MCE: Mixed Cantonese and English Audio Dataset

Appendix
Paper ID: 7361

1 Supplementary Experiments

The example below demonstrates the recognition performance of Cantonese datasets and English-Cantonese mixed-language datasets. The recognition results for various pre-fine-tuning model sizes are shown in Table 1. For models with a relatively larger number of parameters, such as Large and Medium models, they can understand the meaning of sentences relatively accurately. However, they struggle to effectively recognize spoken content, and instead, partially translate some English into simplified Chinese. This indicates that the model can recognize words with significant pronunciation differences between the two languages, but potential issues arise due to inconsistent quality of the training dataset. For instance, if the training set contains English audio from movies and TV shows while the subtitles (annotations) are in Chinese, it may result in inaccuracies. On the other hand, for the Whisper-small model with fewer parameters, the presence of English words interspersed in the sentences introduces significant interference, making it unable to accurately grasp the meaning of sentences. The above experiments demonstrate that the Whisper model cannot effectively recognize mixed-language audio information. Only when the model size reaches a certain scale do the models exhibit a certain level of contextual understanding, enabling them to speculate about the meaning of sentences. However, due to the uneven quality of the training dataset, accurate restoration of speech is hindered.

Before fine-tuning, the models not only exhibit poor recognition capabilities for mixed-language audio but also tend to output Cantonese as simplified Chinese. Although they can translate the sentences to convey the general meaning, it still affects the accuracy of the sentences. The main reason for these issues is that the Chinese dataset is significantly larger than the Cantonese dataset, and it has higher annotation quality. This causes the pre-fine-tuned Whisper model to lean towards translating Cantonese into the closely related simplified Chinese. To address these deficiencies, the fine-tuned Whisper-MCE model proposed in this paper, based on the MCE dataset, can more accurately recognize Cantonese. Table 2 shows a comparison of the recognition performance between the fine-tuned Whisper-MCE and the pre-fine-tuned models. After fine-tuning on the MCE mixed-language dataset, the performance of the model on the Cantonese dataset has significantly improved. Surprisingly, even the model with fewer parameters, Whisper-MCE-Small, outperforms Whisper-Medium in recognition accuracy.

Table 3 displays the recognition performance on the test set of the MCE dataset. The Whisper-Medium model, with fewer parameters, lacks contextual understanding and tends to recognize English words as Cantonese with similar pronunciation. For instance, the pronunciation of ”phoon” is close to the Cantonese word ”歡” (huan). The Whisper-Large-V2 model exhibits a certain level of contextual understanding, allowing it to relatively accurately translate the meaning of sentences. However, due to its excessively large number of parameters, it incurs high computational costs and inference delays, limiting its applicability to domains insensitive to computational intensity and latency. After fine-tuning, the Whisper-MCE-Small model not only accurately understands the meaning of sentences but can also recognize different languages. Additionally, it is observed that in the Hong Kong and Macau regions, the words ”Kyoto” and ”Tokyo” have similar pronunciations. Given the higher frequency of ”Tokyo” in the dataset, the model tends to prefer recognizing it over ”Kyoto.” This observation underscores the importance of high-quality datasets in training speech recognition models.

Model	Audio Information	Speech Recognition Result
Whisper-Large-v2	你睇下我schedule, meeting, 又有training, 仲要同事set target, 今晚要OT咯, Hold住.	你看看我的行程, 又有訓練, 又有會議, 還要幫同事設定目標, 今晚要OT了, 先等一等。
Whisper-Medium		你看看我的行程, 又有訓練, 又有會議, 還要幫同事設定目標, 今晚要OT, 先鋪著。
Whisper-Small		你看看我的訊息, 又有訊息, 又有訊息, 還有幫我同時去做訊息, 今天要OT了, 先等我先。
Whisper-Large-v2	香港真係美食天堂, 入邊有好多唔同小食, 例如魚蛋、燒賣、碗仔翅, 我覺得真係好好食, 同埋我都同屋企人一齊去唔同旅遊景點， 例如迪士尼海洋公園等等，真係一次流連忘返經歷。	香港真是美食天堂,裡面有很多不同的小吃,例如魚蛋、燒賣、碗仔翅, 我覺得真的很好吃, 還有我和家人也去了不同的旅遊景點, 例如迪士尼、海洋公園等等, 真是一次流連忘返的經歷。
Whisper-Medium		香港真是美食天堂裡面有很多不同的小吃, 例如魚蛋、燒米、碗仔翅, 我覺得真是很好吃, 還有我和家人一起去了不同的旅遊景點, 例如迪士尼、海洋公園等等, 真是一次流連忘返的經歷。
Whisper-Small		香港真的是沒吃天堂, 裡面有很多不同的小吃, 例如魚蛋、燒米、碗仔翅, 我覺得真的很好吃, 還有我跟家人一起去不同旅遊景點, 例如迪士尼、海洋公園等等, 真的是一次流連亡犯的經歷。

Table 1: Whisper Performance in Single Language and Mixed Language Speech Recognition

Model	Audio Information	Speech Recognition Result
Whisper-Large-v2	我今晚要去酒吧係咪好難搵啊? 呢間餐廳菜式真係唔錯, 特別係佢燒鵝. 呢度交通都算方便, 不過搭巴士要留意返工時間. 呢個問題唔係咁簡單, 要多加思考先得解決. 我聽日要去開會, 要準備好筆記同報告.	我們今晚要去的酒吧是不是很難找呢? 這間餐廳的菜式真的不錯, 特別是他的燒鵝. 這裡溝通也算方便, 不過搭巴士要留意上班時間. 這個問題不是那麼簡單, 要多加思考才能解決. 我明天要去開會, 要準備好筆記和報告.
Whisper-Medium		我們今晚要去的酒吧是不是很難找呢? 這間餐廳的菜式真的不錯, 特別是它的燒鵝, 這裡高通也算方便, 不過坐巴士要留意上班時間, 這個問題不是那麼簡單, 要多加思考才能解決, 我聽的是要去開會, 要準備可不忌同報告.
Whisper-MCE-Small		我今晚要去酒吧係咪好難搵呀? 呢間餐廳菜式真係唔錯, 特別係佢燒鵝, 呢度高通都算方便, 不過搭巴士要留意返工時間, 呢個問題唔係咁簡單, 要多加思考先得解決, 我聽要去開會,要準備可不記同報告.
Whisper-Large-v2	唔使唸啦, 你功課做哂未啊? 今日天氣好熱, 要多水啊. 我今晚去邊度食飯好呢? 啊, 你真係好識講笑啊. 今日天氣唔錯, 出行下街啦. 啊, 唔好咁懶啦, 收拾下屋企啦.	不用想了, 你的功課我做完了嗎? 今天天氣很熱, 要多喝水. 我們今晚去哪裡吃飯好呢? 啊? 你真的很會說笑啊今天天氣不錯, 出來逛逛街吧啊? 不要這麼懶, 收拾一下家吧.
Whisper-Medium		不用想了, 你的功夫我做完沒有今天天氣很熱, 要多喝水我們今晚去哪裡吃飯好呢? 你真是好會說笑今天天氣不錯, 出來逛逛街吧不要這麼, 收拾一下家.
Whisper-MCE-Small		唔使諗啦, 你功課做晒未呀? 今日天氣好熱, 要多飲水呀. 我今晚去邊度食飯好呢? 啊? 你真係好識講笑呀? 今日天氣唔錯, 出行街啦? 啊? 唔好咁冷呀? 收拾下屋企啦.

Table 2: Whisper and Whisper-MCE Performance in Single Language and Mixed Language Speech Recognition

Model	Audio Information	Speech Recognition Result
Whisper-Large-v2	我睇天氣報告話下個月可能會有typhoon, 我要做好萬全準備.	我剛剛看天氣報告說下個月可能會有颱風, 我們要做好萬全的準備.
Whisper-Medium		我剛剛看天氣報告說下個月可能會有開歡,我們要做好漫傳的準備.
Whisper-MCE-Small		我睇天氣報告話下個月可能會有typhoon, 我要做好萬全準備.
Whisper-Large-v2	我准Summer去日本Kyoto, 欣下花火festival魅力啦!	我準備在夏天去日本的東京, 欣賞花火節的魅力.
Whisper-Medium		我準備在夏天去日本的Tokyo, 拍下花火Festival的命力.
Whisper-MCE-Small		我準備Summer去日本Tokyo, 欣賞下花火Festival魅力喇!

Table 3: Whisper and Whisper-MCE Performance in Single Language and Mixed Language Speech Recognition