STT Evaluation Report

Date of Evaluation: 2025-05-16
Evaluation Tool: SpeechBrain
Mode: Comparing Ground Truth (.vtt) against Hypothesis (.txt)
Metrics Calculated: WER, WRR, SER.

Executive Summary:

This report summarizes the performance of the "google" Speech-to-Text (STT) method.

The "google" method was evaluated across 240 files. It exhibited a very high average Word Error Rate (WER) of 160.51% and a negative average Word Recognition Rate (WRR) of -60.51%. A significant factor contributing to this poor performance appears to be failures in automatic language detection. Many files resulted in "no speech recognized" outputs, incorrect language transcription (e.g., Mandarin hypothesis for Cantonese ground truth), or API errors, leading to extremely high error rates for those specific files. All files processed for the "google" method resulted in a Sentence Error Rate (SER) of 100%.

No metric calculation errors were reported.
The OpenCC Chinese script converter was not initialized, so no Chinese script conversion (e.g., Simplified to Traditional) was performed during this evaluation. This did not directly cause the Google STT errors but is a noted configuration detail.

1. Performance Summary:

Method	File Count (Valid Metrics)	Average WER	Average WRR	Average SER	Files_With_Metric_Errors
google	240	1.6051	-0.6051	1.0000	0

2. Detailed Analysis of the "google" Method:

The "google" method demonstrated significant inaccuracies across the 240 test files. The average WER of 1.6051 (or 160.51%) and an average SER of 1.0000 (or 100%) indicate a very low transcription quality.

Key Issues Observed:

- **Failure of Automatic Language Detection:** This appears to be a primary cause for the high error rates.
  - **Misidentification of Language:** In several instances, the ground truth (GT) was in one language (e.g., Cantonese), while the hypothesis (HYP) was transcribed in another (e.g., Mandarin) or was unintelligible.
    - **Example (Pair 1):**
      - GT (Cantonese): '樓價都會受影響 樓價還大 立場企理說是 他是下行兩個情況的預期出現 如果你說下行兩個情況 第一個下行就是 那個關稅政策就會令到 銀行的利息收...'
      - HYP (Mandarin/Gibberish): '我老家对手银行我家狗狗官属于正常人没没领取家说一下好吧...'
      - Metrics: $WER = 0.9714, SER = 1.0000$
    - **Example (Pair 7):**
      - GT (Cantonese): '大家不想怕 匯豐的提供下行兩個景況 就讓大家去看 預測前面香港的老位怎樣 大家就聽聽這樣 假設他要說 全球的關稅行動升級 和地緣政治關係進一...'
      - HYP (Mandarin/Gibberish): '微风掠过海龟头冠山东经济管理专业博通正television...'
      - Metrics: $WER = 0.9910, SER = 1.0000$
  - **"No Speech Recognized":** A substantial number of files resulted in the STT system reporting "no speech recognized", even when the ground truth contained speech. This directly leads to a WER close to or at 1.0000 and an SER of 1.0000 for these files.
    - **Example (Pair 2):**
      - GT: '沒有在分行線上留過簽字 是不允許使用郵寄表格的方式提交護照的 或許這些是櫃員跟我說 一定要本人到香港才能辦理的原因吧 如果真是這樣 之前線上...'
      - HYP: 'no speech recognized...'
      - Metrics: $WER = 1.0000, SER = 1.0000$
    - **Example (Pair 4):**
      - GT: 'exactly so this is what ive been trying to argue that chinas fundament...'
      - HYP: 'no speech recognized...'
      - Metrics: $WER = 0.9588, SER = 1.0000$
  - **API Call Errors:** Failures resulting in an "error during api call" were observed. What were the common characteristics of files encountering these API call errors, and what was their specific impact on the overall error metrics?
- **Noise Impact:** While language detection is a major issue, the logs also show varying noise levels (e.g., noisy_0, noisy_25, noisy_100). Higher noise levels generally correlate with poorer

performance (higher WER or "no speech recognized"), which is expected, but the language detection failures seem to be the dominant error factor.

3. Conclusion:

The evaluation highlights significant performance issues with the "google" STT method under the tested conditions. The predominant factor for the high error rates ($AverageWER = 1.6051$, $AverageSER = 1.0000$) is the failure of its automatic language detection mechanism. This resulted in numerous instances of incorrect language transcription, "no speech recognized" outputs, or API errors.