STT Evaluation Report

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

Executive Summary:

This report summarizes the performance of two Speech-to-Text (STT) methods: "google" and "perfect_reference". The "perfect_reference" method, as expected, achieved perfect scores, serving as a baseline.

The "google" method was evaluated across 240 files. It exhibited a very high average Word Error Rate (WER) / Character Error Rate (CER) of 160.51% and a negative average Word Recognition Rate (WRR) / Character Recognition Rate (CRR) 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 for either method.
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 (CER)	Average_WRR (CRR)	Average_CER	Average_SER	Files_With_Metric_Errors
google	240	1.6051	-0.6051	1.6051	1.0000	0
perfect_reference	1	0.0000	1.0000	0.0000	0.0000	0

Note: The summary log uses "Average_WER (CER)" and "Average_WRR (CRR)" suggesting that for character-based languages, WER might be reported as CER and WRR as CRR. The separate "Average_CER" column has the same value as "Average_WER (CER)" for Google.

2. Detailed Analysis of the "google" Method:

The "google" method demonstrated significant inaccuracies across the 240 test files. The average CER 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: $WER(CER) = 0.9714, SER = 1.0000$
    - **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: $WER(CER) = 0.9910, SER = 1.0000$
- **"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 CER 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: $WER(CER) = 1.0000, SER = 1.0000$
    - **Example (Pair 4):**
        - GT: 'exactly so this is what ive been trying to argue that chinas fundament...'
        - HYP: 'no speech recognized...'
        - Metrics: $WER(CER) = 0.9588, SER = 1.0000$
- **API Call Errors:** Several files, particularly longer audio chunks (60s), resulted in an "error during api call". These errors also contribute to the high overall error rates as the hypothesis becomes a long error string, yielding very high CER values (e.g., >1.0).
    - **Example (Pair 11):**
        - GT: '哈囉各位好 今天我們來講一下pulse信用卡的用卡心得 這是一張匯豐銀行 香港匯豐銀行 發行的一張銀聯信用卡在中國大陸非常的火 那麼它為什麼...'
        - HYP: 'error during api call for sampledtestcasetc1chunk60s0ejp6yyu5bonoisy10...'
        - Metrics: $WER(CER) = 5.4920, SER = 1.0000$
    - **Example (Pair 16):**
        - GT: 'because of the deepseek because of the potential of integrating ai int...'
        - HYP: 'error during api call for sampledtestcasetc1chunk60srwlmeanaqfqnoisy0r...'
        - Metrics: $WER(CER) = 1.2674, SER = 1.0000$
- **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 CER or "no speech recognized"), which is expected, but the language detection failures seem to be the dominant error factor.
- **Inconsistent Performance:** Even when some speech was recognized and the language was seemingly correct, the accuracy varied significantly.
    - **Example (Pair 3 - Relatively Better):**
        - GT: '於是當場幫我在app申請投資賬戶 還說會有一封平遊信件寄給我 需要我收到後拍照在app提交 投資賬戶才會啟動 我突然反映過來 以平遊的效率恐...'
        - HYP: '你是当场帮我在一批申请复制账户来说会有一份凭留信件寄给我需要我收到后台照在app提交投资账户才会启动我突然反应过来的朋友的下一个月的时候不到...'
        - Metrics: $WER(CER) = 0.4458, SER = 1.0000$
    - **Example (Pair 21 - Relatively Better for Short Audio):**
        - GT: '所以剛好我們就符合這個上市界的 這樣一個用卡的需求 因為我們基本上都在大陸...'
        - HYP: '水刚好我们就符合这个上世界的这样一个用卡的需求用什么用我们基本上都在打啊...'
        - Metrics: $WER(CER) = 0.4857, SER = 1.0000$

3. Detailed Analysis of the "perfect_reference" Method:

Only one file was processed using the "perfect_reference" method.

- **Example (Pair 22):**
    - GT: '所以剛好我們就符合這個上市界的 這樣一個用卡的需求 因為我們基本上都在大陸...'
    - HYP: '所以剛好我們就符合這個上市界的這樣一個用卡的需求因為我們基本上都在大陸...'

- Metrics: $WER(CER) = 0.0000, WRR(CRR) = 1.0000, CER = 0.0000, SER = 0.0000$

As expected, this method yielded perfect scores ($CER = 0$, $SER = 0$), confirming it acts as a correct baseline for the evaluation process.

4. Conclusion:

The evaluation highlights significant performance issues with the "google" STT method under the tested conditions. The predominant factor for the high error rates ($Average\_CER = 1.6051$, $Average\_SER = 1.0000$) 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.