% (Many "No speech recognized")
- **Numbers in Noise (English-US-Numbers):**
    - 25% Noise: Avg Number Accuracy ~75.00%
    - 50% Noise: Avg Number Accuracy ~50.00%
    - 75% Noise: Avg Number Accuracy ~31.25%
    - 100% Noise: Avg Number Accuracy ~16.67%

C. Mandarin:

- **25% Noise:** Avg WER 53.40%, Avg WRR 46.60%
- **50% Noise:** Avg WER 82.66%, Avg WRR 18.20% (Many "No speech recognized")
- **75% Noise:** Avg WER 93.75%, Avg WRR 6.25% (Mostly "No speech recognized")
- **100% Noise:** Avg WER 93.90%, Avg WRR 6.10% (Mostly "No speech recognized")
- **Numbers in Noise (Mandarin-Numbers):**
    - 25% Noise: Avg Number Accuracy ~18.75%
    - 50% Noise: Avg Number Accuracy ~0.00%
    - 75% Noise: Avg Number Accuracy ~0.00%
    - 100% Noise: Avg Number Accuracy ~0.00%

English demonstrated better noise robustness at lower noise levels (25%, 50%) compared to Cantonese and Mandarin for both general speech and number recognition. All languages struggled significantly at 75% and 100% noise levels.

4. Overall Conclusion

The STT service shows varied performance depending on language, accent, acoustic conditions, and specific features like punctuation. While English (US) performance is generally strong in favorable conditions, improvements are needed for other languages, particularly Cantonese and Mandarin, especially concerning accent robustness and performance in noisy environments. The auto-punctuation feature requires significant enhancement across the board. Latency is generally acceptable for successful transcriptions. The provided logs had limitations in directly assessing profanity filtering (TC-5) and correlating latency with specific audio input lengths (TC-6).