STT Performance Analysis Report (Google Method)

1. Executive Summary

This report analyzes the performance of the Google Speech-to-Text (STT) method based on the provided evaluation summary and log data. The analysis covers multilingual support, accent robustness, domain-specific vocabulary, auto-punctuation, profanity filtering, transcription speed, and noise robustness.

Overall, the Google STT method demonstrates strong performance for **English-US, particularly in clean environments and with standard accents**. It achieves low Word Error Rates (WER) and high Word Recognition Rates (WRR) under these conditions. However, its performance significantly degrades for **Cantonese and Mandarin**, especially with accents, domain-specific terms, and in noisy environments. Punctuation accuracy is also a notable challenge for non-English languages.

2. Detailed Performance Analysis

TC-1: Multilingual Support

Summary Results:

- Cantonese-HK: WER 59.42%, WRR 40.58%
- English-US: WER 13.32%, WRR 86.68%
- Mandarin: WER 66.25%, WRR 55.62% (Note: Summary table for Mandarin in TC-1 shows 33.75% WER and 66.25% WRR, but the detailed table shows 66.25% WER and 55.62% WRR. The report will use the values from the detailed table if there's a discrepancy, assuming the detailed table is more accurate or a typo in the summary. For TC-1, the values are 66.25% WER and 55.62% WRR for Mandarin. However, the provided summary table for TC-1 only has Cantonese and English-US. The Mandarin result is from TC-3 for vocabulary. Let's assume TC-1 Mandarin has a high WER as well, consistent with Cantonese.)
  Correction based on provided summary: The TC-1 summary table only lists Cantonese-HK and English-US. Mandarin is listed in TC-3. The report will stick to the provided summary for TC-1.

Analysis:
The STT method performs significantly better for English-US compared to Cantonese-HK. The high WER for Cantonese suggests difficulties in accurately transcribing this language.
Log Examples (Illustrative):

- **Good Performance (English-US):**
  - TestCase: TC-1, LanguageSubFolder: English-US, GroundTruthFile:

1AnPurpl81c_noisy_0_1AnPurpl81c.whisper.auto.gt_audio_8.vtt, PredictionFile: 1AnPurpl81c_noisy_0_1AnPurpl81c.whisper.auto.gt_audio_8.google.txt, WER: 0.00, WRR_Percent: 100.00
    - **GT_Text_Raw:** "But back then, there was only a minority group of elites who were proficient in English."
    - **Pred_Text_Raw:** "But back then there was only a minority group of Elites who were proficient in English."
    - **Reasoning:** Perfect transcription for a standard English-US sentence.
- **Poor Performance (Cantonese-HK):**
    - TestCase: TC-1, LanguageSubFolder: Cantonese-HK, GroundTruthFile: FmpXfFwzvP8_noisy_0_FmpXfFwzvP8.whisper.auto.gt_audio_11.vtt, PredictionFile: FmpXfFwzvP8_noisy_0_FmpXfFwzvP8.whisper.auto.gt_audio_11.google.txt, WER: 83.33, WRR_Percent: 16.67
        - **GT_Text_Raw:** "Security Code 沒有一個安全、驗證碼的問題" (Security Code, there isn't a security, verification code issue)
        - **Pred_Text_Raw:** "安全啊驗證碼" (Security ah verification code)
        - **Reasoning:** Significant word loss, misinterpretation, and missing the English phrase "Security Code." The punctuation and sentence structure are completely lost.
    - TestCase: TC-1, LanguageSubFolder: Cantonese-HK, GroundTruthFile: SbduiuP_gek_noisy_0_SbduiuP_gek.whisper.auto.gt_audio_140.vtt, PredictionFile: SbduiuP_gek_noisy_0_SbduiuP_gek.whisper.auto.gt_audio_140.google.txt, WER: 100.00, WRR_Percent: 0.00
        - **GT_Text_Raw:** "最低的佔比 死尾了, 當然了, 做兩餸飯了," (The lowest proportion, it's dead last, of course, doing two-dish rice,)
        - **Pred_Text_Raw:** "第一季已累" (The first quarter is already tired/accumulated)
        - **Reasoning:** Completely incorrect transcription, indicating a severe failure in recognition.

TC-2: Robustness Across Accents

Summary Results:

- Cantonese_Mandarin_Accent: WER 100.00%, WRR 0.00%
- English_Indian_Accent: WER 13.13%, WRR 86.87%
- English_SouthEastAsian_Accent: WER 31.05%, WRR 79.53% (Note: Summary table shows 68.95% WRR, but 100-31.05 = 68.95. Let's assume 79.53% WRR is a typo and it should be 68.95%)

Correction based on provided summary: English_SouthEastAsian_Accent has WRR 79.53%. This means 100-31.05 is not 79.53. The table is likely correct as is. My prior assumption was wrong. Let's use 79.53%

- Mandarin_Cantonese_Accent: WER 100.00%, WRR 0.00%

Analysis:
The system struggles immensely with heavily accented Cantonese and Mandarin, resulting in complete transcription failure. However, it handles English with an Indian accent relatively well, comparable to standard English-US. The performance for English with a Southeast Asian accent is moderate, with a noticeable increase in WER compared to Indian-accented English.
Log Examples (Illustrative):

- **Good Performance (English_Indian_Accent):**
  - TestCase: TC-2, LanguageSubFolder: English_Indian_Accent, GroundTruthFile: CU5R9c3Wc60_noisy_0_CU5R9c3Wc60.whisper.auto.gt_audio_105.vtt, PredictionFile: CU5R9c3Wc60_noisy_0_CU5R9c3Wc60.whisper.auto.gt_audio_105.google.txt, WER: 0.00, WRR_Percent: 100.00
    - **GT_Text_Raw:** "Suppose this is the point where it has the maximum highest accuracy starting somewhere randomly."
    - **Pred_Text_Raw:** "Suppose this is the point where it has the maximum highest accuracy starting somewhere randomly."
    - **Reasoning:** Perfect transcription despite the Indian accent.
- **Poor Performance (Cantonese_Mandarin_Accent):**
  - TestCase: TC-2, LanguageSubFolder: Cantonese_Mandarin_Accent, GroundTruthFile: JNAeatPbndE_noisy_0_JNAeatPbndE.whisper.auto.gt_audio_0.vtt, PredictionFile: JNAeatPbndE_noisy_0_JNAeatPbndE.whisper.auto.gt_audio_0.google.txt, WER: 100.00, WRR_Percent: 0.00
    - **GT_Text_Raw:** "是雞肉墨丸三小辣收到" (It's chicken meatball, three small spicy, received)
    - **Pred_Text_Raw:** "吃雞我買丸三小辣" (Eat chicken I buy balls three small spicy)
    - **Reasoning:** Completely wrong transcription, demonstrating the system's inability to handle this accent.
- **Moderate Performance (English_SouthEastAsian_Accent):**
  - TestCase: TC-2, LanguageSubFolder: English_SouthEastAsian_Accent, GroundTruthFile: 1AnPurpl81c_noisy_0_1AnPurpl81c.whisper.auto.gt_audio_15.vtt, PredictionFile: 1AnPurpl81c_noisy_0_1AnPurpl81c.whisper.auto.gt_audio_15.google.txt, WER:

100.00, WRR_Percent: 0.00
    - **GT_Text_Raw:** "Like gonna rain now, bring an umbrella la. That's only 7 words."
    - **Pred_Text_Raw:** "English. You could say I'm going to read now bring on Bella. That's only 7."
    - **Reasoning:** This example shows a high WER (100%) for a specific clip, indicating issues with Singlish colloquialisms ("la") and specific phrases. The overall average WER for this accent (31.05%) suggests varied performance. "gonna rain now" became "going to read now", "umbrella la" became "on Bella".

TC-3: Domain Vocabulary Support

Summary Results:

- Cantonese: WER 69.75%, Avg Vocab Accuracy 36.53%
- English: WER 10.95%, Avg Vocab Accuracy 93.33%
- Mandarin: WER 57.37%, Avg Vocab Accuracy 66.67%

Analysis:
The system performs well with English domain-specific vocabulary (HSBC terms). However, it struggles significantly with Cantonese and Mandarin domain terms, with vocabulary accuracy below 70% for both.

Log Examples (Illustrative):

- **Good Performance (English - HSBC Vocabulary):**
    - TestCase: TC-3, LanguageSubFolder: English, GroundTruthFile: RwlmeANAQFQ_noisy_0_RwlmeANAQFQ.whisper.auto.gt_audio_0.vtt, PredictionFile: RwlmeANAQFQ_noisy_0_RwlmeANAQFQ.whisper.auto.gt_audio_0.google.txt, WER: 14.29, WRR_Percent: 85.71, VocabularyAccuracy_Percent: 100.00
        - **GT_Text_Raw:** "Welcome to Perspectives from HSBC, thanks for joining us and now onto today's show."
        - **Pred_Text_Raw:** "Welcome to perspectives from HSBC, thanks for joining us. And now on to Today's Show,"
        - **GT_Vocabs_Found:** "HSBC"
        - **Pred_Vocabs_Matched:** "HSBC"
        - **Reasoning:** Correctly identified "HSBC". Minor errors in capitalization and punctuation contribute to WER but vocab accuracy is high.
- **Poor Performance (Cantonese - HSBC Vocabulary):**
    - TestCase: TC-3, LanguageSubFolder: Cantonese, GroundTruthFile: SbduiuP_gek_noisy_0_SbduiuP_gek.whisper.auto.gt_audio_153.vtt,

PredictionFile: SbduiuP_gek_noisy_0_SbduiuP_gek.whisper.auto.gt_audio_153.google.txt, WER: 83.33, WRR_Percent: 16.67, VocabularyAccuracy_Percent: 100.00 (Note: Vocab accuracy is 100% here but WER is high, which is interesting. This specific log shows "匯豐" was matched.)
    - **GT_Text_Raw:** "擔憂, 匯豐, 股價曾經在一天之內, 暴跌了15%," (Worry, HSBC, stock price once within a day, plummeted 15%,)
    - **Pred_Text_Raw:** "匯豐估價" (HSBC valuation/stock price)
    - **GT_Vocabs_Found:** "匯豐"
    - **Pred_Vocabs_Matched:** "匯豐"
    - **Reasoning:** While "匯豐" (HSBC) was recognized, the overall sentence transcription is poor, leading to a high WER. The average vocabulary accuracy for Cantonese is low (36.53%), indicating many other instances where terms were missed.
- TestCase: TC-3, LanguageSubFolder: Cantonese, GroundTruthFile: YBLuZ09hr6o_noisy_0_YBLuZ09hr6o.whisper.auto.gt_audio_8.vtt, PredictionFile: YBLuZ09hr6o_noisy_0_YBLuZ09hr6o.whisper.auto.gt_audio_8.google.txt, WER: 94.12, WRR_Percent: 5.88, VocabularyAccuracy_Percent: 0.00
    - **GT_Text_Raw:** "Red Card在疫情前匯豐很大力推" (Red Card before the pandemic HSBC promoted heavily)
    - **Pred_Text_Raw:** "就係呢個城市呢就咁推" (It's this city that promotes like this)
    - **GT_Vocabs_Found:** "匯豐, RED, red card"
    - **Pred_Vocabs_Matched:** (empty)
    - **Reasoning:** Key terms "Red Card" and "匯豐" (HSBC) are completely missed, resulting in 0% vocabulary accuracy for this clip and a very high WER.

TC-4: Auto Punctuation Feature

Summary Results (Average Segmentation Accuracy %):

- Cantonese-HK (various conditions): 0.00% - 12.50% (mostly very low)
- English-US (various conditions): 0.00% - 100.00% (100% in clean, degrades with noise)
- Mandarin (various conditions): 0.00% - 32.05% (generally low)

Analysis:
Auto-punctuation works perfectly for English-US in clean audio and reasonably well with light noise (25%). However, its accuracy drops significantly with increased noise. For Cantonese and Mandarin, punctuation accuracy is extremely low across almost all conditions, including

clean audio. The system also struggles with punctuation when accents are present (e.g., English_SouthEastAsian_Accent: 45.31%, Cantonese_Mandarin_Accent: 0.00%).

Log Examples (Illustrative):

- **Good Performance (English-US, Clean):**
  - TestCase: TC-4, SourceTestCaseVTT: TC-1, LanguageSubFolder: English-US, GroundTruthFile: 3BBtS1ir4tA_noisy_0_3BBtS1ir4tA.whisper.auto.gt_audio_15.vtt, SegmentationAccuracy_Percent: 100.00
    - **GT_Text_Raw:** "When I go to Philippines, I love watching TV and all the funny advertisements. Kuya, germs!"
    - **Pred_Text_Raw:** "When I go to Philippines, I love watching TV and all the funny advertisements kuya. Germs."
    - **Reasoning:** Correctly placed comma and period. (Note: "Kuya, germs!" vs "kuya. Germs." shows a slight difference in interpretation but main punctuation is present).
- **Poor Performance (Cantonese-HK, Clean):**
  - TestCase: TC-4, SourceTestCaseVTT: TC-1, LanguageSubFolder: Cantonese-HK, GroundTruthFile: 4FEN95kaPIo_noisy_0_4FEN95kaPIo.whisper.auto.gt_audio_17.vtt, SegmentationAccuracy_Percent: 0.00
    - **GT_Text_Raw:** "總之就是 那個營利受到拖累 同時預期信貸損失" (In short, that profit was dragged down, and at the same time, expected credit losses) (Note: GT itself lacks punctuation here in the log snippet, but GT_Segments_Count is 3, implying punctuation was expected).
    - **Pred_Text_Raw:** "總之就係話佢型你就過嚟啦同時預期信貸損失" (In short, it means if he's stylish you come over, and at the same time expected credit losses)
    - **Reasoning:** The prediction is a single run-on sentence with no punctuation, and the transcription itself is poor. The 0% accuracy indicates a complete failure to segment or punctuate as per ground truth.
- **Poor Performance (English-US, Noisy):**
  - TestCase: TC-4, SourceTestCaseVTT: TC-7, LanguageSubFolder: English-US\noisy_100, GroundTruthFile: CU5R9c3Wc60_noisy_100_CU5R9c3Wc60.whisper.auto.gt_audio_27.vtt, SegmentationAccuracy_Percent: N/A (The summary table shows 0.00% for English-US\noisy_100)
    - **GT_Text_Raw:** "parameters and see which is the lowest error, let's say, okay? This is grid search approach. Whereas random"
    - **Pred_Text_Raw:** "Parameters. Okay, this is great source approach. Let us

run."
    - **Reasoning:** While some punctuation is present (periods), it doesn't match the ground truth structure (missing comma, question mark, and incorrect sentence breaks). High noise (100%) severely impacts punctuation.

TC-5: Profanity Filtering

Summary Results:

- No valid data to aggregate for this STT method after filtering.

Analysis:
The effectiveness of profanity filtering could not be assessed due to a lack of valid data in the test set for the Google STT method.

TC-6: Transcription Speed and Latency

Summary Results:

- Average Actual Latency (s): 1.930
- System-Reported Latency (s): Data not available

Analysis:
The average actual latency is 1.930 seconds for audio clips of 5-10 seconds. This is a reasonable speed for non-real-time applications. The lack of system-reported latency prevents a comparison between perceived and actual processing times. Individual log entries show ResponseSpeed_s varying, for example, from 0.531s to 4.984s.

TC-7: Noise Robustness

Summary Results (Average WRR %):

- Cantonese-HK-Numbers: 0.00% (100% noise) to 10.34% (25% noise)
- Cantonese-HK: 1.66% (100% noise) to 12.38% (25% noise)
- English-US-Numbers: 23.66% (75% noise) to 52.88% (25% noise)
- English-US: 18.44% (100% noise) to 49.92% (25% noise)
- Mandarin-Numbers: 2.18% (100% noise) to 39.09% (25% noise)
- Mandarin: 14.33% (100% noise) to 46.80% (25% noise)

Analysis:
Noise significantly degrades performance across all languages.

- **English-US** shows better resilience to noise compared to Cantonese and Mandarin, maintaining around 50% WRR at 25% noise. However, even for English, WRR drops to around 20-30% at higher noise levels (75-100%).
- **Cantonese and Mandarin** performance is very poor in noisy conditions, with WRR often falling below 15% even at 25% noise, and near 0% at 100% noise.

Log Examples (Illustrative):

- **Relatively Better (English-US, 25% Noise):**
  - TestCase: TC-7, LanguageSubFolder: English-US\noisy_25, NoiseLevel_Percent: 25%, GroundTruthFile: CU5R9c3Wc60_noisy_25_CU5R9c3Wc60.whisper.auto.gt_audio_39.vtt, WER: 0.00, WRR_Percent: 100.00
    - **GT_Text_Raw:** "the searches, more, more variety will be there. However, the problem is, both of these approaches"
    - **Pred_Text_Raw:** "The searches more, more variety will be there. However, the problem is both of these approaches"
    - **Reasoning:** Excellent performance for this specific clip despite 25% noise. The average WRR of 49.92% for this category indicates varied results.
- **Poor Performance (Cantonese-HK, 100% Noise):**
  - TestCase: TC-7, LanguageSubFolder: Cantonese-HK\noisy_100, NoiseLevel_Percent: 100%, GroundTruthFile: 4FEN95kaPIo_noisy_100_4FEN95kaPIo.whisper.auto.gt_audio_100.vtt, WER: 100.00, WRR_Percent: 0.00
    - **GT_Text_Raw:** "是啊 很明顯的 是 因為我們玩 我們這一代是最後機會玩" (Yes, it's obvious, yes, because we play, our generation is the last chance to play)
    - **Pred_Text_Raw:** "No speech recognized."
    - **Reasoning:** Complete failure to transcribe in the presence of 100% noise.
- **Poor Performance (Mandarin, 75% Noise):**
  - TestCase: TC-7, LanguageSubFolder: Mandarin\noisy_75, NoiseLevel_Percent: 75%, GroundTruthFile: 0Ejp6yyU5bo_noisy_75_0Ejp6yyU5bo.whisper.auto.gt_audio_15.vtt, WER: 81.25, WRR_Percent: 18.75
    - **GT_Text_Raw:** "它是銀聯卡 所以我們在中國大陸的話" (It's a UnionPay card, so if we are in mainland China)
    - **Pred_Text_Raw:** "所以我们在这个这么大陆的话。" (So we are in this so mainland China.)
    - **Reasoning:** Significant errors and word loss due to 75% noise. "銀聯卡" (UnionPay card) is missed.

3. Conclusion and Recommendations

Strengths:

- **English-US Transcription:** Performs well with standard US English, especially in low-noise environments and with some accents like Indian English.

- **English Domain Vocabulary:** Good recognition of specific English terms (e.g., HSBC).
- **English Punctuation (Clean Audio):** Accurate punctuation for English-US in clean conditions.
- **Latency:** Acceptable average transcription speed for non-real-time tasks.

Weaknesses:

- **Non-English Languages:** Significantly poorer performance for Cantonese and Mandarin in terms of basic transcription accuracy (WER/WRR), domain vocabulary, and punctuation.
- **Accents (Non-English):** Fails almost completely with Cantonese (Mandarin accent) and Mandarin (Cantonese accent). Performance also degrades for English with certain accents (Southeast Asian).
- **Noise Robustness:** Performance degrades substantially across all languages as noise levels increase. Cantonese and Mandarin are particularly vulnerable.
- **Punctuation (Non-English & Noisy English):** Auto-punctuation is unreliable for Cantonese and Mandarin, and for English in noisy conditions.

Recommendations:

1. **Improve Cantonese and Mandarin Models:** Significant investment is needed to enhance the base models for Cantonese and Mandarin to improve WER/WRR, domain vocabulary recognition, and punctuation accuracy.
2. **Enhance Accent Robustness:** Focus on training models with a wider variety of accents, especially for Cantonese and Mandarin, where performance is currently unacceptable with cross-language accents.
3. **Boost Noise Robustness:** Implement or improve noise reduction pre-processing and train models on more diverse noisy datasets to enhance performance in real-world environments.
4. **Refine Punctuation Models:** Develop more sophisticated punctuation models, particularly for Cantonese and Mandarin, that are less reliant on clean audio and can handle linguistic nuances.
5. **Domain Adaptation:** For specific applications (like HSBC), further fine-tuning or providing custom vocabulary lists for Cantonese and Mandarin could improve domain term recognition.
6. **Profanity Filter Testing:** Ensure future evaluations include sufficient data to test profanity filtering capabilities.

This analysis indicates that while the Google STT method is proficient for certain English use cases, substantial improvements are required for it to be a reliable solution for Cantonese and Mandarin, especially in challenging acoustic conditions or

with accented speech.