The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

針對特定領域之中英語碼語音辨識系統
Mandarin-English Code-Switching Speech Recognition System for Specific Domain

邱川溥 **Chung-Pu Chiou**, 林厚安 **Hou-An Lin**, 陳嘉平 **Chia-Ping Chen**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
m103040061@nsysu.edu.tw, m093040066@nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

摘要

本文將介紹利用自動語音辨識 (Automatic Speech Recognition, ASR) 技術處理帶有特定領域的語音內容。我們將以 Conformer 端到端模型做爲系統架構，並利用純中文資料進行初步訓練，再以遷移式學習（Transfer learning）技術對系統以中英語碼轉換 (Mandarin-English Code Switching) 資料進行一次微調，最後利用帶有特定領域的中英語碼轉換資料對模型進行最終微調，使其在特定領域的語音辨識上達到一定的效果。我們以不同微調方式進行實驗，最終錯誤率從 82.0% 降到 34.8%。

Abstract

This paper will introduce the use of Automatic Speech Recognition (ASR) technology to process speech content with specific domain. We will use the Conformer end-to-end model as the system architecture, and use pure Chinese data for initial training. Next, use the transfer learning technology to fine-tune the system with Mandarin-English code-switching data. Finally, use the Mandarin-English code-switching data with a specific domain makes the final fine-tuning of the model so that it can achieve a certain effect on speech recognition in a specific domain. Experiments with different fine-tuning methods reduce the final error rate from 82.0% to 34.8%.

關鍵字：語音辨識、語碼轉換、語音識別、語言模型、遷移式學習

Keywords: Speech Recognition、Code switching、Language model、Transfer learning

1 緒論

近幾年在全球疫情的影響下，許多會議以及課程等事項都逐漸以遠距的方式來執行。而在課程方面，有些老師或機構都紛紛成立自己的 Youtube 頻道，並將原先實體授課的內容紀錄下來再上傳到 Youtube 中，此做法能夠避免在疫情嚴重的情況下造成群聚的風險，學生也能夠更方便的學習知識。

雖然將課程放上 Youtube 能夠方便學生學習，但影片的錄製方式也會直接影響到播放的聲音品質。當影片錄製是直接在課堂上進行收音時，整段影片音訊會充滿各種環境噪音，學生也因此不容易聽清楚老師所講授的內容。對於以上問題可以利用人工添加字幕的方式來解決，不過此方法非常耗時耗力。我們希望利用語音辨識（Automatic Speech Recognition, ASR）系統來提昇上字幕的效率，但課程中難免會受到課程領域而有不同中文和英文專有名詞的影響，因此本論文提出針對特定領域的中英語碼轉換語音辨識（Mandrin-English Code-Switching Speech Recognition System For Specific Domain）系統。

本文基於 (Lin and Chen, 2021) 的實驗方法，但在基礎架構上採用 Conformer (Gulati et al., 2020b) 端到端（End-to-End, E2E）架構來進行實驗。

在實驗中，首先以中文資料集先對 ASR 模型進行訓練，當作具有中文能力的基礎語音辨識系統。另外，我們從 Youtube 擷取一些教育相關內容的語音資料，並分成 Education 資料集與 Course 資料集。接著使用由 (Wang and Zheng, 2015) 所提出的遷移式學習（Transfer Learning）對中文 ASR 模型以帶有中英語碼轉換的 Education 資料集進行一次微調，使其能學習處理中英夾雜的語音，另外也使用帶有領域資訊的 Course 資料集進行一次微調，來與前者進行比較。最後利用以 Education 資料集微調後再以 Course 資料集微調的二次微調與上述兩者比較，藉此使我們的 ASR 模型達到能夠處理特定領域中專有名詞的能力。

在接下來的章節中，第二章會詳細介紹我們的實驗架方法，第三章爲使用的資料集與實驗設置，第四章將呈現我們的實驗結果和結果分析，第五章爲我們本次實驗的結論與見解。

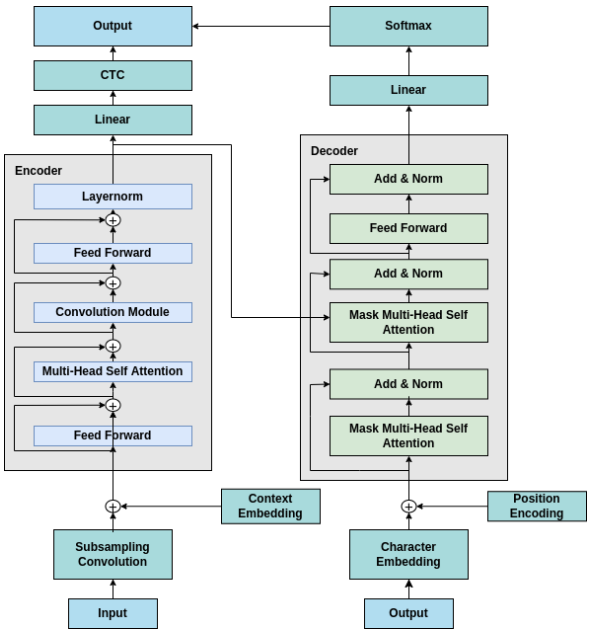

圖 1. Conformer 架構，連續時序分類器 (Connectionist Temporal Classification, CTC) 和 Transformer 解碼器將接收到 Conformer 編碼器的輸出。

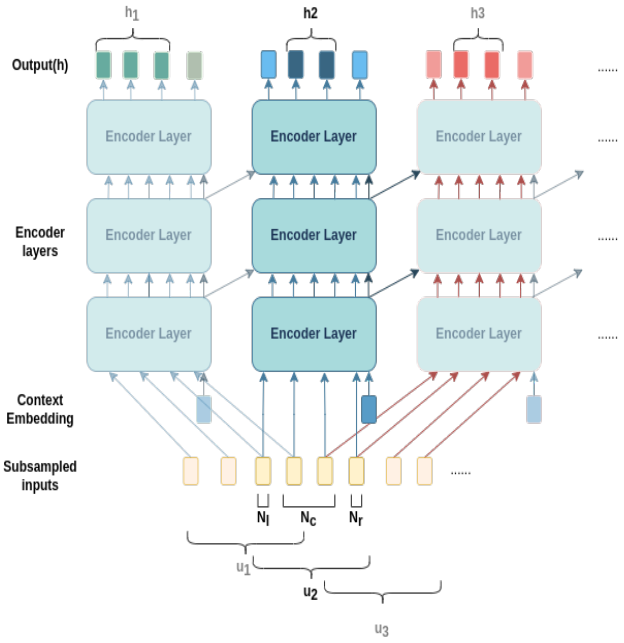

圖 2. Contextual Block Processing 示意圖。其中 u_i 爲 block，並包含著數個 frames，這些 frames 將被標記成過去、目前、未來 $\{N_l, N_c, N_r\}$ 三個部份。在訓練過程中，較後面的 block 將繼承前面 block 的訊息。

2 實驗方法

我們使用 Conformer 端到端架構聯合訓練連續時序分類器 (Connectionist Temporal Classification, CTC) (Graves et al., 2006) 當作 ASR 模型，而在編碼器與解碼器上結合 (Lin and Chen, 2021) 所利用到的 Contextual Block Processing (Tsunoo et al., 2019) 和 Blockwise Synchronous Beam Search (Tsunoo et al., 2020) 使 ASR 模型擁有 Streaming 的效果，其中 Contextual Block Processing 如圖 2 所示。實際訓練過程中以中文資料集來訓練一個中文語音辨識系統，再利用遷移式學習以 Education 資料集進行一次微調當作基準，並以 Course 資料集做二次微調後來與基準做比較，另外也單獨以 Course 資料集行一次微調加入比較。以下將介紹我們使用的模型架構以及訓練方法。

2.1 端到端模型

本篇論文我們所使用的 Conformer 端到端 ASR 模型架構如圖 1 所示，其利用 Conformer 區塊取代了 Transformer (Vaswani et al., 2017) 編碼器的部分。Conformer 架構如圖 1 所示。

2.2 編碼器

輸入的 80 維梅爾頻譜圖（mel-spectrogram）資料首先會經過由兩層捲積神經網路（Convolutional neaural network）和 ReLU 激活函數所組成的降採樣模塊（Subsampling Convolution），kernel 大小爲 3 及 stride 爲 2，channel 數爲 256，其中 channel 爲自注意力機制（Self-attention）特徵的維度。經過降採樣後的音訊序列資料會輸入到數個 Conformer 區塊 (Gulati et al., 2020a) 中，其結合自注意力機制和卷積，前者學習交流 global 等級的資訊，而後者捕捉 local 等級的資訊。Conformer 區塊之架構如圖 1 左側所示，輸入 a_i 進到第 i 個 Conformer 區塊產生輸出 x_i 之數學表示式爲:

$$\tilde{a} = a_i + \frac{1}{2}\text{FFN}(a_i)$$
$$a'_i = \tilde{a}_i + \text{MHSA}(\tilde{a}_i)$$
$$a''_i = a'_i + \text{Conv}(a'_i)$$
$$x_i = \text{Layernorm}(a''_i + \frac{1}{2}\text{FFN}(a''_i))$$

FFN 指的是 Feed Forward 模組，MHSA 指的是 Multi-Head Self-Attention 模組，Conv 指的是 Convolution 模組。最後編碼器輸出爲 X_e 。

2.3 解碼器

Transformer 解碼器將接收到 Conformer 編碼器的輸出 X_e 和序列的 token IDs Y[1:u] = Y[1],...,Y[u]，此 token IDs Y[1:u] 及編碼器輸出 X_e 將被用來計算序列的後驗機率（poste-

資料集	音檔數	總時長（小時）
NER-Trs-Vol1	21,089	126.65
AISHELL-1	20,000	24.82
AISHELL-2	20,000	19.87
科技大擂台	24,102	50.50
total	85191	221.84

表 1. 中文資料集的音檔數以及總音檔時長

資料集	音檔數	總時長（小時）
訓練集	2301	9.36
驗證集	254	1.05
測試集	1047	2.55
total	3602	12.96

表 2. Education 資料集的音檔數以及總音檔時長

rior probabilities):

$$[p_{\text{s2s}}(Y[2]|Y[1], X_e), ..., p_{\text{s2s}}(Y[u+1]|Y[1:u], X_e)] = \text{softmax}(Z_d W_{\text{att}} + b_{\text{att}})$$

$$p_{\text{s2s}}(Y|X_e) = \prod_u p_{\text{s2s}}(Y[u+1]|Y[1:u], X_e)$$

Z_d 爲編碼器的輸出，$W_{\text{att}} \in \mathbb{R}^{d_{\text{att}} \times d_{\text{char}}}, b_{\text{att}} \in \mathbb{R}^{d_{\text{char}}}$ 爲可學習之參數，d_{char} 爲字元數量。Transformer 解碼器如圖 1 右側所示。

2.4 訓練方法

我們採用聯合訓練 CTC 的方式。連續時序分類器 (Connectionist Temporal Classification, CTC) 會將每個語音特徵與輸出字元做對齊，聯合訓練 CTC 能使學習速度提昇，並且使模型更快速的收斂 (Kim et al., 2017)。在訓練階段損失函數結合解碼器和 CTC 的負對數機率 (Kim et al., 2017; Nakatani, 2019)，如下所示：

$$L_{\text{mtl}} = -\alpha \log p_{\text{s2s}}(Y|X_e) - (1-\alpha) \log p_{\text{ctc}}(Y|X_e)$$

p_{ctc} 爲 CTC 的後驗機率，α 爲超參數，能夠用來調整 CTC 與模型之間的比例。

2.5 遷移式學習

遷移式學習 (Wang and Zheng, 2015) 是將之前訓練好的模型當做基礎，稱爲其爲預訓練模型，接下來的訓練將繼承預訓練模型已訓練好的參數再進一步使用新資料去進行微調（Fine-tune），因此遷移式學習能在資料量小的情況下，依然使模型具有其他領域能力的效用。

資料集	音檔數	總時長（小時）
訓練集	17145	15.18
驗證集	2143	1.86
測試集	2144	2.80
total	21432	19.84

表 3. Course 資料集的音檔數以及總音檔時長

3 實驗設置與資料集

在本篇論文使用的資料集分爲三種，其中包含中文資料集和兩個教育資料集，而兩個教育資料集皆包含中英語碼轉換的內容，另外教育資料集大部份是直接在教室進行錄製，因此也包含著各種環境噪音。另外，我們使用 ESPnet2 (End-to-End Speech Processing tookit) (Watanabe et al., 2018) 這套端到端語音處理工具包來進行 ASR 相關實驗。

3.1 中文資料集

中文資料集總共由四個部份所組成，此資料集的資訊以表 1 表示。

(1) NER-Trs-Vol: 由國立教育廣播電台所提供，資料內容爲談話性節目及新聞報導的朗讀式語音，總時長爲 126.8 小時，共 21,089 筆資料。

(2) AISHELL-1：由 AISHELL 公司所提供 (Bu et al., 2017)，內容紀錄著 400 位來自中國不同地區的人的語音，而其語音內容包含 11 個不同領域，像是智慧家庭、無人駕駛等。我們將其文本從簡體字轉爲繁體字並隨機抽出 20000 筆來當作訓練資料。

(3) AISHELL-2：由 AISHELL 公司所提供 (Du et al., 2018)，內容紀錄著 1991 位來自中國不同地區的人的語音，而其語音內容包含 12 個不同領域，像是智慧家庭、無人駕駛等。我們將其文本從簡體字轉爲繁體字並隨機抽出 20000 筆來當作訓練資料。

(4) 科技大擂台（Formosa Grand Challenge）：由國研院科技政策研究與資訊中心提供，其語音內容爲華語能力測驗，並且分成文章、題目和選項。此資料集總時長爲約 400 小時，我們利用其中的部份問題及選項加入到訓練集中，總時長爲 50.5 小時，共 24,102 筆資料。

3.2 教育資料集

首先是 Education 資料集，此資料集是在 Youtube 上的一些教育相關內容且其語音包含中英語碼轉換。部份資料集語音是在安靜的環境下錄製，其他則是直接在教室裡面錄製，因此較爲吵雜，Education 資料集詳細資料以表 2 表示。

另外，Course 資料集爲同一位教授現場講授資料結構課程的語音，因此資料中有各種環境雜音，內容也包含中英語碼轉換。我們將原先在 Youtube 上的課程的影片音訊與經校正過的字幕檔做讀取，再將音訊從原先的 48,000 採樣率轉爲 16,000 採樣率以便訓練，最後利用字幕檔的時間標籤將完整音訊依照對應文本切割爲多個片段音訊，其中取得 2144 筆當作最終測試集，Course 資料集詳細資料以表 3 表示。

3.3 資料增強

我們在資料前處理上使用了兩種資料增強的方法，分別爲速度擾動（Speed Perturbation）(Ko et al., 2015) 以及 SpecAugment (Park et al., 2019) 來解決我們資料量過少的問題。速度擾動能將原始音訊資料經過三個不同倍率來產生不同速度的新資料，而我們採用的倍率爲 0.9、1.0、1.1，此方法能有效的增加資料量。SpecAugment 則是對梅爾頻譜圖分別進行時間扭曲（Time Warping），也就是在時間軸上對頻譜圖的特定區塊進行平移，以及在時間（Time）與頻率（Frequency）軸上對頻譜圖進行遮罩的動作。

3.4 端到端模型設置

我們所採用的端到端架構爲 Conformer ，其中包含 12 層的 Conformer 編碼器以及 6 層的 Transformer 解碼器，Conformer 中深度卷積的 kernel 大小爲 15，另外使用了 (Tsunoo et al., 2019, 2020) 的方式使所有輸入的 frames 重疊一半，且在 block 中的過去、目前、未來三個部份以 $\{N_l, N_c, N_r\}$ 表示，我們的設置爲 $\{8, 16, 16\}$。在 (Tsunoo et al., 2019) 提到的 Contextual Embedding Vector 中，我們採用了將 block 中的 frames 取平均的方式來當作初始值，並利用 Position Encoding 區分不同 block 的序列。此外，我們在多任務學習方法（multitask learning）的超參數 α 設爲 0.3，解碼階段的超參數 λ 與 γ 分別設爲 0.5 和 0.3。採用 Adam Optimizer (Kingma and Ba, 2017) 當作優化器。

模型	CER(%)
Conformer-FT-Education	82.0
Conformer-FT-Course	35.4
Conformer-FT-Education-Course	34.8

表 4. 利用 Course 測試集比較在 Conformer 模型上利用不同資料集做微調的表現，其中 FT 爲 Fine-Tune 簡稱。

REF：接 下 來 是 s p a c e <space> c o m p l e x i t y
微調 1：Education 資料集
HYP：接 下 來 的 s p e a t e <space> c o * * n e s i * s
Eval： S I S DDS S DS
微調 2：Course 資料集
HYP：接 下 來 *** s p l a c e <space> c o m p l a c i t y
Eval： D I S S
微調 3：Education 資料集、Course 資料集
HYP：接 下 來 *** s p a c e <space> c o m p l a c i t y
Eval： D S S

圖 3. REF、HYP、Eval 分別爲 Reference、Hypothesis、Evaluation。另外，D 代表刪字錯誤，S 代表換字錯誤，I 代表插字錯誤，<space> 爲空格。

4 實驗結果

我們以 Course 測試集來對兩個模型進行評分，並使用字元錯誤率（CER）當做評斷標準，實驗結果以表 4 表示。首先可以看到在單純使用有中英語碼轉換的 Education 資料集進行微調後，模型在字元錯誤率表現爲 82.0%，使用 Course 資料集微調後錯誤率降到 35.4%，而經過 Education 資料集與 Course 資料集的二次微調後，字元錯誤率來到了 34.8%。另外，由圖 3 的微調 2 與微調 3 可以觀察到有加入 Education 資料集的微調方式仍然在英文能力上有幫助。

4.1 領域影響

圖 3 爲我們分別從三種不同的微調方式取出對同一筆資料的預測結果，由圖中可以觀察到只有經 Education 資料集微調過後的模型在 Hypothesis 中，英文專有名詞部份的辨識率非常差，此原因爲 Education 資料集並無包含資料結構領域的內容，因此常出現在資料結構領域的英文專有名詞要被辨識出來是比較困難的。而有經過 Course 資料集微調後的模型在 Hypothesis 中，英文專有名詞有更大的機率辨識出正確結果。

4.2 資料品質影響

由於教育資料集大部份音訊是以現場授課的形式呈現，因此裡面包含著學生的說話聲、環境音等，再加上收音裝置與錄音軟體對音訊品質的影響，以上各種因素也直接影響了模型效能，因此在錯誤率方面仍然有進步空間。

5 結論

由本次實驗能發現在特定領域中時常出現專有名詞的狀況，例如在課程上可能會有至少一半的句字出現專有名詞，若 ASR 模型在訓練過程中未學習過此領域的資訊，此因素將會使效能銳減。另外，在實際情況下專有名詞也常會以英文的形式出現，因此我們使用遷移式學習的方式先使模型在資料缺乏的情況下，能夠擁有語碼轉換與特定領域的能力，而在我們的實驗結果上有顯著的改善。

另外，由於教育資料集大部份帶有環境雜音的關係，因此錯誤率仍然還有很大的進步空間，其中我們也嘗試過使用語言模型輔助，但也許是語言模型訓練資料的問題造成實驗上錯誤率不減反增，因此我們仍然會對語言模型部份持續實驗，並嘗試加入語言分類器來增加中英語碼轉換的效能。

References

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. 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, and Ruoming Pang. 2020a. Conformer: Convolution-augmented transformer for speech recognition.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020b. 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.

DP Kingma and J Ba. 2017. Adam: A method for stochastic. *optimization*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*.

Hou-An Lin and Chia-Ping Chen. 2021. Exploiting low-resource code-switching data to mandarin-english speech recognition systems. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 81–86.

Tomohiro Nakatani. 2019. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proc. Interspeech*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Emiru Tsunoo, Yosuke Kashiwagi, Toshiyuki Kumakura, and Shinji Watanabe. 2019. Transformer asr with contextual block processing. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 427–433. IEEE.

Emiru Tsunoo, Yosuke Kashiwagi, and Shinji Watanabe. 2020. Streaming transformer asr with blockwise synchronous beam search.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.