The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Data centric approach to Chinese Medical Speech Recognition
以資料爲中心的中文醫療語音辨識技術開發

鍾聖倫 **Sheng-Luen Chung;** 李憶萱 **Yi-Shiuan Li**
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; m10807310@gapps.ntust.edu.tw

丁賢偉 **Hsien-Wei Ting**
衛生福利部臺北醫院神經外科
Department of Neurosurgery
Taipei Hospital, Ministry of Health and Welfare
Taipei, Taiwan
ting.ns@gmail.com

摘要

針對中文醫療語音辨識技術，本研究以資料爲中心的觀點 (data centric)，按照機器學習中開發與佈署 (MLOps) 的流程進行研究。首先，本研究按語義完整性切割出 Chinese Medical Speech Corpus (ChiMeS)；其次是語音辨識模型的優化：在固定 Joint CTC/Attention 的自動語音辨識 (Automatic Speech Recognition，ASR) 網路架構後，針對語料庫極端受限的挑戰，利用波形之資料增量提升辨識效果。整體來看，爲了促進中文醫療語音辨識的發展，本研究的具體貢獻有三，分別是：(一) 收集並標註 ChiMeS 語料庫，其爲時 14.4 小時，共 7,225 句語音。(二) 針對中文醫療語音辨識應用，訓練好的 Joint CTC/Attention ASR 模型，其在 ChiMeS-14 的測試集上的字符錯誤率 (Character Error Rate，CER) 和關鍵字錯誤率 (Keyword Error Rate，KER) 分別爲 13.65% 和 20.82%。以及 (三) 提供用來評估其他 ASR 模型績效的測試平台。細節請參 ChiMeS 入口網站 (https://iclab.ee.ntust.edu.tw/home)。

關鍵字：深度學習、中文醫療語音庫、語音辨識

Abstract

Concerning the deleopment of Chinese medical speech recognition technology, this study re-addresses earlier encountered issues in accordance with the process of Machine Learning Engineering for Production (MLOps) from a data centric perspective. First is the new segmentation of speech utterances to meet sentences completeness for all utterances in the colllected Chinese Medical Speech Corpus (ChiMeS). Second is optimization of Joint CTC/Attention model through data augmentation in boosting recognition performance out of very limited speech corpus. Overall, to facilitate the development of Chinese medical speech recognition, this paper contributes: (1) The ChiMeS corpus, the first Chinese Medicine Speech corpus of its kind, which is 14.4 hours, with a total of 7,225 sentences. (2) A trained Joint CTC/Attention ASR model by ChiMeS-14, yielding a Character Error Rate (CER) of 13.65% and a Keyword Error Rate (KER) of 20.82%, respectively, when tested on the ChiMeS-14 testing set. And (3) an evaluation platform set up to compare performance of other ASR models. All the released resources can be found in the ChiMeS portal (https://iclab.ee.ntust.edu.tw/home).

Keywords: Deep learning, Chinese medical speech corpus, Speech recognition

1 緒論

機器學習與運作 Machine Learning and Operations (MLOps)(Spjuth et al., 2021) 的概念是：針對以深度學習爲架構的任務，通常仰賴大量的資料進行訓練與測試、需要不斷的驗證以進行錯誤分析、並要針對不夠理想的結果設法提升效能，以最終能夠發展成爲可上市產品的系統化技術開發流程。本研究所要探討的目的即爲：以深度學習技術爲基礎，發展中

文醫療語音辨識成文字的技術發展過程中，所涉及到 MLOps 流程中幾個關鍵點的考量。

傳統語音辨識技術 (Oh et al., 2008)，主要建構在以聲學模型 (acoustic model，AM)、拼音模型 (pronunciation model，PM) 以及語言模型 (language model，LM) 三個模型的基礎上。這些與語言學專業相關的模型需要個別訓練，過程繁瑣複雜。近年來，多虧 Deep Speech(Hannun et al., 2014) 等基於深度學習之端對端語音模型的出現，新的語音辨識技術可以讓深度學習網路直接學習到語音和字符之間的對應關係，而不再像傳統技術中需要獨立訓練再整合的處理，並可得到優異的辨識效果。

目前語音辨識技術大多針對一般日常應用情境的語句，如語音助理、會議紀錄的轉譯等。針對此類應用，文獻上有相當數量的語料庫，可支援相關語音模型的訓練。大部份商用的語音辨識技術，透過大規模語料庫的訓練，在常用對談的情境中，可得到相當好的辨識效果。相較而言，對於類似像醫護專業的應用情境來說，文獻上或是商用產品均較缺乏相關醫學語料庫以及語音辨識技術的報導。然而，醫護專業場景中，如有自動語音辨識的輔助，可帶來很大的便利性。以護理交班爲例，醫護人員在下班前要將其所負責病患的病歷內容交接給另外一位護理人員。護理師除了口述之外，還需要手動輸入至電子健康紀錄 (Electronic Health Record，EHR) 中，手工作業繁瑣且費時。本中文醫療語音辨識技術研究的目的之一，即在於協助護理人員在口述交班時，利用語音辨識技術，將口述的內容直接轉譯成文字，輔助輸入 EHR，以降低護理人員紙筆輸入、手動輸入病歷、以及操作電子設備的時間。

發展中文醫療語音辨識技術所面臨到最主要的挑戰是缺乏專業語料庫。首先，以深度學習爲主的語音辨識技術，需要大量的資料集進行訓練與測試。而在專業醫療領域上，語料庫的收集更是困難，以致於目前在華語地區尚未有以中文爲主的醫學語料庫的發佈。其次，一般中文醫護的語音雖然以中文爲主，但多輔以英文的專業術語，內含雙語 (sentential code-switching) 的語料庫對於語音辨識是一項挑戰。爲了解決上述兩個問題，本研究遵循著 MLOps 的流程：定義問題、蒐集資料庫、訓練模型，以及將模型部署於其他場域，並以資料爲中心進行研究。本研究之貢獻在於：(1) 收集並標註一個語意完整的中文醫學語料庫 (ChiMeS)；(2) 針對此醫學語料庫額外使用波形增量 (Ko et al., 2015) 的方法，得到 CER 和 KER 分別爲 13.65% 和 20.82% 的 ASR 模型；(3) 並提供 ChiMeS 入口網站，提供語料庫以及公平測試的平台。

本論文以下第二節的文獻審閱分爲語料庫與 ASR 的技術演進進行介紹。第三節的 ChiMeS-14 語料庫會詳述資料集的錄製來源、訓練與測試之資料統計數字，還有本研究所提供的入口網站。第四節介紹 Joint CTC/Attention 語音辨識架構，以及波形增量的應用。第五節爲實驗結果與分析效能之討論。第六節則是結論與未來研究之方向。

2 文獻審閱

2.1 語料庫

針對不同領域的語音辨識任務，需要不同的訓練資料。以下介紹從 2017 年至文獻上所發佈，與中文醫療語音梢有關連的語料庫，包括在臺灣、大陸與新馬地區的華語語料庫，以及英文版本的醫學語料庫。

當作全世界最重要的語言之一，中文有相當的地域性。臺灣與中國雖然語言都使用中文，但在字體的運用、發音，以及字義的表示上，兩個地區都有所不同。臺灣最具代表性的中文語料庫，如 FSW 語料庫 (Liao et al., 2020)，其爲透過臺灣國家教育廣播電台的錄製，提出總共約 610 小時。此語料庫中含有 98,089 句以及 14,631,829 個字符。若是忽略不計算 735 個廣播新聞，大約至少有 120 人參與共 800 場訪談的錄音。本語料庫並當作 2018 之 Formosa Grand Challenge 競賽之用。2019 年時，Common Voice(Ardila et al., 2019) 收集了 38 種不同的語言，其中臺灣口音的中文涵蓋約 43 小時，到了 2020 年更收集到 78 小時，錄製人員也從 949 增加至 1,444 位。

另一方面，2018 年在中國所發佈的 AIShell-2(Du et al., 2018) 含有接近 100 萬個句子，包含了 1,000 小時的語音資料，總共有 1,991 位人員參與錄音，爲目前最大量的中文語料庫。在 2020 年間，DiDiSpeech(Guo et al., 2021) 收集了 6,000 人的語音，相較其他語料庫含更多錄音者，有更多元的音色。爲了滿足不同語音辨識任務的需求，將此語料庫切割成兩個不同的子集：DiDiSpeech-1 和 DiDiSpeech-2 。4500 位錄製者形成 DiDiSpeech-1 的 480,571 句子；而另外的 1,500 位錄製 171,361 句的 DiDiSpeech-2。

同樣在有許多華人的新加坡和馬來西亞，人們通常使用中英混雜的方式對談。2018 年釋出 SEAME(Lee et al., 2018)，爲 192 小時的語料庫，有 157 位錄音者，共有 162,290 句。

主要內容以採訪與對話爲主，在文本中以字(word-level)爲單位進行標註，並且含有語言標籤以及額外的六種分類標籤，分別爲目標語音、語助詞、其他語言、縮寫與專有名詞、口語化以及非語言訊號。

特別針對醫療語音辨識，2017 年時，Google (Chiu et al., 2017) 錄製超過 90,000 筆醫師與病患的醫療臨床對話，語料庫中最短的對話約爲 10 分鐘，而最長達到 2 個小時左右。此外，資料集中共有 100 名以上的醫師參與錄音，在訓練集與測試集的區分上，並不會有醫生的聲音重疊的情形，也就是出現在訓練集的醫生，就不會出現於測試集中。此英語內容的醫療語料庫極爲龐大，但並沒有釋出。

2.2 ASR 演進

傳統語音辨識技術中，kaldi(Povey et al., 2011) 的架構主要由：聲學模型、發音模型以及語言模型等三個模型所組成。其中，聲學模型主要是學習聲音上的特性；發音模型又稱爲詞典 (lexicon) 模型，其對應音素 (phoneme) 與形素 (grapheme) 的關係，協助聲學模型將發音映射到字形序列上；而語言模型則是學習文本中字與字之間的關係，提供傳統 ASR 語意上的幫助。由於這些模型需要個別訓練，因此需要仰賴語音的專業知識來對模型進行建模與優化。

近年來，以深度學習爲基礎的端對端語音辨識技術可直接學習如何語音特徵轉換成目標字符，免去了需要先針對傳統三個模塊個別建模與訓練的繁瑣流程，而有蓬勃發展。端對端語音辨識架構的本質是序列對序列的轉換，在將長度不固定的語音映射至可變長度輸出字符的技術上，主要分爲 Connectionist Temporal Classification(CTC) 和注意力 (attention) 方法。其中，爲了解決語音辨識輸入與輸出序列長度不固定的問題，CTC(Graves et al., 2006) 對每一幀的輸入都會有相對應獨立的輸出結果，透過 CTC 縮減的方式，將重複的字符先刪減，再移除空白標籤，最後得到短序列輸出的程序，來學習語音與相對應的標籤序列如何自動對齊。

另外一方面，基於注意力 (Attention) 機制的 ASR 也是透過編碼器和解碼器組成語音辨識架構：將輸入不固定長度的語音轉換成固定長度的語音特徵向量並得到隱藏狀態 (hidden state)，再利用基於 RNN 的解碼器，將語音的編碼器隱藏狀態解碼成序列表示的預測結果。和 CTC 機制不同的是，注意力機制的解碼器會同時參考編碼器中的所有隱藏狀態，可以捕捉時間跨度更長的前後文 (context) 的關係，如：拼音模型中的片語和語言模型的句型關係。注意力機制的代表架構如 Google 提出的 Listen, attend and spell (LAS)(Chan et al., 2016)，其中，稱爲 Listener 的金字塔型雙向 LSTM 將輸入序列轉變成高維特徵，而 Speller 在基於注意力的機制下，藉此學習前後文的關係，考慮過去輸入的情況下，輸出所有學過的字符機率分佈。

於 2017 年所發表的 Joint CTC/Attention (Kim et al., 2017) 語音識別模型結合 CTC 與注意力機制，提高架構的穩定性且使訓練能夠更快速收斂。藉由共用編碼器的方式，將聲學序列萃取後轉換成高維的特徵，在解碼器的部分則引用上述的 CTC 和注意力機制進行推斷以推測出最終結果。文獻上之後，也有許多針對基於結合這兩種方法的語音辨識架構，但主要是針對模型架構做調整與改善。舉例來說：(Zhu and Cheng, 2020) 的編碼器由雙向 LSTM (BLSTM) 組成的三角型架構組成，在沒有卷積層的情況下，來提升特徵萃取的能力。另外，(Li et al., 2019) 使用兩個獨立的編碼器先個別獲取聲音特徵以及各自的注意力分數，再將個別的注意力分數引入 Hierarchical Attention Network (HAN) 進行整合，得到更有效的訊息。

3 ChiMeS 語料庫與入口網站

語音資料庫採集不易，而專業領域情境下之語音更是如此。本研究針對 516 份住院病歷表朗讀語音檔，先按照語意完整性進行語句的切割與標註。之後，確認將此語料庫所切分出來的訓練集與測試集中錄音人員沒有重複。而對於評估語音辨識績效的指標，除了 CER 之外，由於專業語音中關鍵字相對重要，我們另外定義關鍵字錯誤率 Keyword Error Rate (KER)。相關語料庫、辨識評測標準以及工具一併公布至入口網站。

3.1 收集與標註

中文醫療語料庫 Chinese Medical Speech Corpus (ChiMeS) 共爲時約 14.4 小時，其由衛生福利部台北醫院的 15 位女性護理師，根據 516 份匿名化處理的住院病患病歷表，以交班時講話的方式進行語音的錄製，再經按完整語義切割與中英文譯文的標註而成。每份病歷表內容主要包括：匿名化病患之病史資料、入院狀態、目前病情，與每天狀況更新等四個部分。病歷來源包括：外科、泌尿科、耳鼻喉科、眼科、重症及安寧患者。此語料庫之音檔皆爲 wav 格式，採樣率 (sampling rate) 和取樣解析度 (bit resolution) 分別爲 16K Hz 及 16-bit，錄音文本則使用 UTF-8 編碼格式。

我們使用 ELAN(https://archive.mpi.nl/tla/elan) 標註工具，將專業護理師朗讀病歷表之語音，按照文本的語意完整性，以及大約時間長度爲 5 秒至 15 秒的條件下，將每份病歷表的朗讀切割成許多份語意完整的語音 (utterance) 檔案。由於病歷表中所記載之內容通常爲筆記型的摘要，不一定符合嚴謹的文法定義，我們儘量依據語意的完整性，先進行切分與文字標註。其中，語意的完整，指的是當語音內容描述完病患的一項狀態、一個診斷等完整的內容，即認定此爲完整的語意表達，進而做語音的分割，完成一句完整短語音，其流程如圖1所示。

語料庫中的字 (word) 或形素 (grapheme) 的標註，不僅決定 ASR 最終輸出字符的單位，也決定到最後 ASR 的績效評比，像是 CER 或是 WER 的計算。中文醫療語料庫標註的困難主要的原因在於其爲句中語系切換 (intra-sentential code switching)：雖然語音主體是中文句型，但文句中穿插有顯著比例的英文專業術語。針對句中語系切換語料庫的標註，不同語系的形素單位也有不同：在拼音系統中一般採用拼音字母，以英文而言，即爲 26 個英文字母，等同於 26 個形素。相對而言，中文字比較偏向表意符號 (ideogram)，每個中文字符本身即爲一個釋文代表。針對以字母標註的語料庫，ASR 除了要辨識各個字母，還要決定在連續字母之輸出後考慮是否爲一個單詞，並以空格做詞與詞間的區隔。而以表意文字標註的語料庫，不同的表意字符可能會發生同音字的情況，因此 ASR 架構需要仰賴前後文才能判定目前對應的字符輸出。爲了使 ASR 解碼一致，針對句中語系切換的 ChiMeS 語料庫，我們標註的原則是：中文以一個中文字符爲單位，而英文則以一個英文的單音節 (mono-syllable) 爲單位。本研究採用 How many syllables (https://www.howmanysyllables.com) 做爲將一個多音節的英文詞 (word) 拆成由數個由英文單音節所串成標註的依據。舉例來說：‘glucose ’ 分解成 ‘glu’ ‘cose’；而在英文縮寫上使用大寫表示，如 CRP。此外，由於病歷表中含有大量數值的內容，都使用中文標註，如：‘10.3’ 標註爲 ‘十點三’。

3.2 語料庫切分與評測依據

針對深度學習模型訓練與測試需要，ChiMeS 語料庫按約 4 : 1 的比例，切分爲錄音員不重覆的訓練集與測試。另外，音檔的命名方式以日期與第幾分病歷表爲標示，其次才爲此語句在整份病歷表中之編號，如：0522_01_1.wav。

配合跨科別測試，我們另外由 ChiMeS-14 的語料庫中，取出一個子集 ChiMeS-5，如表 1所示，其參與之錄製人員較少，病歷表涉及的科別涵蓋較小，只有外科、泌尿科、耳鼻喉科和眼科。而在人員錄製的病歷表份數上，訓練集中 01 和 02 號的份數不變，而編號 03、04 和 06 號護理人員，分別只含 86、33 和 9 份的病歷數；測試集的 05 號則是有 33 份的病歷表在子資料集 ChiMeS-5 中。在表 2爲 ChiMeS 的時長、句數、含中文字符與英文音節的總字數、平均時間長度等相關統計。值得一提的是，總字數的字符與音節分佈在 ChiMeS-14 中爲 167,409 和 48,110；而 ChiMeS-5 則爲 65,807 和 17,534。而不同字總數之字符和音節數在 ChiMeS-14 中爲 1,608 與 689；而 ChiMeS-5 則爲 1,268 和 553。

表 1: 語料庫分佈

語料庫	ChiMeS-14	ChiMeS-5
訓練護理師編號	{01~15}-{05,11,12}	{01~06}-05
測試護理師編號	{05,11,12}	05
訓練病歷表份數	394	166
測試病歷表份數	122	33

表 2: 語料庫細節

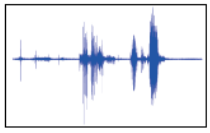
語料庫	ChiMeS-14	ChiMeS-5
時長	14.4hr (867mins)	5.5hr (335mins)
句數	7,225	2,987
總 tokens 數	215,519	83,341
不同 tokens 總數	2,297	1,821
平均時長 (secs)	7.2	6.7
平均 tokens 數	29.8	27.9
OOV 數	104	109

由於醫療語音辨識中，專業醫學術語的辨識上極爲重要，我們從 ChiMeS-14 中特別由人工萃取出 707 個與醫學相關的關鍵字，包括：病名、注射液、手術、傷口、藥物以及檢查項目，等六大類，如表 3 所示，各類醫學術語如：心臟病、limadol、扁桃腺切除、燙傷、vena 和核磁共振等中英文語詞。這些術語的正確辨識也將做爲專業語音辨識的績效依據。

表 3: 醫療關鍵字類型

分類	病名	注射液	手術	傷口	藥物	檢查項目	總數
數量	354	60	99	19	60	115	707

針對 ChiMeS 語料庫切分之測試集，我們



(a) 完整病歷表朗讀語音

女性六十七歲診斷是{co}{lon}CA沒有高跌沒有高壓沒有過敏史{on}{full}{diet} 這次就是因為{co}{lon}CA術後過去病史有{co}{lon}CA術後甲狀腺腫瘤骨刺膽結石開過刀 這次就是因為{co}{lon}CA術後在左{port}A左鎖骨有放一個左{port}A這次是為了第八次化療病那個{port}A回血都很不錯啦齁目前就是打五FU九點四然後在二十八號已經都打完了病人沒有什麼不舒服就辦理出院

(b) 標註文本

女性六十七歲診斷是{co}{lon}CA沒有高跌沒有高壓沒有過敏史{on}{full}{diet}

這次就是因為{co}{lon}CA術後過去病史有{co}{lon}CA術後甲狀腺腫瘤骨刺膽結石開過刀

這次就是因為{co}{lon}CA術後在左{port}A左鎖骨有放一個左{port}A

這次是為了第八次化療病那個{port}A回血都很不錯啦齁目前就是打五FU九點四

然後在二十八號已經都打完了病人沒有什麼不舒服就辦理出院

(c) 按語意切割語音

圖 1: 標註流程

使用兩種語音辨識的評判指標：CER、KER。其中，CER 的定義如式1 所示：

$$CER = \frac{S + D + I}{N} \times 100\% \qquad (1)$$

在上式中，N 爲 ground truth 中正確標註的字符數目，而 S 、D 和 I 分別對應預測結果中，對應：替換 (錯判)、刪除 (漏掉) 和插入 (新加) 等三項錯誤結果的字符數。其中，中文的字符以一顆顆中文字爲單位，而英文則使用英文單音節爲單位進行計算。舉例而言，‘glucose ’ 分解成 ‘glu’ ‘cose’，在計算時視爲兩個字符，這剛好跟其中文意譯的 ‘血糖’ 由兩個字符構成等權重。

除了 CER 的評測外，本研究也提供 KER 的計算，其公式如下式2：

$$KER = \frac{S_k + D_k + I_k}{N_k} \times 100\% \qquad (2)$$

KER 的計算概念與 CER 類似，針對表 3的關鍵字列表，我們將 ground truth 中正確標註以及預測結果中所出現的關鍵字進行對齊以比較差異。其中，N_k 爲所有測試集中的 ground truth 中所出現正確關鍵字的數量，而 S_k 、D_k 和 I_k 分別爲預測結果中被替換、刪除與插入的錯誤之關鍵字的數目。

3.3 ChiMeS 入口網站

爲促進中文醫學語音辨識技的提供，我們建置了 ChiMeS 入口網站 (https://iclab.ee.ntust.edu.tw/home)，並於其中公佈相關的語料庫、利用 ChiMeS-14 所訓練的 ASR 語音辨識模型，以及提供比較 ASR 模型績效的測試平台。

ChiMeS 入口網站共含五個部分：首頁、資料集、語音模型、評測以及線上實例展示。首頁包含後續頁面的簡介；語料庫分頁中，提供授權後可下載之 ChiMeS 語料庫的訓練集和測試集的音檔和標註的文本。評測平台提供 ChiMeS-5 與 ChiMeS-14 測試集，以及評估工具，可提供其他 ASR 的解決方案的測試績效。此外，我們也提供使用 ChiMeS-14 訓練之 Joint CTC/Attention 進行即時語音辨識。透過直接上傳音檔或是使用麥克風進行即時錄音，即可得到 ASR 轉譯結果。最後線上實例展示分頁中，一樣透過錄音或者上傳音檔的方式，即時得到辨識結果，並可額外針對結果進行修改。

4 醫療語音辨識模型

本研究參考 (Hori et al., 2017) 之 Joint CTC/Attention 架構做爲中文醫療語料庫的基線解決方案。此外，還額外使用波形增量的方法，在有限的醫療病歷語料庫下，提升語音辨識的正確率。本節將分爲三個部分介紹。首先，概述本實驗之語音辨識基準方案在訓練與測試的流程。接下來說明 Joint CTC/Attention 架構訓練方法。最後，詳述有關資料增量方法的產生與應用。

在語音辨識的流程中，首先將波形資料進行預處理，形成長度爲 T 的聲學特徵序列 $X = (x_1, ..., x_T)$，再透過語音辨識模型輸出所有字符分佈機率，並在最後選出最大機率之句子作爲對應長度爲 U 的辨識結果 $Y = (y_1, ..., y_U)$。在訓練階段，本研究採用 Joint CTC/Attention 網路，根據 ground truth 正確標註來計算由 CTC 與注意力模型的損失函數，調整網路參數，並在解碼過程中，考量注意力模型與 CTC 之間的比重，來組合兩種解碼機制。另外一方面，在測試階段，在考量以上架構的輸出時，也會針對 CTC 和注意力機制，進行權重比例的分配。最後是採用光束搜

索 (Beam search) 的解碼方式，從所有的可能性中選擇出一個機率最大的句子作爲預測結果。

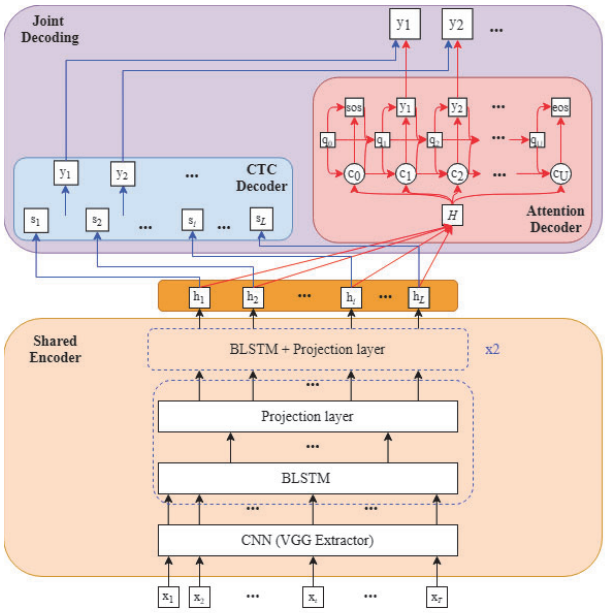


圖 2: Joint CTC/Attention

4.1 共同編碼器

我們將語料庫內的第 i 筆語音資料 w_i，其對應之逐字正確標註表示爲 y_i^*，這樣，語料庫中所有的資料可寫成 $W = \{(w_1, y_1^*), (w_2, y_2^*), .., (w_i, y_i^*)\}$。由於輸入 Joint CTC/Attention 網路架構的資料爲頻譜圖，因此將音檔透過快速傅立葉轉換 (Fast Fourier transform，FFT) 從時域轉換成頻域的形式，變成聲音特徵序列 $X = (x_1, ..., x_T)$。Joint CTC/Attention 網路架構共用一個編碼器，用來萃取語音特徵與學習時序關係，將長度 T 的序列 X 轉換成長度較短的高維表示，即編碼器的隱藏狀態爲 $H = (h_1, ..., h_L)$，如式3：

$$H = Encoder(X) \tag{3}$$

4.2 CTC 解碼器

CTC 機制能夠將長序列對應至短序列，並透過特殊的縮減方法來訓練語音模型。最主要的概念是針對每個語音幀都會有相對應的輸出字符，且不同字符的集合由字典 V 表示，其含中文字符與英文單音節，並有 blank 標籤 (<blank>) 代表發音模糊或是無發音的狀態。若要得知序列 Y 之機率，必須考慮刪減與移除前所有可能輸出 $S = (s_1, ..., s_L)$ 組合，公式如下式4：

$$p(Y|X) = \sum_{S \in Align(X,Y)} p(S|X) \tag{4}$$

在訓練階段，網路反向傳播誤差以用來更新權重參數。其中，CTC 的損失計算輸入序列與正確標註 Y^* 的差異，以在訓練時回傳與更新參數來學習對應關係。損失函數數值越小，代表此序列和正確標註越相似，模型學得越好。如式5所示：

$$L_{CTC} \triangleq -lnp(Y^*|X) \tag{5}$$

4.3 注意力解碼器

基於注意力機制之解碼器架構藉由式 6遞迴的方式，參考所有跨度的隱藏狀態 H，運算出在位置 u 的輸出字符分佈 y_u。如式 7所示，Attention 方法和 CTC 方法最大差別在於注意力除了將輸入聲音特徵序列 X 納入考量之外，也會參考過去輸出 $y_{1:u-1}$，以達到考慮前後文關係的完整序列 Y 預測。

$$y_u \sim AttentionDecoder(H, y_{1:u-1}) \tag{6}$$

$$p(S|X) = \prod_u p(y_u|X, y_{1:u-1}) \tag{7}$$

其損失函數計算如式 8，利用 cross-entropy 的準則學習編碼器的 $H = (h_1, ..., h_L)$ 和注意力解碼器的輸出 $Y = (y_1, ..., y_u)$ 兩種不同長度的序列。此外，爲了確保注意力編碼器能夠準確的學習到前後文的關係，使用強迫學習 (Teacher forcing) (Chang et al., 2019) 的方式，將過去的正確標籤序列 $y_{1:u-1}^*$ 作爲第 u 步的輸入資訊，進行網路的訓練。

$$L_{Attention} = -ln \sum_u p(y_u^*|X, y_{1:u-1}^*) \tag{8}$$

爲了計算出前後文向量 (Context Vector)，必須計算 location-based 的注意力權重 (Chorowski et al., 2015)，將編碼器的隱藏狀態 h_l 與基於注意力機制解碼器的隱藏狀態 q_{u-1} 融合，並透過所有跨度的隱藏狀態得出前後文向量。

最後，本研究所採用 Joint CTC/Attention 的 ASR 架構是將兩種不同的解碼方法使用 λ 參數調和，將 CTC 目標函式額外加入與注意力模型所組成的端對端架構一同進行訓練，藉此共享編碼器隱藏狀態。

$$L_{Joint} = \lambda L_{CTC} + (1 - \lambda) L_{Attention} \tag{9}$$

4.4 資料增量

針對基於深度學習之語音辨識模的訓練，收集大量的數據難度較高，而少量的訓練資料又會導致辨識效果無法達到最佳化。因此我們使用波形增量 (Ko et al., 2015)，針對原始訓練集音檔額外增加三倍音檔資料量，再將這些音檔轉換成頻譜的形式送進 ASR 進行訓練。

我們使用 Sound eXchange 語音編輯軟體 (http://sox.sourceforge.net/) 進行音頻的修改。爲了模擬不同人說話速度快慢、聲音高低和音量大小聲的差異，我們隨機在語速上調整 70% 至 120%；音高則是在-500 到 500 音分 (cent) 間；音量隨機放大 10dB 至減少 20dB 的範圍間；時移上則是移動 0 到 10ms 左右。除此之外，爲要強健模型面對吵雜語音的能力，參考文獻 (Amodei et al., 2016)，增加 10 到 15dB 噪音比的白噪音。透過隨機調整並組合以上五種聲音的特性，進而模擬不同人之聲紋與吵雜的背景音，增加訓練集資料的多樣性，以加強 ASR 的辨識能力。

5 實驗結果

爲了反映此模型的語音辨識能力，我們針對 ChiMeS 語料庫進行模型的訓練與測試，並比較使用波形增量前後之辨識結果。本研究實驗中有關 ASR 的參數設置，編碼器中的特徵萃取部分爲 VGG 萃取器之 4 層 CNN 外加兩層池化層的組合；而學習語言關係的部分則是由單層共有 640 個細胞 (cell) 的雙向 LSTM 共 3 層組合而成，且在每一層雙向 LSTM 後都會接續一層線性投影層 (linear projection layer)。而注意力解碼器則爲一層單向擁有 320 細胞的 LSTM 架構。Joint CTC/Attention 在訓練時，使用 0.0001 的學習率 (learning rate)、批量大小 (batch size) 爲 24，以及 Adam 優化器 (optimizer) 搭配梯度裁減的設置，CTC 和注意力之權重數值設爲 0.5。在測試時，光束寬度使用 6，來進行最佳結果的選擇。

爲了清楚說明不同組合條件下的實驗結果，我們引入以下的命名方式，以標註：(1) 所採用的 ASR 網路架構、(2) 針對哪一個語料庫進行訓練、(3) 訓練過程中是否有使用數據增量、以及 (4) 最後是否在相同語料庫的測試集上做測試。實驗的命名方式爲：ASR 模型 _ ChiMeS_ 資料增量/測試於不同語料庫的測試集上。舉例來說：使用 Joint CTC/Attention 模型，訓練 ChiMeS-14 並搭配波形增量，命名爲 JCA_14_w；若是在沒有使用增量的情況下訓練 ChiMeS-5 訓練集，並測試於 ChiMeS-14 測試集，則爲 JCA_5/14。

針對 Joint CTC/Attention 架構中兩種解碼機制，我們比較不同權重下所訓練模型的語音辨識效果，如表4。在只使用注意力方法的條件下，也就是 $\lambda = 0$ 時的效果極差，原因在於 ChiMeS 多以長句子爲主，若是前面出現辨識錯誤，易出現後續連續錯誤的情形。而在同時包含 CTC 和注意力解碼機制下，ASR 的預測相差不多，對於 λ 值並不敏感。最後，只包含 CTC 的方法，也就是 $\lambda = 1$ 時，因爲缺乏輸出間的語意關係，結果會略差於融合兩種解碼機制的結果。

表 4: 實驗結果

λ	JCA_14	JCA_14_w
0	92.01%	83.41%
0.2	21.57%	16.71%
0.5	19.82%	13.65%
0.8	21.51%	17.39%
1	33.49%	19.97%

受限語料庫大小有限的緣故，我們也比較有無使用波形增量的 ASR 辨識率，如表 5 所示，當使用波形增量訓練 Joint CTC/Attention 時，其 CER 可由不使用波形增量的 19.82%，大幅減少 6.17%；同時，在 KER 也可以展現出 10.82% 的改善。

表 5: 增量前後之實驗結果

評測標準	CER	KER
JCA_14	19.82%	30.90%
JCA_14_w	13.65%	20.82%

我們利用表 6 展示測試集中第 11 號錄音者所錄製的其中一份病歷表辨識結果，比較有無使用波形增量之效果。並將結果中翻錯、多翻和少翻分別使用 紅色、藍色刪除線 和綠色<>表示。

6 結論

本研究提出一個由專業護理人員所緣製的中文醫療語料庫 ChiMeS，並提供一使用端對端的深度學習架構，在原始的訓練集下，利用波形增量的方法，加入額外音檔，有效提高語音辨識的效能。

未來的研究有三個方向：首先是針對醫學語料庫進行更全面的收集：不只要包含更多錄製人員，更要增加不同科別的病歷表內容，豐富語料庫的資料，以有助於在新佈署之場域辨識效果的穩定性。再來嘗試使用其他較新的端對端語音辨識架構，例如：Transformer(Dong et al., 2018)、Conformer(Gulati

表 6: 增量前後之實例

實驗	文字	CER
Ground Truth	{co}{lon}{can}{cer} 沒有高跌沒有高壓沒有過敏史 DM{diet} 一天一千五百卡有 DM 腹膜炎 {co}{lon}{can}{cer} 過去病史 然後此次是因爲發現 {co}{lon}{can}{cer} 尚未開刀 在左鎖骨放 {port}A 預計行第二次化療入院 左邊有一條 {port}A 到十月三號 {su}{gar} 測 QIDAC 沒事	-
JCA_14	<{co}>{com}{can}{cer} 沒< 有 >高跌沒有高壓沒有過敏史 D<M>{diet} 一天一千五帶卡有 DM 腹後炎<{co}><{lon}>看開過去病史 < 然 >< 後 >讓他是因爲< 發 >排線 {con}{can}{cer}上胃開大 內科說不放懷A 預計形 DN 吃排量入院 < 左 >這 B一條{po}A{gas}十A三{per}{su}{gar} 測 QIDAC名 {tive}	41.94%
JCA_14_w	{co}{lon}{can}{cer} 沒有高跌沒有高壓沒有過敏史 DM{diet} 一千一千五百卡有 DM 腹膜炎 {co}{lon}{can}{cer} 過去病史 < 然 >他此< 次 >日因爲發現 {co}{lon}{can}{cer}從維採大 乁左鎖骨放 {port}AA 底性第二次化療入院 左邊< 有 >一條痛A 到十月三號 {su}{gar}< 測 >QIDAC 沒事	15.97%

et al., 2020) 等提升辨識效果。最後，面臨醫療術語的多樣化，利用轉移學習 (Kunze et al., 2017) 的方式加快訓練速度，得到更佳的效果。

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Xuankai Chang, Yanmin Qian, Kai Yu, and Shinji Watanabe. 2019. End-to-end monaural multi-speaker asr system without pretraining. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6256–6260. IEEE.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, et al. 2017. Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, et al. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6968–6972. IEEE.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. *arXiv preprint arXiv:1706.02737*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. 2015. Audio augmentation for speech recognition. volume 2015-January, pages 3586–3589.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. 2017. Transfer learning for speech recognition on a budget. *arXiv preprint arXiv:1706.00290*.

G. Lee, T.-N. Ho, E.-S. Chng, and H. Li. 2018. A review of the mandarin-english code-switching corpus: Seame. volume 2018-January, pages 210–213.

Ruizhi Li, Xiaofei Wang, Sri Harish Mallidi, Shinji Watanabe, Takaaki Hori, and Hynek Hermansky. 2019. Multi-stream end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:646–655.

Yuan-Fu Liao, Yung-Hsiang Shawn Chang, Yu-Chen Lin, Wu-Hua Hsu, Matus Pleva, and Jozef Juhar. 2020. Formosa speech in the wild corpus for improving taiwanese mandarin speech-enabled human-computer interaction. *Journal of Signal Processing Systems*, 92(8):853–873.

Y.R. Oh, M. Kim, and H.K. Kim. 2008. Acoustic and pronunciation model adaptation for context-independent and context-dependent pronunciation variability of non-native speech. pages 4281–4284.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

O. Spjuth, J. Frid, and A. Hellander. 2021. The machine learning life cycle and the cloud: implications for drug discovery. *Expert Opinion on Drug Discovery*.

Tao Zhu and Chunling Cheng. 2020. Joint ctc-attention end-to-end speech recognition with a triangle recurrent neural network encoder. *Journal of Shanghai Jiaotong University (Science)*, 25(1):70–75.