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

# Chinese Medical Speech Recognition with Punctuated Hypothesis
# (釋文內含標點符號之中文醫療語音辨識技術)

鍾聖倫 **Sheng-Luen Chung;** 范晉桓 **Jin-Huan Fan**
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; m10807507@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

語音辨識可協助醫護專業人員節省其使用電子醫療系統的文書處理時間。中文醫療語音的內容爲中文爲主，之間摻入以英文的專業術語，所以本質上可視爲句中雙語 intra-sentential code switching speech corpus。先前在中文醫療語料庫 (Chinese Medicine Speech Corpus: sChiMeS) 上進行的語音辨識的譯文並沒有標點符號，不易人的閱讀，也不利後續機器翻譯等應用。據此，本論文提出一階段即可回復標點符號的語音辨識技術。爲此，在資料庫的準備上，將原本沒有標點符號標註的 ChiMeS 語料庫的標註加入標點符號成爲 psChiMeS，然後進行訓練。同時，爲取得更好的語音辨識效果，我們採用 ESPnet 框架中以 conformer 爲基礎並且彙整 CTC 辨識機制的 ASR 網路架構。在 psChiMeS-14 的語料庫上，本論文所提的方法得到 10.5% 的 CER 以及 13.10% 的 KER。對比之前在 Joint CTC/Attention 架構上最好的結果爲：15.70% 的 CER 以及 22.50% 的 KER。本技術可作爲發展線上醫療語音辨識系統的基石。

關鍵字：深度學習、語音辨識、醫療語料庫

## Abstract

Automatic Speech Recognition (ASR) technology presents the possibility for medical professionals to document patient record, diagnosis, postoperative care, patrol records, and etc. that are now done manually. However, earlier research aimed on Chinese medical speech corpus (ChiMeS) has two shortcomings: first is the lack of punctuation, resulting in reduced readability of the output transcript, and second is the poor recognition error rate, affecting its application put to the fields. Accordingly, the contributions of this paper consist of: (1) A punctuated Chinese medical corpus psChiMeS-14 newly annotated from ChiMeS-14, which is the collection of 516 anonymized medical record readouts of 867 minutes long, recorded by 15 professional nursing staff from Taipei Hospital of the Ministry of Health and Welfare. psChiMeS-14 is manually punctuated with: colons, commas, and periods, ready for general end-to-end ASR models. (2) A self-attention based speech recognition solution by conformer networks. Trained by and tested on psChiMeS-14 corpus, the solutions deliver state-of-the-art recognition performance: CER (character error rate) 10.5%, and KER (Keyword error rate) of 13.10%, respectively, which is contrasted to the 15.70% CER and the 22.50% KER by an earlier reported Joint CTC/Attention architecture.

***Keywords:*** deep learning, speech recognition, Chinese medical speech corpus

## 1 緒論

### 1.1 動機

醫療語音辨識有助於醫療專業人員進行病歷紀錄、巡房與診斷追蹤等。相較於一般大眾日常會話或是智慧家庭週邊商品所考量的情境語音，醫療語音的特殊性在於其中英文混雜的術語、筆記式的片斷句型、以及區域性醫護人

員特殊的發音等特性。這些挑戰致使醫療情境中的語音無法直接使用一般語音辨識技術當作解決方案。

針對中文醫療語音辨識技術，先前的研究成果的貢獻有三項，分別是：(一) ChiMeS 語料庫，其爲時 14.4 小時，共 7,225 句語音。(二) 訓練好的 Joint CTC/Attention ASR 模型，其在 ChiMeS-14 的測試集上最好的字符錯誤率 (Character Error Rate：CER) 和關鍵字錯誤率 (Keyword Error Rate：KER) 分別爲 12.85% 和 17.62%。以及 (三) 評估其他 ASR 模型針對醫療語音辨識的基本績效測試平台。

大多數的語音辨識模型，在訓練時不包含標點符號。然而，由 (Garg et al., 2018)、(Zhang and Zhang, 2020)、(Li et al., 2021) 所提出的文獻可以得知，標點符號不但能提高可讀性，同時也有助於翻譯的效能提升。另一方面，近兩年新的語音辨識技術的出現，也能對當前對 ChiMeS 的辨識績效更多提升的空間。上述內容統整出目前特別針對中文醫療語音辨識的兩項挑戰：(1) 先前的醫療語音辨識技術並沒有考量標點符號，導致辨識出的文本可讀性低；(2) 過去所採用的醫療語音辨識模型的 CER 仍有改善的空間。據此，本論文的貢獻如下：

(1) 含標點符號標註之中文醫療語料庫 (psChiMeS)：

參考教育部所頒定《重訂標點符號手冊》並依照醫療文本特性適時調整規則，我們針對 ChiMeS 的文本重新進行標註，得到 psChiMeS，其可作爲後續針對中文醫療音辨識之端對端訓練的語料庫。

(2) 到目前爲止辨識績效最好的醫療語音辨識模型 Joint CTC-Conformer：

利用基於自我注意力機制 (Self-Attention) 和卷積網路 (Convolution) 機制的 Joint CTC-Conformer 語音辨識模型，在 psChiMeS 上的訓練與測試，實現了 10.5% 的 CER，以及 13.10% 的 KER，對比之前報導的 Joint CTC/Attention 經數據增強後所得的 15.7% 的 CER 和 22.50% 的 KER。

### 1.2 論文架構

本論文的第二節回顧自動化標註標點符號的技術和近年語音辨識模型的架構發展。第三節詳細介紹目前較先進的醫療語音辨識模型的相關技術與運作流程。第四節爲實驗結果與分析討論，包含語音辨識的效能比較；以及強調加註標點符號語音辨識的效能。最後，第五節爲本研究之結論與未來研究方向。

## 2 文獻審閱

### 2.1 自動化標註標點符號

針對語音辨識的釋文額外需要內含標點號的問題，文獻上主要作法爲兩階段，即先取得不含標點符號的文字串，然後再加上標點符號。以下先介紹對應這第二階段的自動化標註標點符號技術。之後再介紹一般端對端的 ASR 技術。

文獻上，自動化標點符號標註是將沒有標點符號的句子或 ASR 輸出的結果，進行額外填入語意上需要停頓的標點符號還原 (punctuation restoration)，也就是將不含標點符號的文本輸入模型後，輸出爲含標點符號的文本。2017 年 (Salloum et al., 2017) 提出針對醫學報告的文本進行標點還原，其所採用的技術是使用 RNN 加上 Attention 的機制來進行標點符號恢復的任務。其作法是將相同字首或字尾的詞彙更換成統一表示方式來降低訓練詞彙量。此技術在逗號、句號、冒號的標點回復效果上都取得不錯的效能。2018 年由 (Garg et al., 2018) 所提出的自動標註標點符號模型，目標是爲了讓電子教學平台影片在做語言翻譯時，可以先進行標點符號的預測再送入翻譯模型，以提升翻譯品質，其提出的 CNN 加上雙向 LSTM 預測模型有最佳的效能。

2020 年 Amazon 針對醫療語音的辨識任務，進行標點符號的預測以及正確大小寫 (Sunkara et al., 2020) 的恢復。文中使用 BERT (Devlin et al., 2018) 預訓練模型對其任務進行微調，此架構用兩階段方式來訓練，最後在標點符號預測及恢復正確大小寫的任務上與 LSTM 架構相比，效能提升 3 ∼ 4%。同樣在 2020 年，百度 (Zhang and Zhang, 2020) 爲了解決即時翻譯沒有輸出標點符號無法找出句子邊界的問題，使用 ERNIE (Zhang et al., 2019) 預訓練模型，以多分類的訓練方式來預測句子的邊界，最後 ASR 在搭配此模型運行下，BLEU 翻譯評測指標 (Papineni et al., 2002) 可以提升大約 2%。2020 年 Google 提出在翻譯實驗中 (Li et al., 2021)，透過句子邊界增量 (Sentence Boundary Augmentation)，在多個資料集上 BLEU 都有所提升，證明好的句子邊界不僅能提高可讀性，也能使翻譯品質有所提升。

### 2.2 端對端 ASR 模型

#### 2.2.1 CTC

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，並學習語音與相對應標籤序列中如何自動對齊。2017 年由百度提出的 Deep Speech 2 (DS2)(Amodei et al., 2016)，編碼器是由 CNN 與雙向 LSTM 組成，用來提取每個語音時序上的前後文關係。摘要前後文資訊的的隱藏層狀態 (hidden state) 會被用來當作後續 CTC 輸出的依據。另一方面，解碼器爲單純一層全連接層 (fully connected layer)，不再需要額外使用詞典將音素映射至形素。端對端 ASR 的架構只需要一個架構進行訓練即可將語音轉換成文字。然而 CTC 的缺點是：針對每個輸入幀都被視爲獨立的對應輸出，所以常需搭配語言模型，來補足前後文關係較不足的問題。

#### 2.2.2 Joint CTC-Attention

2017 年由 (Kim et al., 2017) 等學者提出同時具備 Attention 機制和 CTC 的 Joint CTC-attention 語音辨識模型。此模型主要分成三個部分：第一，Encoder 採用共用形式，主要由 CNN 和 LSTM 組成並在解碼階段將所有時序的隱藏層狀態輸出至 CTC Decoder 和 Attention Decoder。第二，CTC Decoder 在每一時間下的音頻輸入與對應字符爲條件獨立，因此在每一幀的音源輸入皆會有對應的文字輸出，最後以刪除重複的字和空白標籤 (Blank) 的縮減方式來對齊。第三，Attention Decoder 透過 attention 分數運算來獲取上下文關係。最後，透過合併使用這兩種 Decoder 的優點，並以 $\lambda$ 來當作損失函數的調和參數，如式 1 所示，這樣所獲得得 Joint CTC-attention 模型能在當時達到更好的語音辨識效果。

$$L_{joint} = \lambda L_{CTC} + (1-\lambda) L_{Attention} \quad (1)$$

#### 2.2.3 Self-Attention 機制用於 ASR 模型

在 2017 年，Google 提出 Transformer (Vaswani et al., 2017)，此模型是一個加入自我注意力機制 (Self-Attention) 的 Seq2Seq 模型。採用此機制的模型在訓練時可以對所有輸入的時序資料進行矩陣的平行運算，加速訓練。Transformer 在翻譯任務和 NLP 任務等時序資料處理的問通，都有很好的表現。之後，有學者將 Transformer 模型引入語音辨識。在 2018 年由 (Dong et al., 2018) 等學者首先提出 Speech-Transformer，其使用 Transformer 架構來取代 RNN 或 LSTM 等 Seq2Seq 模型，不僅在 Wall Street Journal (WSJ) 語料庫上可以取得到字錯誤率 (Word Error Rate：WER ) 10.9% 的效能，所需的訓練時間也僅爲原本使用 RNN 或 LSTM 架構的 30%。

在 2020 年由 (Miao et al., 2020) 等學者所提出將 CTC 機制合併加入 Transformer 架構中，此架構可以保有 Transformer 在擷取大跨度的上下文特徵優異的特性，同時也結合 CTC 每一時間下預測的獨立性。最後在 HKUST 中文語料庫上，相較於 Joint CTC-Attention 模型的 CER 降低了 4%。

雖然 Transformer 對於大跨度的上下文特徵有不錯的表現，但對於局部特徵擷取較爲弱勢。因此 2020 年由 Google 另外在 Transformer Encoder 中加入 Convolution 機制，稱作 Conformer Encoder (Gulati et al., 2020)，其能有效的擷取全局與局部的音頻特徵。在搭配一層的 LSTM Decoder 架構下，可以在 LibriSpeech 語料庫取得 WER 2.1% 的效能；而搭配語言模型更取得 WER 1.9% 的效能。

## 3 研究方法

### 3.1 一階段直接標註標點符號的語音辨識技術

相較於一般文獻均是採用兩階段將語音轉譯成有附加標點符號的方法，本論文直接採用更爲簡潔與迅速的一階段訓練方式，如圖 1 所示。據此，我們首先針對原沒有加標點符號 ChiMeS 語料庫中，原音檔 $w_i$ 所對應原沒有加標點符號的文本標註，重新進行人補註標點，加上包括：冒號、逗號與句號的文本，我們以 $y_i^*$ 表示。接下來是建置同時合併使用 Self-Attention 機制的 conformer，以及 CTC 編碼器而成 Joint CTC/Conformer 架構的 ASR 網路。在此端對端的新架構上，直接使用有加標點符號的譯文 psChiMeS 作爲訓練的語料庫，而直接將額外標註的標點符號 (冒號、逗號與句號) 視同原本輸出字符字典中與中文字和英文單音節同樣的字符來處理，當作整個 ASR 網路的輸出字符處理。

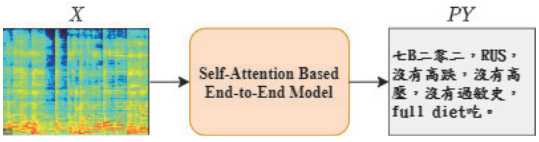


圖 1: 一階段訓練示意圖

### 3.2 重新標記之具標點符號語料庫

在本研究中，我們所採用的中文醫療語料庫 Chinese Medical Speech Corpus (ChiMeS)

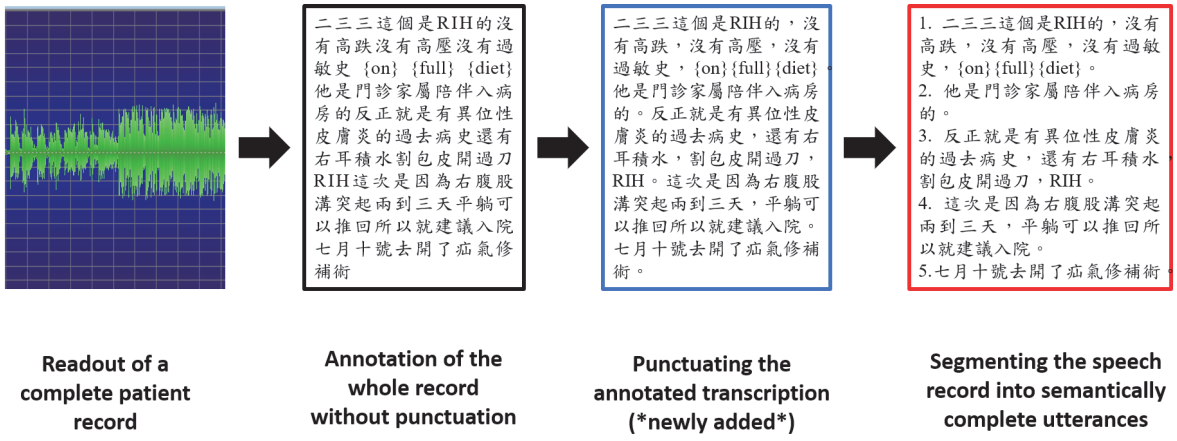


圖 2: psChiMeS 文本標註方式

共爲時約 14.4 小時，其由衛生服利部台北醫院的 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，再按完整語義切割與中英文譯文的標註。如圖 2 所示，在標註時，我們是先將病歷表的朗讀檔先進行標註，得到沒有標點符號的譯文 (黑框處)。接著我們參考了教育部的標點符號手冊並且根據醫療文本的特性訂定了醫療文本標點符號規則，如表 1。將先前所標註的譯文加入標點符號，得到具標點符號的譯文 (藍框處)。

每一份病歷表的內容包括：病患匿名化之基本資料、入院狀態、目前病情，與每天狀況更新等四個部分。針對深度學習模型訓練與測試需要，我們透過前後文語意的完整性，將完整的病歷表譯文切割成多個句子 (紅框處)。最後，將 ChiMeS 語料庫切分爲訓練集與測試集，兩者的比例分配約爲 4：1，得到訓練集爲 5,682 句，而測試集爲 1,543 句。訓練集與測試集中沒有彼此重複的護理師。詳細的語料庫分布表如表 2, 其中 psChiMeS 爲含標點符號語料庫。

表 1: 標點符號標註規則

| 標點符號 | 規則 |
|---|---|
| 逗號 (，) | 前後意思銜接 |
| | 用來隔開檢查數值或是病名 |
| 冒號 (：) | 接下來要表示各個檢查數值 |
| | 接下來要敘述病人過去病史 |
| 句號 (。) | 語意上，敘述結束 |
| | 切分不同日期之間交班紀錄 |

表 2: sChiMeS 和 psChiMeS 資料分布表

| 語料庫 | sChiMeS | psChiMeS |
|---|---|---|
| 特性 | 語意完整 | 語意完整<br>有標點符號 |
| 句數<br>(訓練/測試) | 7,225<br>(5,682/1,543) | 7,225<br>(5,682/1,543) |
| 平均時長<br>(秒/句) | 7.2 | 7.2 |
| 平均字數 | 29.8 | 33.3 |
| 總時長<br>(分數) | 867.86 | 867.86 |

### 3.3 Joint CTC-Conformer 模型

如圖 3 所示，Joint CTC-Conformer 是由 Conformer 編碼器，搭配 CTC 解碼器和 Transformer 解碼器所組成，其中編碼器的部分是由多個 Conformer Block 組成，每個 block 中依序包含前饋式網路 (Feed-Forward Netword)、多頭式注意力機制、卷積模組以及另一層前饋式網路。以下介紹 Joint CTC-Conformer 中的每個模組。

#### 3.3.1 多頭式注意力機制 (Multi-Head Self-Attention，MHSA)

Self-Attention 的運算方式爲縮放點積注意力機制 (Scaled Dot-Product Attention)，其將每一時序特徵透過三個不同的線性層 (linear layer) 分別轉換爲 $Q$、$K$、$V$ 後，送入縮放點積注意力機制。而縮放點積注意力機制運算如式 2，透過平行運算方式同時將所有時序的 $Q$ 對各別時序的 $K$ 兩兩做點積，接著將點積結果除以縮放因子 $\sqrt{d_k}$，然後再送入 softmax 得到相加爲 1 的注意力權重 (Attention Weight)，最後再使用此權重乘上 $V$ 得到輸出

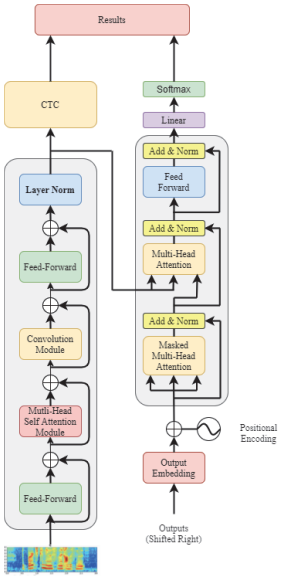


圖 3: Conformer ASR 架構圖

結果，再將每個時序的輸出結果合併就可得到注意力圖 (Attention Map)，其中使用縮放因子 $\sqrt{d_k}$ 的目的是在於讓 $Q$ 和 $K$ 點積後的數值不會因爲 $Q$ 和 $K$ 的維度太大，而造成數值過大，進而影響 softmax 的運算導致梯度變小，影響訓練的結果。

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

多頭式注意力機制，如式 3 所示，是由多個縮放點積注意力機制所組成。同樣將每一時序特徵透過三個不同的線性層 (linear layer) 分別轉換爲 $Q$、$K$、$V$，接著再進一步將 $Q$、$K$、$V$ 各別切分成 $h$ 等分。多頭的概念是類似 CNN 卷積層中各通道對應的 convolution kernel 的效果，而每一顆頭 (head，h) 會各別關注不同來自不同時序的資訊，然後將每個時序切分後的 $Q$、$K$、$V$ 序列送入縮放點積注意力機制進行平行運算最後得到注意力圖。其中 $W^Q$、$W^K$、$W^V$ 爲參數矩陣，由訓練所得到。

$$MHSA(Q, K, V) = Concat(head_1, ..., head_H)W^o$$
$$where\, head_h = Attention(QW_h^Q, KW_h^K, VW_h^V) \quad (3)$$

### 3.3.2 卷積模組 (Convolution Module)

Conformer 中的 Convolution Module，有優異的局部特徵擷取能力，可以將相鄰語音的前後文特性塑模出來，像是能有效地重組醫療相關的關鍵詞，其架構主要先經由一層的點卷積層 (1-D pointwise convolution) 可以將通道 (channel) 加倍，然後依序是 GLU 激勵函數、一維深度卷積層 (1-D depthwise convolution)、Batch Normalize，接著使用 Swish 激勵函數，最後在接上點卷積層。

### 3.3.3 Self-attention 解碼器

此節我們說明 Self-Attention 解碼器的損失函數。來自編碼器的隱藏層狀態，經過解碼器多層的 Self-Attention 之後，會透過線性層與 softmax，輸出每個時序的聲學特徵序列與相對應字典中每個字符的預測機率。接著將 Ground Truth 轉換爲 one-hot 編碼形式，最後再與每個時序的預測機率進行交叉熵損失函數 (cross-entropy loss) 計算。公式如式 4 所示：

$$L_{self-attention} = -\sum_{u=1}^{U} y^* ln(y_u) \quad (4)$$

其中，$y^*$ 表示爲正確標註 Ground Truth，$y_u$ 爲與 $y^*$ 相同字符的預測機率。最後將所有時序的計算結果相加。由於訓練時希望損失函數 (Loss Function) 最小化，因此取負對數 (Negative logarithm) 來計算損失函數。

### 3.3.4 CTC 解碼器

接著介紹 CTC 如何將隱藏層狀態 $H$ 預測後的長度縮減成和 $y^i$ 的長度相同。令 $U$ 爲語料庫中所有預測的字典集合，包含中文字以及英文 Syllable 的所有類別，在 CTC 中會另外在字典中加入 Blank (-) 用來表示無發音、發音不清晰或模糊的類別。而語音辨識模型對於每一時序的輸入都會有相對應的輸出，以本研究爲例，即由 Self-Attention Based Encoder 所輸出每一時序的 hidden state 都會得到相對應的類別概率 $p_t(s_t|x)$，其中 $s_t$ 是在 $t$ 時刻從字典 $U$ 中所預測的字符。透過網路輸出將各時序所預測的類別概率相乘可以得到各預測字串的機率，如下式 5 所示：

$$p(S|x) = \prod_{t=1}^{T} p_t(s_t|x) \quad (5)$$

其中 $x$、$T$ 分別代表語音特徵序列和時序長度；$p_t(s_t|x)$ 表示爲在第 $t$ 個時序所輸出爲某字符的機率，$S$ 爲預測字串，而 $p(S|x)$ 則可理解成每一時序預測字符機率相乘後的預測字串的機率，也就是將每一時序所得到的 $p_t(s_t|x)$ 機率相乘所得到的預測字串機率。

由於 CTC 每一時序皆有相對應的輸出，其可能含有很多重複的預測字符與 blank，使得輸出字串的長度遠大於 $y_i^*$ 的長度，因此在 CTC 中會透過特殊的刪減機制讓預測字串長度更接近 $y_i^*$。據此，CTC 的刪減規則會先將

重複字符移除，接著將預測字串中的 blank 移除，最後得到刪減後的字串，例如：「醫醫-療–語語音音」經過 CTC 刪減後會得到「醫療語音」。

在解釋 CTC 的基本原理之後，我們如下說明如何定義 CTC 的損失函數，以計算輸入的特徵和 $y_i^*$ 之間的損失，並藉此回調網路參數以提高辨識效果。爲了讓訓練和預測時，能讓預測結果更接近 $y_i^*$。損失函數定義如式 6 所示：

$$L_{CTC} = -log \sum_{s \in Align(x,y^*)} \prod_{t=1}^{T} p_t(s_t|x) \quad (6)$$

其中 $x, y^*$ 分別代表輸入語音特徵序列和相對應的 Ground Truth；$Align(x, y^*)$ 爲所有預測組合經過 CTC 刪減後與 $y_i^*$ 相同的組合，最後將所有屬於 $Align(x, y^*)$ 的組合機率相加。而由於訓練時希望損失函數 (Loss Function) 最小化，因此取負對數 (Negative logarithm) 來計算損失函數。

#### 3.3.5 共同解碼機制 Joint Decoding

爲了讓語音辨識模型在解碼時能同時兼具 CTC 能專注語音局部的特性，以及 Self-Attention 能保持較大跨度前後文關係的優勢，本論文所提的網路架構在解碼階段採取共同解碼機制，也就是透過 $\lambda$ 當作調和參數來調整 CTC loss 以及 self-attention 的 cross-entropy loss 的權重比例，如式 7 所示：

$$L_{total} = (1-\lambda)L_{self-attention} - \lambda L_{CTC} \quad (7)$$

## 4 實驗與結果

我們對照 Joint CTC-Attention 模型與本研究所提出 Joint CTC-Conformer 模型，有使用與沒有使用資料增量時，在有標點符號 psChiMeS-14 進行訓練，對於輸出必需有標點符號標註的辨識效能。同時，我們也比較也進一步探討增加標點符號的標註對 ASR 效能的影響，也就是同樣採用 Joint CTC-Conformer 網路架構，但訓練與測試的語料庫爲 ChiMeS-14 與 psChiMeS-14 的差異時，CER 與 KER 的影響。

### 4.1 Attention 與 Conformer 的比較

關於實驗評測指標，除了使用語音辨識常見的字符錯誤率 (Character Error Rate：CER) 之外，考量醫療情境中，醫學術語的正確辨識極爲重要，我們也加入了關鍵字錯誤率指標 (Keyword Error Rate：KER)，其中，關鍵字是從訓練文本中另外經由人工擷取出六大類共 707 個醫療相關的關鍵字，其數量如表 3 所列。按此定義的 KER 計算公式如式 8 所示：

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 |

$$KER = \frac{S_k + D_k + I_k}{N_k} \times 100\% \quad (8)$$

KER 的計算概念與 CER 類似，針對 Tabel 3 的關鍵字列表我們將 ground truth 的正確標註，以及預測結果中所出現的關鍵字，額外提取出來進行兩者的對齊比較。其中 $N_k$ 爲所有正確的關鍵字數量，而 $S_k$、$D_k$ 和 $I_k$ 預測的關鍵詞與正確關鍵字比對後發現有替換 (翻錯)、刪除 (漏掉) 與插入 (多加) 動作等三種錯誤的個別次數。

此外，針對 Out Of Keyword (OOK) 也就是未曾出現在訓練集，卻出現在測試集的關鍵詞，我們也另外定義 OOK-KER 的評測指標，如式 9 所示：

$$OOK - KER = \frac{S_{ook} + D_{ook} + I_{ook}}{N_{ook}} \times 100\% \quad (9)$$

其算法基本上與 KER 的算法相同，僅是 OOK-KER 是針對從未出現於訓練集的關鍵字進行評比：$N_{ook}$ 爲所有測試集中未出現在訓練中 ground truth 正確關鍵字的數量，而 $S_{ook}$、$D_{ook}$ 和 $I_{ook}$ 則爲比對測語音中，針對 OOK 的預測發生替換、刪除與插入這三種錯誤的次數。本研究將使用 CER、KER 和 OOK-KER 作爲評測標準來探討 ASR 績效，三種指標皆是數值越低，表示 ASR 效能越佳。

我們利用表 4 展示分別有使用波形增量的 Joint CTC-Attention 以及有使用速度增量之 Joint CTC-Conformer 訓練網路，在測試測試集中第 11 號錄音者所錄製的其中一份病歷表辨識結果。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。

如表 5 所示，兩模型在 baseline 不使用任何增量的條件下，Joint CTC-Conformer 的 CER 和 KER 分別優於 Joint CTC-Attention 大約 6.04% 和 9.51%，而 OOK-KER 也優於 Joint CTC-Attention 大約 14.3%。當各別使用數據增量後，Joint CTC-Attention 使用 wave 增量，也就是透過對原音檔上的音量 (volume)、音調 (pitch) 以及語速 (speed) 等進行隨機調整，並將因檔增加爲原來的 4 倍；而 Joint CTC-Conformer 則是使用語速

表 4: 比較 Joint CTC-attention 與 Joint CTC-Conformer 並模型的測試實例

| 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。 | - |
| Joint CTC-Attention with wave augmentation | {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<，>：他此次< 是 >因爲排現 {co}{lon}{can}{cer}<，>上胃開刀，乀左鎖骨放 {port}A<，>預計先第二次化療入院。左邊{an}一套口A<，>到十月三號<，>{su}{gar} 測 QIDAC，沒事。 | 16.81 |
| Joint CTC-Conformer with speed augmentation | {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<。>：他此次是因爲發現 {co}{lon}{can}{cer}，上位開大，乀左鎖骨放 {port}AA急性，DM次化療入院。左邊有一條 {port}A<，>打十月三< 號 >，{su}{gar} 測 QIDAC<，>沒< 事 >。 | 15.25 |

增量 (speed perturbation)，也就是透過調整原音檔的語速 (0.9 和 1.1 倍語速)，將音檔增加爲原來的 3 倍。從數據上顯示，Joint CTC-Attention 在加入 wave 增量之後，相較於不使用的 baseline，其 CER 和 KER 分別下降 4.74% 和 7 %，而 OOK-KER 則持平。而 Joint CTC-Conformer 在加入語速增量後，相較於沒有使用的 baseline，其 CER、KER 和 OOK-KER 也分別下降了 3.9 %、6.89% 和 3.34 %。

由上述數據可知，在使用數據增量後，對於語音辨識效能提升是很有幫助。若進一步比較兩模型的效能，可發現 Joint CTC-Conformer 在 baseline 的條件下，三項指標都優於使用 wave 增量的 Joint CTC-Attention；在使用增量的情況下，Joint CTC-Conformer 的 OOK-KER 更是優於 Joint CTC-Attention 大約 17.69 %，再次驗證了 Conformer 加入 Convolution 的機制能有效補捉局部特徵，提升醫療關鍵字的辨識效能。

表 5: 不同架構在 psChiMeS-14 上的效能比較

| ASR | Joint CTC-Attention | | Joint CTC-Conformer | |
|---|---|---|---|---|
| Aug. | baseline | wave | baseline | speed |
| CER(%) | 20.44 | 15.70 | 14.40 | 10.50 |
| KER(%) | 29.50 | 22.50 | 19.99 | 13.10 |
| OOK-KER(%) | 76.85 | 76.85 | 62.50 | 59.16 |

我們特別說明不曾出現在訓練集中 OOK 的辨識。相較於一般 ASR 不可能測試出不曾出現在訓練集的字符 (OOV)，對於醫療關鍵字，其本質上比較像是詞的概念，也就是由幾個獨立的字或是英文單音節所串接而成，只要一個由多個字是多個英文單音節所組成的關鍵詞 (keyword)，其各別組成的字或單音節曾出現在訓練集中，好的 ASR 就有機會將這從來沒有聽過的關鍵詞重組出來。在本研究所提出的 Joint CTC-conformer 架構中，不爲全錯的 OOK-KER 即代表：即使不曾在訓練集中出現的關鍵字，也有可能被正確辨識出來。也就是說，儘管沒有額外的 PM (pronunciation model) 以及 LM (language model) 的塑模，本研究所提出的 ASR 架構，仍可以從訓練集中出現過字符的前後文開係拼湊出先前不曾看過的關鍵字。表 6 與表 7 分別是一些被正確辨識出中文與英文 OOK 的例子。

### 4.2 增加標點符號的標註對 ASR 的影響

此外，我們也好奇，加入標點符號標註的文本對於端對端的語音辨識模型在辨識績效上的影響。我們固定使用 Joint CTC-Conformer 模型，比較當使用無標點符號的 ChiMeS 與有標點符號標註的 psChiMeS 這兩種語料庫，進行訓練與後續測試效果的比較。如表 8 所示，當 Joint CTC-Conformer 使用加上標點符號的 psChiMeS 訓練時，會比當使用沒有標點符號標註的 ChiMeS 訓練，CER 上升了 2.1 % 和 SER 上升大約 11%。我們解釋的原因是：標點符號占了 psChiMeS-14 整份文本中字數的大約 10%，而由整體標點符號正確辨識率的 F-Score 來看，約爲 83%，因此可以推知模型在 CER 上升 2% 的原因，主要應該是由標點符號所影響；而 SER (sentence error rate) 部分，由於 SER 計算方式較爲嚴格，也就是只要一句裡錯一個字即視爲全錯，再加上加入標點符號訓練所造成的標點符號錯誤，因此導致 SER 的上升。

表 6: 中文可辨識 OOK 範例

| OOK | Conformer Results |
|---|---|
| 膽管癌 | 七 B 五二二，男性其實歲診斷是膽管癌，UTI，然後他沒有高跌，是高壓的病人，沒有過敏史。 |
| 腸造口手術 | 他預計五月二二號進去開刀房做胰臟結腸切除跟吻合術，然後再加腸造口手術 |
| 心包膜積水 | 阿沒有心包膜積水齁，然後 IO 有都還好話，然後沒 S 二十九號辦出院。 |
| 輸尿管碎石 | 然後他的過去病史有：DM，尿路結石有開剖腹採，有輸尿管碎石，還有右側輸尿管狹窄開過刀。 |
| 肝囊腫 | buscopan，buscopan 打過，有 follow A 加 P，A 腹部的 CTL 已沒有顯影的，就是肝囊腫，然後腸炎 X ray 沒有事，EKG sinus tachycardia。 |

表 7: 英文可辨識 OOK 範例

| OOK | Conformer Results |
|---|---|
| kascoal | 然後他有那個自備藥有一些藥 kascoal 齁，還有一些 primperan nexium 這些都把他停掉了。 |
| bladder CA | 男性八十二水診斷是 bladder CA，沒有高跌，沒有高壓，沒有過敏史，是 on full diet。 |
| CAD | 男性五素水診斷是 CAD ESRD，沒有高跌，沒有高壓。 |
| on levophed pump | 八月十七號因爲血壓低，給他 on levophed pump，之後血壓 OK 就 on 服。 |
| 右 lung tumor | 男性六十五歲，他的診斷是右 lung tumor，沒有高跌，沒有高壓，沒有過敏史，目前是 on soft diet。 |

我們另外分析，標註文本加入標點符號後，對於原本中文字與英文單音節 (mon-syllable) 辨識的影響。爲此，我們將 psChiMeS-14 測試結果中的標點符號去除，重新計算前述三項效能指標，並與以文本中本身就沒有標點符號標註的 sChiMeS-14 訓練後模型的的測試結果進行比較。如同表 8中的第三列示，CER 並沒有太大的差異；SER 的部分些微下降了 0.45%，而 KER 下降了 1.14%。從數據上可以觀察出，模型使用加入標點符號的文本進行訓練，不僅不會造成中文字與英文單音節辨識效果下降，反而因爲加入標點符號後前後語意更爲完整，所以讓醫療關鍵字辨識率有些微的提升。

表 8: Joint CTC-Conformer 不同語料庫效能比較

| Corpus | Joint CTC-Comformer | | |
|---|---|---|---|
| | CER | SER | KER |
| sChiMeS-14 | 8.42 | 76.21 | 13.59 |
| psChiMeS-14 | 10.5 | 87.60 | 13.10 |
| psChiMeS-14 (remove punctuation) | 8.40 | 75.76 | 12.45 |

## 5 Conclusion

爲了改良中文醫療語音辨識的結果也提供標點符號的需求，本研究透過重新標註原語料庫文本 ChiMeS-14 而得到有標點符號的訓練集 psChiMeS。然後，利用以 self-attention 機制爲基礎的 conformer 模型與 CTC 合併搭建的 Joint CTC-Attention ASR 的模型，在 psChiMeS-14 語料庫進行訓練與測試，我們獲得到目前爲止最好的語音辨識績效。一般說來，利用加上標點符號文本，在 Joint CTC-Conformer 的架構上進行端對端的訓練，基本上不會對 CER, SER 或是 KER 有明顯的影響。ASR 僅將這些額外標註的標點符號視同與中文字或是英文單音一樣的字符。

最後，由於目前病歷表的讀出較爲口語，未來將針對辨識結果中贅字虛字的校正與制式文書的轉譯謄寫，如：嗯、齁、痾等贅字可進行刪除。而醫療文本終有許多筆記型的陳述方式省略完整內容，不易一般人明瞭，因此未來也會使用後處理的機制進行轉譯，如高跌表示將轉譯爲高跌倒危險群等，方便大眾閱讀。此外，由 (Salloum et al., 2017) 等學者所提出的研究中，將語音辨識的結過再進一步透過

NLP 模型轉換成診斷報告。我們受到了啓發，未來將嘗試利用醫療文本中關鍵字建構智慧型的醫學診斷以及出院報告等，讓病歷報告可以自動化生成。

## 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Bhrigu Garg et al. 2018. Analysis of punctuation prediction models for automated transcript generation in mooc videos. In *2018 IEEE 6th International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pages 19–26. IEEE.

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*.

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.

Daniel Li, I Te, Naveen Arivazhagan, Colin Cherry, and Dirk Padfield. 2021. Sentence boundary augmentation for neural machine translation robustness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7553–7557. IEEE.

Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. 2020. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Wael Salloum, Gregory Finley, Erik Edwards, Mark Miller, and David Suendermann-Oeft. 2017. Automated preamble detection in dictated medical reports. In *BioNLP 2017*, pages 287–295.

Monica Sunkara, Srikanth Ronanki, Kalpit Dixit, Sravan Bodapati, and Katrin Kirchhoff. 2020. Robust prediction of punctuation and truecasing for medical asr. *arXiv preprint arXiv:2007.02025*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ruiqing Zhang and Chuanqiang Zhang. 2020. Dynamic sentence boundary detection for simultaneous translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 1–9.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.