The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

A preliminary study on Hakka speech recognition by using the Branchformer
使用 Branchformer 之端至端客語語音辨識初探

Jia-Jyu Su, Dong-Min Li, Chen-Yu Chiang
National Taipei University, New Taipei City, Taiwan
{s711181105, s410886005}@gm.ntpu.edu.tw, cychiang@mail.ntpu.edu.tw

摘要

本文爲 2023 客語辨識挑戰對於端到端客語語音辨識初探，使用 Branchformer 作爲模型並在 ESPnet 下執行，說明爲何選用 Branchformer 作爲辨識模型與其特點，並實驗字元和拼音兩個不同目標對於辨識效果的好壞，並驗證中文語料是否對客語的辨識有幫助。

Abstract

This paper is a preliminary study on Hakka speech recognition using the end-to-end Branchformer framework provided by the ESPnet. Two types of recognition targets were tested: character and Hakka pinyin. The experimental result showed that the Branchformer for the Hakka speech recognition pre-trained with the large Mandarin speech corpus Aishell-1 can improve the recognition performance by using the Branchformer trained by the Hakka speech corpus only.

關鍵字：客語辨識、語音辨識、漢語變體
Keywords: Hakka speech recognition, ASR, Varieties of Chinese

1 Introduction

近幾年，中華民國政府爲了對於母語進行保存，特別制定了國家語言發展方向，因此對於民眾常用的母語，都有讓國中小學安排母語課程，這些母語課程包括閩南語、客語、原住民語、以及其他東南亞新住民語言。在此時空環境下，政府爲了要能倡導母語的使用、教學、保存以及再發展，舉辦了這次的客語語音辨識大賽。

嚴格來講，客語是屬於漢語系的方言，漢語最主要可以分爲 7 大方言系，客語是屬於全球漢語使用人口數的第 7 名 (Wikipedia, 2023)。

如同其他漢語方言，客語沒有特別的文字紀錄方式，原因是長期的歷史以及政治因素，並沒有任何系統性並且可大量被使用的文字紀錄方法，加上目前年輕人口逐漸沒有使用方言的場域，只剩下教育部使用政策方法來推行客語等等的母語教育。

客語又可以分爲許多次方言，比如海陸腔、四縣腔、大埔腔、饒平腔、紹安腔和南四縣腔，這些腔調的差異有出現在基礎音節的音素組合上，也出現在聲調於音高以及音節長度上的韻律差異。

以客語的使用價值來看，最重要是在於文化資產的保留，客語就是最能夠表示客家文化的知識體系以及語言系統空間。若要活化從文化資產，便需要科技元素的融入，語音辨識的競賽便是一種很直觀套入的形式。

以技術價值來討論，客語語音科技的發展於產業的產值創造發揮有限，無法和主流社會經濟生活需求的國語以及英語、日語一樣，但建立客語語言學習所需要的輔助工具，比如利用語音辨識技術來做爲客語發音檢正確與否的鑑測技術，是可以發展的方向。

本論文介紹了使用 Branchformer (Peng et al.) 之端至端客語語音方法。使用 Branchformer 做爲系統建立工具的原因是 Branchformer 做爲編碼器將抽取全域和局部資訊的部分分爲兩個分支，以更好的提取資訊讓準確度上升，除此之外 Branchformer 也有較好的訓練穩定度。

本論的組織如下：首先介紹 banchformer 的模型架構及其特點，接著實驗環境設定與實驗方法，最後附上實驗討論與結語。

2 Branchformer 介紹

Branchformer 是對編碼器去設計，將提取全域與局部資訊的模塊分開爲兩個平行的分支 (branch)，目的是讓模型可以提取更多範圍 (various ranged) 之間的相關性，分支的設計讓 Branchformer 有以下優點：模型設計有彈性、可根據目標客製分支、模型較好解釋。因爲雙分支的緣故，Branchformer 可以根據需

求更換個別分支所使用的架構，例如爲降低複雜度在注意力模塊使用 fastformer，雙分支的設計讓兩個分支在合併的時候可以被可學習的權重控制，讓模型學習到在不同的狀況下，哪個分支的資訊較爲重要，使模型更加彈性，另外雙分支可以在推理階段時將注意力端關閉來加速處理。

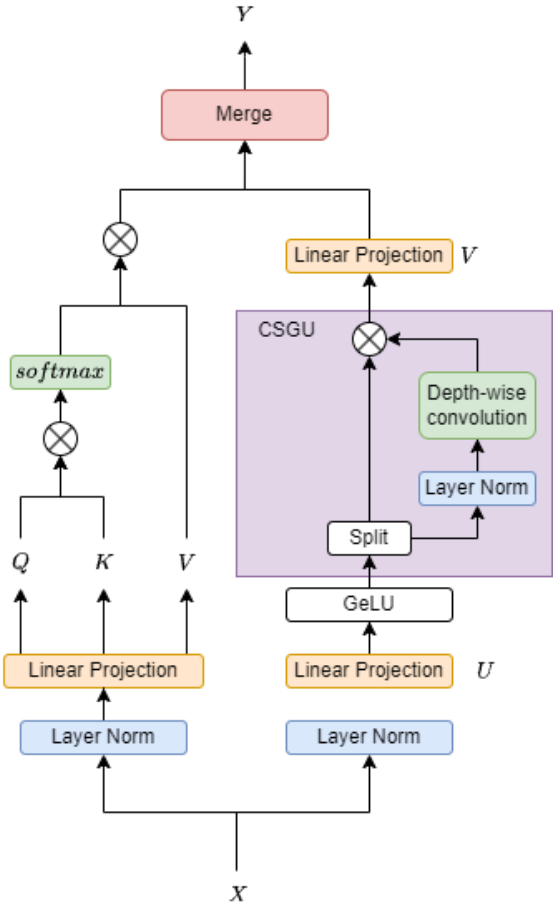

Figure 1: The architecture of Branchformer model

2.1 Attention Branch

注意力模塊透過注意力機制對整個序列提取資訊，輸入 $X \in \mathbb{R}^{T\times d}$，$T$ 是序列長度，d 是特徵維度，MHSA (Multi-headed Self Attention) 首先會將輸入轉換成 $Q, K, V \in \mathbb{R}^{T\times d}$ (query、key、value) 三個矩陣，且投影內的參數爲可學習的，Q 與 K 會做內積並經過 $softmax$ 得到一組權重代表每個位置資訊的重要性，最後跟 V 相乘得到輸出。

$$\mathrm{Attention(Q,K,V)} = \mathrm{softmax}(\frac{\mathrm{QK^T}}{\sqrt{d}})\mathrm{V} \quad (1)$$

在 attention 的數學式中，$\mathrm{softmax}(\frac{\mathrm{QK^T}}{\sqrt{d}})$ 的矩陣相乘，由於 d 是常數，而 T 則爲輸入長度，而複雜度與 T 承平方關係，當輸入越長，輸出的速度越慢。

MHSA 實際上輸入會經過 h 次的投影，這些投影是平行化的，因此在最後需要將每個 attention head 的輸出組合起來再投影成原本的大小，才是 MHSA 最後的輸出。

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{concat(head_1 ... head_h)W^O 0} \quad (2)$$

$$\mathrm{head}_i = \mathrm{Attention(QW}_i^Q, \mathrm{KW}_i^K, \mathrm{VW}_i^V) \quad (3)$$

此處 $\mathrm{W}_i^Q, \mathrm{W}_i^K, \mathrm{W}_i^V \in \mathbb{R}^{d\times d/h}$，爲投影矩陣，將 $\mathrm{Q,K,V}$ 投影到較低維度；$\mathrm{W}^O \in \mathbb{R}^{d\times d}$，將各個 head 組合的結果投影轉換爲最後的輸出。

2.2 Convolution Branch

捲積模塊透過帶有捲積閥的多層感知機 (MLP) 提取局部的上下文資訊，其藉由深度方向的捲積和一個線性閥來實現，cgMLP 比起 Conformer 的捲積模塊效能要好，其主要組成是由 Gaussian error Linear Unit (GeLU)、convolution spatial gating unit (CSGU) 與投影轉換層所組成。

cgMLP 首先將輸入 $X \in \mathbb{R}^{T\times d}$ 通過 layer-norm，之後經過許多層到最後的輸出：

$$\mathrm{Z} = GeLU(XU) \in \mathbb{R}^{T\times d_{hidden}} \quad (4)$$

$$\tilde{\mathrm{Z}} = CSGU(Z) \in \mathbb{R}^{T\times d_{hidden}/2} \quad (5)$$

$$\mathrm{Y} = \tilde{Z}V \in \mathbb{R}^{T\times d} \quad (6)$$

其中 $\mathrm{U} \in \mathbb{R}^{d\times d_{hidden}}$,$\mathrm{V} \in \mathbb{R}^{d_{hidden}/2\times d}$，爲兩個通道投影，隱藏層的維度通常會大於輸入的維度，這樣的設計相似於點對點的 feed-forward 層。

cgMLP 的另外一個要件爲 CSGU，它包含了一個線性閥並採用了深度方向的捲積來捕捉局部關係，他的輸入 $\mathrm{Z} \in \mathbb{R}^{T\times d_{hidden}}$ 會在特徵維度被均等分成 $\mathrm{Z}_1, \mathrm{Z}_2 \in \mathbb{R}^{T\times d_{hidden}/2}$，之後只有 Z_2 會沿著時間維度做深度方向的捲積：

$$\mathrm{Z}_2' = \mathrm{DWConv(LayerNorm(Z_2))} \quad (7)$$

最後的輸出是將 $\mathrm{Z}_1, \mathrm{Z}_2'$ 做元素之間的相乘，得到 $\tilde{\mathrm{Z}} = \mathrm{Z}_1 \bigotimes \mathrm{Z}_2'$，將自己相關的資訊跟自己做內積，是一種 self-gating，透過自身資訊去決定該位置的資訊是否向前傳遞，而這實際上也是一種線性閥，因爲在相乘之前不會經過非線性激活層。

2.2.1 複雜度分析

在 cgMLP 模塊主要有兩個通道投影和 CSGU，其複雜度分別爲 $O(Tdd_{hidden})$、$O(Tdd_{hidden}/2)$、$O(TKd_{hidden}/2)$，其中 K 是 kernel size 爲一常數，全部看下來 cgMLP 的複雜度只跟序列長度 T 成線性關係。

2.3 Branch Merge

本節說明 Branchformer 如何將兩個分支的資訊結合，當作編碼器的輸出向前傳遞給解碼器，一種是將兩個序列接在一起後再降維至原本的長度，另一種則是讓模型學習如何合併才是最好的，結果顯示直接連接的方式比起銓重平均更好，推測是因爲這種方法將所有資訊都傳遞出去，較多的資訊量讓模型效果較好，而另一種方式雖然會限制資訊的傳遞量，但對於研究觀察來說可以更好的解釋模型的行爲與其學習的內容。

2.3.1 Concatenation

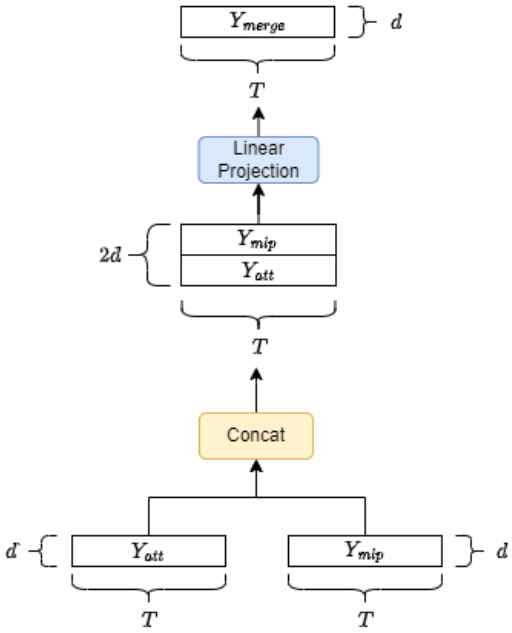

Figure 2: Concatenation merging

直接連接的方式易於實作，將兩個分支的輸出 $\mathrm{Y}_{\mathrm{att}}, \mathrm{Y}_{\mathrm{mlp}} \in \mathbb{R}^{T \times d}$ 直接沿著特徵的維度相接成 $\mathrm{Y}_{\mathrm{concat}} \in \mathbb{R}^{T \times 2d}$，接著乘上一個轉換矩陣投影到原本的維度，圖二說明此方式的實際流程，用數學式表示如下：

$$\mathrm{Y}_{\mathrm{merge}} = \mathrm{concat}(\mathrm{Y}_{\mathrm{att}}, \mathrm{Y}_{\mathrm{mlp}})\mathrm{W}_{\mathrm{merge}} \in \mathbb{R}^{T \times d} \tag{8}$$

其中轉換矩陣 $\mathrm{W}_{\mathrm{merge}} \in \mathbb{R}^{2d \times d}$ 是可學習的參數。

2.3.2 Weighted Average

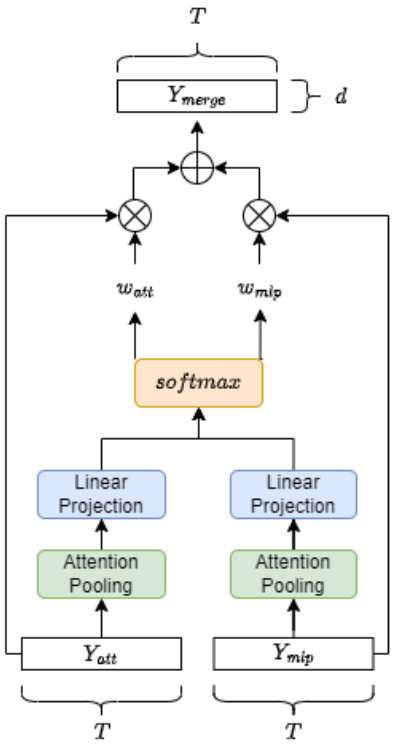

Figure 3: Detail of weighted average merging

爲了讓模型更好解釋與增加修改性，提出了加權平均的做法，這個做法讓模型去學習分支的重要性，結果顯示模型的行爲會傾向一開始取得較多全域的資訊而到中後期才更關注局部資訊，這樣的結果與 conformer 自注意力模塊和捲積模塊的擺放方式吻合。加權平均實際步驟如圖三：

1. 使用 attention pooling 總結每一分支的輸出成單一的向量

$$\mathrm{y}_{\mathrm{att}} = \mathrm{AttPooling}(\mathrm{Y}_{\mathrm{att}}) \in \mathbb{R}^{d} \tag{9}$$

$$\mathrm{y}_{\mathrm{mlp}} = \mathrm{AttPooling}(\mathrm{Y}_{\mathrm{mlp}}) \in \mathbb{R}^{d} \tag{10}$$

2. 將兩個分支總結的向量乘上線性轉換矩陣成單一的數值

3. 將上一步的數值經過 $softmax$ 得到分支權重

$$w_{\mathrm{att}}, w_{\mathrm{mlp}} = \mathrm{softmax}(\mathrm{W}_{\mathrm{att}}\mathrm{y}_{\mathrm{att}}, \mathrm{W}_{\mathrm{mlp}}\mathrm{y}_{\mathrm{mlp}}) \quad where\ \mathrm{W}_{\mathrm{att}}, \mathrm{W}_{\mathrm{mlp}} \in \mathbb{R}^{1 \times d} \tag{11}$$

4. 將分支呈上權重後相加即爲最後合併的輸出

$$\mathrm{Y}'_{\mathrm{merge}} = w_{\mathrm{att}}\mathrm{Y}_{\mathrm{att}} + w_{\mathrm{mlp}}\mathrm{Y}_{\mathrm{mlp}} \in \mathbb{R}^{T \times d} \tag{12}$$

爲了讓模型在推理階段時速度加快，在推理的時候剪掉注意力分支，因此訓練的時候使用 branch dropout 的技巧，讓注意力分支以一定的機率權重爲 0。

3 實驗結果以及討論

3.1 實驗設定

本次實驗使用 ESPnet (Watanabe et al., 2018) 作爲訓練工具，使用 24 層 Branchformer 當作編碼器與 6 層 transfomer 當作解碼器，本次實驗有使用 Aishell-1 (Bu et al.) 178 小時中文語料作預訓練，並比較無預訓練的組合，實驗中文語料對客語的辨識是否有幫助。

客語語料共 59 小時，共 76 位語者，將資料拆成 train、dev、test 三個集合分辨用於訓練、評估、測試模型，詳細資料如表 1.：

Sets	Duration	Spk	Syllables
train	48	62	324 465
dev	5	7	37 602
test	5	7	36 636

Table 1: 客語語料各資料集音檔長度 (小時)、語者數與音節數

使用之預訓練中文語料 Aisell-1 爲 178 小時中文語料，共有 400 位語者，以 44.1K 取樣率與 16bits 位元深度之高保眞麥克風收音並降取樣至 16k 與 iphone 和 Android 手機進行收音，詳細資料如表 2. 所述。

Sets	Duration	Spk	Sentences
train	150	340	120 098
dev	10	40	14 326
test	5	20	7176

Table 2: Aishell-1 語料庫各資料集音檔長度 (小時)、語者數與句數

3.2 實驗結果及討論

實驗分爲兩個方向，預測目標爲字元或者拼音和是否使用中文的預訓練模型，目的在驗證對於不同預測目標，客語語音辨識在實作上能夠達到如何的效能，另外，使用中文語料預訓練也可以驗證中文對漢語方言的語音辨識是否有幫助。

3.2.1 預測目標爲字元或拼音

這裡我們根據兩種不同的預測目標，分別爲字元與拼音，在同樣的模型下實驗模型的好壞，在本實驗沒有食用預訓練模型。

表 3. 和表 4. 分別爲以自原爲目標和以拼音爲目標的模型效能，由於字元是以單字去計算因此使用 CER (Character Error Rate, 字元錯誤率) 來評估模型效能，而拼音因爲需要以字串表示單一一個音因此需要以 WER (Word Error Rate, 詞錯誤率) 來評估。實驗結果，以測試集等未看過的資料進行討論，在字元作爲目標的情況下，其字元錯誤率爲 5.1%，略遜於以拼音作爲目標的字錯誤率 4.2%，因爲拼音的所有可能少於字元，可以避免選錯字造成的錯誤，因此效果會比較好。

	training	dev	test
CER	0.6	6.9	5.1
WER	1.37	5.5	4.2

Table 3: Error rate of the model in recognizing character and Pinyin

3.2.2 是否使用中文預訓練

本實驗使用 Aishell-1 中文語料進行預訓練，此處由於中文和客語所有可能的漢字不同，因此需要將輸出層替換。

由於 Aishell-1 是以字元作爲目標，因此客語的實驗也以字元爲目標進行，驗證中文對於客語的辨識是有幫助的。

表 5. 與表 6. 分別爲使用預訓練模型與否的模型效能，實驗發現，有使用預訓練的 CER 爲在驗證集上可以達到 3.2% ，測試集可以達到 2.3%，相比沒有使用的組別在驗證及爲 6.9% 和測試集的 5.1%，實驗證明中文語料對於客語辨識任務有不小的幫助。

	training	dev	test
w	0.5	3.2	2.3
w/o	0.6	6.9	5.1

Table 4: CER of Hakka corpus with and without Aishell-1 pretrained model

4 結論

這次實驗初探了客語的語音辨識，使用了目前主流的 ESPnet 作爲訓練架構，並使用在中文上有不錯表現的 Branchformer 當作模型，並在乾淨的語音下進行整個實驗，結果發現在客語的情況之下效果不錯，並且也實驗得到中文對於客語的學習有幫助，藉此可以推測漢語方言也可以以類似的方式來做實驗。經過了這次的實驗希望之後在客語方面的語音辨識可以達到更加實用的場景，再更多條件的環境下進行辨識。

Acknowledgments

很感謝在過程中有江振宇老師的指導，在我遇到困難的時候給我建議和方向，也感謝實驗室的李武豪學長，在遇到技術上的問題時總能給予我幫助，最後要感謝客家委員會和國立陽明交通大學的人工智慧語音研發中心舉辦這次的比賽並提供客語資料庫，讓我在過程中可以學習到很多。

References

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Wikipedia. 2023. Varieties of Chinese — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Varieties%20of%20Chinese&oldid=1178163589. [Online; accessed 05-October-2023].