利用遮罩增強的語言模型校正技術提升中文醫療語音的自動語音識別效果 (Enhancing Automatic Speech Recognition for Chinese Medical Speech Using Masking-Enhanced Language Model Correction)

林璟芸 Jing-Yun Lin; 李旭清 Xu-Qing Li; 鍾聖倫 Sheng-Luen Chung
國立臺灣科技大學電機工程學系
Electrical Engineering Department
National Taiwan University of Science and Technology
Taipei, Taiwan
slchung@mail.ntust.edu.tw; fdsa3304@gmail.com; stanley890314@gmail.com

摘要

本研究優化了一個針對中文醫療語音的自動語音識別系統 (ASR)。現有的 ASR 系統如 Google 的 STT 和 OpenAI 的 Whisper，雖然在一般語音識別中表現良好，但在醫療領域識別準確度不足。為此，本研究提出「遮罩增強語言模型校正技術」，通過遮罩機制對 ASR 輸出進行校正，以提升識別精度。方法包括微調 CKIP BERT 模型構建 MedCKIPBERT，並提出同音異字替換策略。結果顯示，Google STT 的關鍵詞錯誤率從 15.37%降至 13.64%，Whisper 則從 20.82%降至 15.80%。該方法在醫療語音識別中展現潛在價值。

Abstract

This study optimized an Automatic Speech Recognition (ASR) system for Chinese medical speech. Existing ASR systems like Google's STT and OpenAI's Whisper perform well in general speech recognition but lack accuracy in the medical field. To address this, we proposed a "mask-enhanced language model correction technique" that uses masking to correct ASR outputs, thereby improving recognition accuracy. The methods include fine-tuning the CKIP BERT model to build MedCKIPBERT and implementing a homophone substitution strategy. Results showed that the keyword error rate (KER) of Google STT decreased from 15.37% to 13.64%, and Whisper's KER dropped from 20.82% to 15.80%. This approach demonstrates potential value in medical speech recognition.

關鍵字：自動語音識別 (ASR)、中文醫療語音、醫療中文 BERT、遮罩增強的語言模型
Keywords: Speech Recognition (ASR), Chinese Medical Speech, Medical Chinese BERT, Masking-Enhanced Language Model

1 簡介

1.1 醫療語音辨識

隨著 ASR 技術的進步，其在醫療領域的應用日益增多，尤其在自動生成醫護交班病歷和病歷數位化方面展現出顯著效益。

自動生成醫護交班病歷能節省醫護人員手動記錄病人狀況的時間，降低錯誤率，並促進病歷的數位化與標準化。Chung et al. (2021) 利用中文醫療語音資料庫 psChiMeS-14 和 Joint CTC、Transformer 等模型，提升 ASR 系統表現，並建立「ChiMeS+」醫療語音辨識系統。此外，Enarvi et al. (2020) 探討了通過 ASR 與神經網絡摘要生成醫療報告，減少診斷過程中的繁瑣記錄工作。

儘管語音辨識技術在日常生活中已普及，但在醫學、金融、法律等專業領域仍面臨挑戰。首先，專業語料庫蒐集困難且昂貴，需涵蓋多種口音與情境並精確標記。其次，專業術語的發音和拼寫差異增加辨識難度，不同地區用詞差異也影響準確性。此外，語音辨識系統在未見過的醫療科別中可能無法正確識別新關鍵字。中英文交錯時，華人發音的音節化問題也會導致識別錯誤，如將“hemovac”識別為“黑魔法”。這些因素使專業領域的語音辨識應用充滿挑戰。

1.2 貢獻

本論文提出利用 BERT 輔助的遮罩語言模型 (MLM) 校正 ASR 輸出的方法，無需重新訓練 ASR 或蒐集大量專業領域語料庫，只需串接微調過的 BERT 即可顯著提升專業術語的辨識精度。此外，論文還構建了醫學語文模型 MedCKIPBERT，透過微調 CKIP BERT (https://huggingface.co/ckiplab/bert-base-chinese)，支持醫療文本分類、命名實體識別 (NER)、臨床決策支援等醫療 NLP 應用，助力醫療數位化和智能化。

1.3 本文架構

本論文各節內容如下：第二節「串接 MLM 校正架構」回顧文獻，涵蓋 ASR 模型及其糾錯技術，介紹了 Medical Bert 模型 MedCKIPBERT，並討論 CER 與 KER 績效指標。第三節「MLM 的方法」探討了 Mask-MLM 和近似音-MLM 的 ASR 校正技術。第四節「實驗與結果」展示了測試集選擇及其結果，分析了 Google STT 與 Whisper Medium 模型結合 BERT 在醫療語音中的表現，並進行效果對比。最後，第五節「結論」總結研究成果與應用意義。

2 串接 MLM 校正架構

2.1 文獻審閱

本段文獻審閱回顧了 ASR 技術的發展，從傳統架構到端對端模型的成功應用，並探討了領先系統如 Google STT 和 OpenAI Whisper，這些系統在多語言識別中展現出色能力。最後，分析了 ASR 錯誤校正技術，特別是 BERT 模型及其變體在醫療文本中的應用與提升準確性的潛力。

ASR 技術：傳統的 ASR 系統由三個主要模組組成：聲學模型、發音模型和語言模型。聲學模型 (Acoustic model) 負責將語音信號轉換為音素 (phoneme) 的概率分布；發音模型 (Pronunciation model) 則將這些音素轉換為詞或短語，依賴發音詞典來完成此過程；語言模型 (Language model) 根據語料庫中的詞語頻率和上下文，提供詞語出現的概率，使得語句在語法和語義上更合理。這三個模型各自獨立訓練，且需大量的訓練資料支持，然而，若前一個模型出現錯誤，會影響後續模型的準確性，導致整體識別表現不佳。這種架構的限制使得傳統 ASR 系統在應對複雜語音環境時的表現有所不足。

端對端 ASR 模型：隨著深度學習技術的發展，端對端 (End-to-End, E2E) ASR 模型逐漸成為研究的熱點。這些模型通常使用一個單一的深度神經網絡來直接從語音信號預測出文字輸出，從而簡化傳統 ASR 的多組件結構。主要有三種方法：連結時序分類 (CTC) (Amodei et al., 2016) 透過刪除重複字和空白標籤來自動對齊輸出與輸入，而注意力機制 (Attention-based models) (Chan et al., 2016) 允許模型動態“關注”輸入序列的不同部分，根據隱藏層狀態和編碼器的時序資訊決定輸出。自注意力機制 (Self-attention) (Vaswani et al., 2017) 則專注於全局依賴關係，使 Transformer ASR 能平行處理時序資料，提升訓練速度和效率。

領先的 ASR 模型：領先的 ASR 模型包括 Google STT 和 Whisper。Google STT (Zhang et al., 2023) 使用 1200 萬小時音訊數據及 Conformer 架構，擁有 20 億參數，支援 125 種語言。其訓練分為三階段：自我監督學習、多目標預訓練和微調，提升模型在特定應用場景的性能。Whisper (Radford et al., 2023) 是 OpenAI 開發的多語言 ASR 系統，使用 Transformer 架構，訓練於 68 萬小時音訊數據，擁有 15.5 億參數，支援 50 種語言。本研究選擇了 769 百萬參數的 Whisper Medium，能平衡資源消耗與準確度，適合複雜語音環境。

ASR 模型的糾錯與校正技術：自動語音識別 (ASR) 系統在處理自然語言時，經常會因為口音、背景噪音、語音相似性等問題產生錯誤。為了提高 ASR 系統的準確性，研究者們提出了各種校正技術，包括語音模型優化、語言模型整合以及錯誤修正方法。

Z. Fang 等人提出了非自回歸中文 ASR 錯誤修正方法 PhVEC(Fang et al., 2022)，其創新在於引入音韻字符 (Phonological Tokens)，將拼音作為特殊標記插入原句，提升修正準確性。該方法使用錯誤偵測網絡標記錯誤字符，並生成相應拼音進行修正，如將「你表難過」中的「表」修正為「不要」。實驗顯示，PhVEC 顯著降低字錯率 (WER)，推理速度比

傳統方法快 6.2 倍，大幅提升中文 ASR 錯誤修正的效率和準確性。

Mani et al. (2020) 提出利用機器翻譯模型進行 ASR 錯誤校正和領域適應的方法，透過生成大量錯誤-正確句子對來訓練模型，學習從錯誤輸出到正確文本的轉換規則。實驗顯示，該方法在 Google ASR 系統中顯著提升準確性，字錯誤率 (WER) 改善 7%，BLEU 分數提升 4 分，展現了在領域不匹配情況下的校正效果。

Udagawa et al. (2022) 探討利用大規模語言模型 (如 GPT-2、BERT、RoBERTa) 對 ASR 系統生成的多個候選結果 (N-best hypotheses) 進行重新排序 (rescoring)。模型計算 ASR 的分數 (Score_AM) 和語言模型的分數 (Score_LM)，並以線性組合得到最終得分。實驗結果顯示，雙向語言模型 (如 BERT、RoBERTa) 能顯著改善 ASR 表現並有效減少錯誤率。

Udagawa et al. (2022) 與本篇論文所提出的「BERT 輔助的遮罩語言模型 ASR 校正法」最大的差別在於，(Udagawa et al., 2022) 的方法中，所有的候選結果仍然是基於 ASR 系統的輸出，語言模型 (LM) 僅作為輔助工具，用來計算 ASR 給出的 N-best hypotheses 並對其進行重新排序 (rescoring)。因此，最終的輸出仍然依賴於 ASR 系統曾見過的字彙和詞彙量。相比之下，本研究的校正方法能充分利用語言模型所掌握的詞彙量以及有在醫療情境下經過微調的優勢，直接修正 ASR 系統中因未見過的詞彙或專業術語而產生的錯誤，從而有效減少對專有名詞和專業術語的辨識錯誤。

2.2 Medical Bert 語言模型：MedCKIPBERT

為了實現我們所提出如圖 1 的「利用遮罩增強的語言模型校正技術」(Masking-Enhanced Language Model Correction Technique)，我們需要一個微調 (fine-tune) 在醫療領域上的 BERT，為此我們蒐集醫療文本以作為微調資料。

ChiMed250 醫療文本：我們從網路上各個論文與相關競賽中蒐集了不同的中文醫療資料集，最終彙整成一共 250MB 的醫療文本 (ChiMed250)，其中包含了線上問診對話、醫療詞彙、醫護交班病歷與電子病歷所組成，如圖 2。要特別強調的是，ChiMes250 文本中不管是來自線上資料或是病歷都不包含可識別的個人穩私資訊。

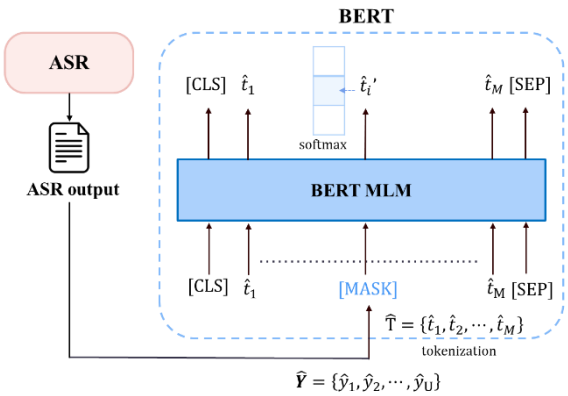

圖 1: 利用遮罩增強的語言模型校正技術

其中有許多語料庫來自於中文醫療資訊處理挑戰榜 CBLUE (Chinese Biomedical Language Understanding Evaluation)。CBLUE 是由中國中文資訊學會醫療健康與生物資訊處理專業委員會在合法開放共享的概念下發起，並由阿里雲天池平台承辦。這個標準旨在推動中文醫學自然語言處理技術的發展，涵蓋命名實體識別、知識抽取、診斷標準化、句子分類以及線上輔助醫療系統的評測 (https://tianchi.aliyun.com/dataset/95414?lang=zh-cn)。

醫療文本 (ChiMed250) 分成兩大類，(1) 249MB 的門診諮詢文本；(2)1.3MB 的醫護交班病歷：

(1) 門診諮詢文本：由線上問診對話與醫療詞彙所組成。

- ChiMed (Tian et al., 2019)：醫療保健平台問答，由從線上醫療保健平台“39 健康網”收集的問答對組成，包含 46,731 個問題及 91,416 個回答 (https://www.39health.com.tw/https://www.39health.com.tw/)。
- CHIP-CDN (CBLUE)：病歷中的診斷實體，關於同一種診斷、手術、藥品、檢查、化驗、症狀等往往會有數百種不同的寫法。標準化為臨床上各種不同說法找到對應的標準說法。
- CHIP-CDN_手術詞 (CBLUE)：手術詞彙，關於同一種手術詞彙標準化為臨床上各種不同說法找到對應的標準說法。
- CHIP-CDN_國際疾病臨床詞 (CBLUE)：關於同一種診斷、手術、藥品、檢查、

門診諮詢文本		
資料集	資料類型	資料量
ChiMed	醫療保健平台問答	184,471 KB
CHIP-CDN	病歷中的診斷實體	1,259 KB
CHIP-CDN_手術詞	手術詞彙	116 KB
CHIP-CDN_國際疾病臨床詞	國際疾病臨床詞彙	1,382 KB
CHIP-MDCFNPC	線上問診對話	18,868 KB
IMCS-V2	線上問診對話	16,981 KB
MedDG	問診對話	27,386 KB
THUOCL	醫學詞彙	384 KB
總資料量: 249.527 MB		

醫護交班病歷		
資料集	資料類型	資料量
CHIP-CDEE	中文電子病歷	272 KB
Text2DT	診療指南、診療本文	95 KB
GPT-Mes2023	醫護交班病歷	953 KB
總資料量: 1.320MB		

圖 2：ChiMed250 醫療文本

化驗、症狀等往往會有數百種不同的寫法。標準化國際疾病臨床詞彙為臨床上各種不同說法找到對應的標準說法。

- CHIP-MDCFNPC (CBLUE)：春雨醫師網站 (https://m.chunyuyisheng.com/) 的線上問診資料，訓練集：5,000，驗證集：1,000，測試集：2,000。
- IMCS-V2 (CBLUE)：線上問診對話，包含 10 種兒科疾病，訓練集：2,472，驗證集：833，測試集：811。
- MedDG (CBLUE)：問診對話，醫生和患者交流的句對話歷史，訓練集：17,864，驗證集：2,747，測試集：1,551。
- THUOCL(https://github.com/thunlp/THUOCL)：醫學詞彙，是由中國清華大學自然語言處理與社會人文計算實驗室整理推出的一套中文詞庫，其中也包含了大量醫學類詞彙，醫學詞條數量共 18,749 條。

(2) 醫護交班病歷：由醫護交班病歷、電子病歷所組成。

- CHIP-CDEE (CBLUE)：中文電子病歷，共 2,485 份電子病歷。
- Text2DT (CBLUE)：臨床診療指南、醫師診療文本，訓練集：300，驗證集：100，測試集：100。
- GPT-Mes2023：將本實驗室與成大醫院的 28 位女性護理師合作蒐集編撰，的醫護交班病歷 M2023 中的訓練集，輸入給 GPT 使其模仿交班病歷特殊形式，生成內容不同的資料。

上述收集的醫療文本中，許多資料集都來自於中國的資源。考量兩岸用詞差異，我們蒐集樂詞網 (https://terms.naer.edu.tw/download/) 上的繁簡醫療詞彙對照表，並對資料進行了轉換。然而，即使經過轉換，仍有部分中國特有的醫療詞彙保留了下來，這些詞彙提供了不同地域性資料和專業術語，進一步豐富了我們的語料庫，並為系統引入了更多新的詞彙和表達方式。

我們將 ChiMed250 語料庫與 BioBERT (Jinhyuk, 2020) 和 EMBERT (Cai et al., 2021) 所使用的語料庫進行比較。BioBERT 的預訓練語料庫規模龐大，超過 21 億字，但全部為英文資料。相比之下，EMBERT 使用 5GB 的中文醫療預訓練資料，主要來自丁香園 (https://www.dxy.cn/) 的醫療問答和論壇數據，專門針對中文醫療場景。

儘管 ChiMed250 語料庫僅有 250MB，其資料來源更專業且多元，涵蓋門診諮詢、線上問診、醫療詞彙、醫護交班記錄和電子病歷等專業醫療文本，覆蓋多個科別及多樣化病歷描述。相比 EMBERT 偏重醫療問答，ChiMed250 提供更廣泛的醫療文本，更全面覆蓋不同的醫療應用場景。

Pretrained BERT CKIP：在獲得醫療文本 (ChiMed250) 後，我們選擇中研院文詞知識庫小組提出的 CKIP BERT (bert-base-chinese) 模型進行微調 (Fine-tuning)。該模型包含 1 億 2 百萬個參數，訓練於中文維基百科和中央通訊社的新聞資料上。表 1 列出了醫療文本資料的大小。

Corpus	Lines	Characters
ChiMed250	1,895,834	87,375,779

表 1：ChiMed250 醫療文本資料統計

MLM 微調 (Fine-tuning)：在具備 250MB 醫療文本、CKIP BERT 預訓練模型和 WordPiece 分詞器後，我們進行微調。即在特定數據集上使用 Masked Language Model (MLM) 技術再訓練 BERT，讓模型更適應特定語言、術語或場景。過程中隨機遮罩部分醫療文本，使用 BERT 進行預測，計算預測結果與原始文本間的損失，並使用梯度下降法更新參數。微調後，模型能更準確處理特定領域語言特徵。圖 3 為 MLM 微調演算法。

Algorithm 1 *BERT Model Fine-tuning with MLM*

Require: *Pretrained BERT model* B_p , *Training dataset* D_{train}
Ensure: *Fine-tuned model* B'
1: ***for*** *each selected* ***s*** *in* D_{train} ***do***
2: *Generate masked of* s_i , *denoted* $M(s_i)$
3: *Compute predictions* $s_i' = M(s_i)$
4: *Compute loss* $L(s_i', s_i)$ *comparing* s_i' *and* s_i
5: *Update* B_p *using* L *to minimize the loss*
6: ***end for***
7: $B' \rightarrow B_p$ // *The fine − tuned model*
8: ***return*** B'

圖 3：MLM 微調 (Fine-tuning) 演算法

具體步驟如下：

1. 訓練資料前處理與遮罩：對於一組訓練資料，首先將句子 $S = \{s_1, s_2, \ldots, s_n\}$ 進行前處理。在過程中會隨機選擇其中 15% 的字符 s_i 進行遮罩處理。在這些被選中的字符中，80% 的字符會被替換為 [MASK] 符號，以 $M(s_i)$ 表示，10%會替換為隨機單詞，另外 10%會保持不變。
2. 模型輸入與預測：將遮罩處理後的句子 $S = \{s_1, s_2, M(s_i), \ldots, s_n\}$ 輸入模型，模型根據上下文訊息和已知詞彙預測被遮罩字符的正確內容，計算預測結果s'_i與真實答案 s_i 的交叉熵損失 (Cross Entropy Loss) $L(s'_i, s_i)$。
3. 損失計算與權重更新：計算預測結果與原始文本之間的損失，並使用梯度下降法 (Gradient Descent) 和反向傳播算法 (Backpropagation) 更新模型參數，以最小化損失 L。
4. 迭代訓練：重複上述步驟，直到模型收斂。

透過將 CKIP 的 pretrained BERT 做 MLM 微調在醫學文本上後，就得到了 MedCKIPBERT。

2.3 CER 與 KER 績效指標

在本研究的所有實驗中將會使用兩個不同的指標：CER 與 KER 來對語音辨識模型進行效能評測。

字符錯誤率 (Character Error Rate, CER)：在醫療語音中，講者常混合使用中文和英文，為了更準確地評估 ASR 績效，我們使用字符錯誤率 (Character Error Rate, CER) 替代傳統的詞錯誤率 (Word Error Rate, WER)。CER 計算每個字符的插入、刪除和替換錯誤，以反映雙語系統在處理中英文混合文本時的真實表現。中文字符以單個字計算，英文則按單音節計算。例如，‘glucose’分為‘glu’和‘cose’，與中文的‘血糖’等權重。此外，標點符號也被視為一個字符進行計算，但我們也使用 CER_NP 來表示不計入標點符號的 CER。

關鍵詞錯誤率 (Keyword Error Rate, KER)：在醫療領域，關鍵詞的準確識別比整體翻譯的相似度更為重要，因此我們使用關鍵詞錯誤率 (Keyword Error Rate, KER) 來取代 BLEU 指標作為 ASR 的績效評估標準。KER 類似於詞錯誤率 (Word Error Rate, WER)，但專注於特定的醫療關鍵詞，計算這些關鍵詞的插入、刪除和替換錯誤率。KER 越低，表示辨識結果越好。為確保評估的準確性，我們彙整了一份醫療關鍵詞表，通過初步篩選、使用 ChatGPT 提取醫療詞彙，並由人工確認後，最終確立這些關鍵詞。

3 MLM 的方法

在訓練或預訓練 BERT 時，我們會隨機遮罩部分字符，強迫編碼器根據前後文預測這些字符並調整其編碼，過程包括遮蔽、預測與調整。然而，本論文的「利用遮罩增強的語言模型校正技術」中，MLM 使用專門訓練醫療文本的 Medical BERT，僅進行字符預測，過程為遮蔽與替代。我們進一步區分為常規遮罩和近音遮罩兩種方式。

3.1 常規遮罩 (Masking)

得到 MedCKIPBERT 後我們就可以，使用本研究提出的 BERT 輔助的遮罩語言模型 ASR 校正法，也就是 MLM 校正，此過程不涉及再訓練，而是使用已經訓練好的 BERT 模型來識別並替換文本中可能的錯誤。這個過程將某些

詞替換為[MASK] 標記，然後讓模型預測最可能的替代詞。通過這種方式，可以自動校正 ASR 系統生成的文本中的錯誤，特別是對於關鍵字和專有名詞的校正效果尤為明顯。

Algorithm 2 MLM Correction Using BERT

Require: *BERT model B , Input text* $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_T\}$
Ensure: *Corrected text* $\hat{Y}'$
Select tokens $\hat{y}_i$ *to be masked based on some criteria*
for *each selected* $\hat{y}_i$ *in* $\hat{Y}$ ***do***
  Generate masked of $\hat{y}_i$ *,denoted* $M(\hat{y}_i)$
  If *confidence* $(\hat{y}_i') > \alpha$:
    $M(\hat{y}_i) =$ *Compute predictions* $\hat{y}_i'$
  else:
    $M(\hat{y}_i) = \hat{y}_i' = \hat{y}_i$
end for
$\hat{Y}' = \{\hat{y}_1', \hat{y}_2', \cdots, \hat{y}_T'\}$
return $\hat{Y}'$

圖 4：MLM 校正 (Correction) 演算法

圖 4 具體步驟如下：

1. 訓練資料前處理與逐步遮罩：將 ASR 輸出的文本 $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_T\}$ 輸入輸入到 BERT 模型中。從序列中的第一個字符 $\hat{y}_1$ 開始做遮罩 $M(\hat{y}_1)$，其餘字符則不動。
2. 預測遮罩：使用 BERT 模型對遮罩後的文本 $\hat{Y} = \{M(\hat{y}_1), \hat{y}_2, \cdots, \hat{y}_T\}$ 進行預測，取最高機率的字符當作預測字符 $\hat{y}_i'$，當預測信心值 $confidence(\hat{y}_i')$ 有大於 α，才將預測結果 $\hat{y}_i'$ 取代 $M(\hat{y}_i)$，否則就保留原來的 token $\hat{y}_i$ 不動。α 在本研究的實驗中設為 0.9。
3. 迭代替代：將預測結果中的替代詞替換回原文本中 $\hat{Y} = \{\hat{y}_1', \hat{y}_2, \cdots, \hat{y}_T\}$，重複上述步驟，直到序列中的每個字符都被遮罩過。

3.2 近音遮罩 (Partial homophones masking)

我們發現在 ASR 的輸出中，多數錯誤源於「同音/近音異字」的錯誤辨識。為了改善這一問題，我們提出了一種近音遮罩技術 (Partial homophones masking)，如圖 5，該技術在遮罩替代時，會優先選取同音或近音中機率最高的作為預測結果。

在預測過程中，系統先判斷字符 (token) 是中文還是英文，並採取不同方法限縮預測範圍至同音/近音異字。對於中文，我們使用 python 函式庫 lazy_pinyin 將中文轉為拼音，並篩選出「同音不同調」的候選字，如「活」(huó) 和「或」、「霍」、「火」等字。對於英文，則使用 Metaphone 函式庫來尋找發音相似但拼法不同的字符，從而提高 ASR 系統的準確性。

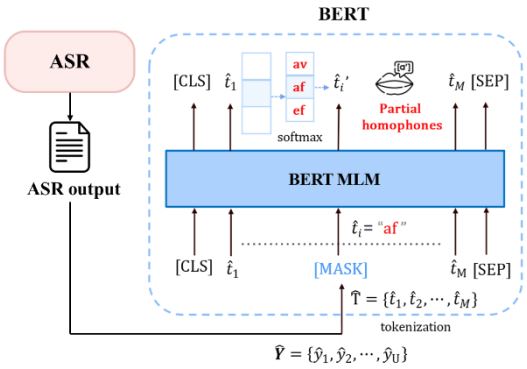

圖 5：近音遮罩框架

4 實驗與結果

為了驗證本論文所提的利用遮罩增強的語言模型校正技術，我們利用 GPT 生成一般醫療情境腳本，請人錄音而集成為醫療語音的測試集，然後利用 Google STT 以及 Whisper Medium 作 ASR 測試，然後對照使用 MedCKIPBERT 的語言模型進行常規遮罩以及近音遮罩的校正效果。

4.1 測試集

模擬醫療語音場景設定在包含專業術語的醫療對話中，由 3 位錄音員錄製了 354 句音檔，這些句子從 ChiMed250 的門診諮詢文本中選取，並利用 GPT 生成風格相似但內容不同的文本，如表 2 所示。有關利用 ChatGPT 生成技術，能利用有限範例文本生產具多樣性同風格的擬真病歷可參考 (Chung et al., 2024)。如此生成的內容保留醫療關鍵字，並確保測試集的獨立性和結果的可靠性，當然也不涉及任到病患的穩私性資訊。

一般醫療文本	鉤端螺旋體病會導致腸道黏膜的損傷，引起蛋白的吸收減少，另外還會導致貧血，會導致低蛋白水腫的情況，所以需要給予抗寄生蟲治療的同時還需要給予支援性的治療，比如給予大量的優質蛋白的攝入，另外需要補充大量的維生素和鐵既有利於貧血的改善，病情就可以逐漸的恢復。
GPT 生成風格相似文本	鉤端螺旋體病引致的腸道黏膜受損，會嚴重影響蛋白質的正常吸收過程，並可能導致貧血以及低蛋白血症引起的水腫現象，在針對寄生蟲的抗感染治療外，強烈建議配合相應的支持性治療措施。

表 2：使用 GPT 生成風格相似的文本

4.2 實驗結果

我們比較了 Google 和 Whisper Medium 的 ASR 系統在醫療語境中的辨識績效，並使用 MedCKIPBERT 進行遮罩增強語言模型校正。MedCKIPBERT 是在 CKIP BERT 基礎上經 WordPiece 切詞微調而成。我們測試了兩種校正技術：(1) 常規遮罩：取最高機率的字符作為預測，以提升準確性；(2) 近音遮罩：根據發音從同音不同調或類似發音候選中取最高機率字符。分別以 mask-MedCKIPBERT 和 homo-MedCKIPBERT 代表。

一般醫療語音	
Google	Whisper Medium
CER: 12.46% KER: 15.37%	CER: 9.62% KER: 20.82%
CER_NP: 8.36% CER_NP -CER: 4.10%	CER_NP: 6.06% CER_NP -CER: 3.56%
mask-CKIPBERT	
CER: 12.42% (-0.04) KER: 15.09% (-0.28)	CER: 9.92% (+0.30) KER: 20.82% (-0.00)
CER_NP: 8.28% (-0.08) CER_NP -CER: -4.14%	CER_NP: 6.22% (-0.16) CER_NP -CER: -3.70%
homo- CKIPBERT	
CER: 12.41% (-0.05) KER: 15.09% (-0.28)	CER: 9.91% (+0.29) KER: 20.82% (-0.00)
CER_NP: 8.26% (-0.10) CER_NP -CER: -4.15%	CER_NP: 6.21% (-0.15) CER_NP -CER: -3.70%

表 3：使用 CKIPBERT 串接 ASR 在一般醫療語音的組態效果總覽

首先比較了兩種不同的自動語音識別系統：Google、Whisper Medium 的測試結果，顯示如上表 3 的第一、二列。Google 儘管 CER 較差，但在關鍵詞錯誤率 (KER) 為最佳。相較而言，雖然 Whisper Medium 在 CER 上效果較佳，但其 KER 較差則顯示出在醫療關鍵詞的識別上明顯較弱。接下來加入語言模型的校正機制，先使用中研院做的 CKIPBERT 不經過微調，直接串接「BERT 輔助的遮罩語言模型 ASR 校正法」進行常規遮罩 (mask) 與近音遮罩法 (homo) 校正，可以看到效果有些微提升、甚或有些微降低，校正效果並不顯著。

CER_NP 是去除標點符號後的 CER，旨在更精確評估 ASR 系統處理純文字的表現。“CER_NP - CER”代表去除標點前後的 CER 差異。我們觀察到 Google 和 Whisper Medium ASR 常無法準確識別標點符號，導致後續 BERT 修正困難。因此，我們分析了去除標點後的字符錯誤率 (CER_NP)。

相較而言，表 4 為使用本研究在醫學領域下建構的 BERT 模型 MedCKIPBERT 來做校正，透過各種組態校正後，Google、Whisper Medium 各的 CER 和 KER 相較於表 3，只用一般文本的語言模型訓練的 CKIPBERT 來作校正的效果，均能有所改善。

一般醫療語音	
Google	Whisper Medium
CER: 12.46% KER: 15.37%	CER: 9.62% KER: 20.82%
CER_NP: 8.36% CER_NP -CER: 4.10%	CER_NP: 6.06% CER_NP -CER: 3.56%
mask-MedCKIPBERT	
CER: 12.00% (-0.46) KER: 13.64% (-1.73)	CER: 9.10% (-0.52) KER: 16.74% (-4.08)
CER_NP: 7.89% (-0.47) CER_NP -CER: -4.11%	CER_NP: 5.58% (-0.48) CER_NP -CER: -3.52%
homo- MedCKIPBERT	
CER: 11.97% (-0.49) KER: 13.64% (-1.73)	CER: 8.96% (-0.66) KER: 15.80% (-5.02)
CER_NP: 7.81% (-0.55) CER_NP -CER: -4.16%	CER_NP: 5.29% (-0.77) CER_NP -CER: -3.67%

表 4：使用 MedCKIPBERT 串接 ASR 在一般醫療語音組態總覽

在 Whisper Medium 經 homo-MedCKIPBERT 校正後，CER 與 KER 的表現效果最佳，CER 表現下降了 0.66%，CER_NP 表現下降了 0.77%，KER 下降了 5.02%。原本辨識效果就不錯的 Google 經過校正後也均能有更好的改善，在 homo-MedCKIPBERT 中效果為最佳：CER 下降了 0.49%，CER_NP 下降了 0.55%，KER 下降了 1.73%，但與一般遮罩的 mask-MedCKIPBERT 比起來，在統計上並沒有顯著的差異。相同的觀察也發生在表 3。

綜合來看，儘管我們的猜測大多數 ASR 錯誤來自於同音/近音異字，但近音遮罩與一般遮罩在校正 ASR 輸出效果上的差異，並不顯著。反而是 MedCKIPBERT 使用了大量預訓練文本切割的 WordPiece 提供更大的字典，這種分詞方法能捕捉更多語義信息，使得校正的效果優於僅用一般文本訓練的 CKIPBERT 模型的校正效果，有助於更準確地識別和校正字符錯誤，特別是與醫療術語相關的 KER 績效上能表現更好。

Ground Truth	盆腔炎是指女性盆腔內的生殖器官和周圍結締組織，包括子宮、輸卵管、卵巢以及盆腔腹膜等，發生炎症的情況。
Google CER: 12.46% KER: 15.37%	盆腔炎是指女性盆腔內的生殖器官和周遭結締組織、包括子宮、輸卵管、卵巢以及盆腔腹膜的、發生癌症的情況。
Google: homo- MedCKIPBERT CER: 11.97% (-0.49) KER: 13.64% (-1.73) CER_NP: 7.81% (-0.55) CER_NP -CER: -4.16%	盆腔炎是指女性盆腔內的生殖器官和周遭結締組織、包括子宮、輸卵管、卵巢以及盆腔腹膜的、發生炎症的情況。

表 5：使用 homo-MedCKIPBERT 串接 Google 在一般醫療語音實例

Ground Truth	盆腔炎是指女性盆腔內的生殖器官和周圍結締組織，包括子宮、輸卵管、卵巢以及盆腔腹膜等，發生炎症的情況。
Whisper Medium CER：9.62% KER：20.82%	盆腔炎是指女性盆腔內的生殖器官和周遭結地組織、包括子宮、舒卵管、卵巢以及盆腔附膜等、發生炎症的情況。
Whisper Medium: homo- MedCKIPBERT CER: 8.96% (-0.66) KER: 15.80% (-5.02) CER_NP: 5.29% (-0.77) CER_NP -CER: -3.67%	盆腔炎是指女性盆腔內的生殖器官和周遭結締組織、包括子宮、輸卵管、卵巢以及盆腔腹膜等、發生炎症的情況。

表 6：使用 homo-MedCKIPBERT 串接 Whisper Medium 在一般醫療語音實例

4.3 實例

以實際辨識測試中發生的句子為例，我們用三種不同顏色分別表達對同類型的錯誤：替換：紅色、插入：綠色<>、刪除：藍色刪除線、Keyword：螢光。在還沒做任何校正前，Google 與 Whisper Medium 在 CER 上的表現相差不多，然而在 KER 方面，Google 的表現明顯更佳。表 5 與表 6，分別為 Google 與 Whisper Medium 使用 homo-MedCKIPBERT 後的校正結果。CER 分別降低了 0.49% 與 0.66%，而 KER 的改善則更顯著，分別降低了 1.73% 與 5.02%。

5 結論

我們在實驗過程中觀察到：首先，在現有 AI 大廠於醫療語音識別績效的比較上，不論是一般語音還是病房交班語音，Whisper Medium 在字元錯誤率 (CER) 方面表現較佳，而 Google 在關鍵詞錯誤率 (KER) 方面更優。這反映了 Whisper Medium 擁有更好的聲學模型 (AM，Acoustic model) 和發音模型 (PM，Pronunciation model)，而 Google 則具備更強的語言模型 (LM，Language model)，這可能是由於其接觸了更多的醫學文本。

其次，MLM 的校正機制在都能改進一般 ASR 的績效，特別是在特殊專業領域中關鍵詞錯誤率 (KER) 方面。其根本原因在於，專業領域語音辨識的最大挑戰在於隱晦的術語，這些術語很難全面地收錄於訓練語音集中，這不僅增加了識別的難度，還會影響前後文的識別判斷。本論文提出的 MLM 校正機制有效提升了關鍵詞的辨識效果。相比於語音庫取得成本高，較全面性覆蓋專業領域術語的文本取得成本低廉，且語言模型的非監督式訓練成本也較低。

總結來說，本研究針對 ASR 在專業場域中的辨識績效欠佳以及專業領域語料庫蒐集困型校正技術」。該方法透過在 ASR 後串接針對特殊應用場域的語言模型作為遮罩語言模型進行校正，在特殊專業領域中有效降低了字

符錯誤率 (CER) 與關鍵詞錯誤率 (KER)，顯著減少了專有名詞和專業術語的辨識錯誤。

本研究的測試語音主是來自線上醫療文本，再經過生成式 AI 的 ChatGPT 所產生的腳本，再由人錄音而成。未來研究方向是針對門診中，醫師對病患的問認與應答，以及病房中之護理交班或是醫師巡房問答等醫療情境中的語音進行測試。

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., . . . Chen, G. (2016). *Deep speech 2: End-to-end speech recognition in english and mandarin.* Paper presented at the International conference on machine learning.

Cai, Z., Zhang, T., Wang, C., & He, X. (2021). *EMBERT: A pre-trained language model for Chinese medical text mining.* Paper presented at the Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). *Listen, attend and spell: A neural network for large vocabulary conversational speech recognition.* Paper presented at the 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP).

Chung, S.-L., Fan, J.-H., & Ting, H.-W. (2021). *Chinese Medical Speech Recognition with Punctuated Hypothesis.* Paper presented at the Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021).

Chung, S.-L., Lin, J.-Y., & Li, X.-Q. (2024). *Applying Generative Language Models to Generate Synthetic Medical Records: ChaVinci.* Paper presented at the Paper presented at the Proceedings of the 36th Conference on Computational Linguistics and Speech Processing (ROCLING 2024).

Enarvi, S., Amoia, M., Teba, M. D.-A., Delaney, B., Diehl, F., Hahn, S., . . . Pinto, J. (2020). *Generating medical reports from patient-doctor conversations using sequence-to-sequence models.* Paper presented at the Proceedings of the first workshop on natural language processing for medical conversations.

Fang, Z., Zhang, R., He, Z., Wu, H., & Cao, Y. (2022). *Non-Autoregressive Chinese ASR Error Correction with Phonological Training.* Paper presented at the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Jinhyuk, L. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*, 1234.

Mani, A., Palaskar, S., Meripo, N. V., Konam, S., & Metze, F. (2020). *Asr error correction and domain adaptation using machine translation.* Paper presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision.* Paper presented at the International Conference on Machine Learning.

Tian, Y., Ma, W., Xia, F., & Song, Y. (2019). *ChiMed: A Chinese medical corpus for question answering.* Paper presented at the Proceedings of the 18th BioNLP Workshop and Shared Task.

Udagawa, T., Suzuki, M., Kurata, G., Itoh, N., & Saon, G. (2022). Effect and analysis of large-scale language model rescoring on competitive asr systems. *arXiv preprint arXiv:2204.00212*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., . . . Wang, G. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.