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

NSYSU-MITLab 之語音辨識系統於 Formosa Speech Recognition Challenge 2023
NSYSU-MITLab Speech Recognition System for Formosa Speech Recognition Challenge 2023

Hong-Jie Hu
NSYSU
jie12504469@gmail.com

Chia-Ping Chen
NSYSU
cpchen@cse.nsysu.edu.tw

摘要

在本論文中將會描述 NSYSU-MITLab 在本次 Formosa Speech Recognition Challenge 2023 (FSR-2023) 所使用的語音識別系統。我們使用了預訓練模型 wav2vec2.0 再加上 Enhanced Branchformer 與我們對其改進的 Dynamic Convolution Enhanced Branchformer ，構成我們在 Track-1 客語辨識漢字任務以及 Track-2 客語辨識拼音任務的參賽模型，並用兩者中表現較好的作爲最終輸出結果的系統。最終，我們將以模型 wav2vec2.0 + Enhanced Branchformer 作爲客語辨識漢字任務的輸出系統，在決賽測試集下 CER 爲 52.2% 。模型 wav2vec2.0 + Dynamic Convolution Enhanced Branchformer 將會作爲拼音任務的輸出系統，同樣在主辦方給予的測試集下 SER 爲 46.8%。

Abstract

In this study, the speech recognition system will be introduced, used by NSYSU-MITLab, for Formosa Speech Recognition Challenge 2023. We use the pretrained model wav2vec2.0 as the frontend module in both the Enhanced Branchformer system and the Dynamic convolution Enhanced Branchformer system. The Dynamic Convolution Enhanced Branchformer system is refined from the Enhanced Branchformer system. We will choose the best CER one in the Taiwanese Hakka Recommended Characters test set, as the final model for Track-1. The system will be the final model for Track-2, with lower WER in the Taiwan Hakka Pinyin test set. Both test sets are released by organizers. Finally, we decided on the wav2vec2.0 + Enhanced Branchformer system for Track-1 which gets 52.2% CER on the final test. The wav2vec2.0 + Dynamic convolution enhanced branchformer system is chosen for Track-2 which gets 46.8% SER on the final test.

關鍵字：語音識別、客家語

Keywords: Auto speech recognition, Enhanced branchformer, Dynamic convolution

1 Introduction

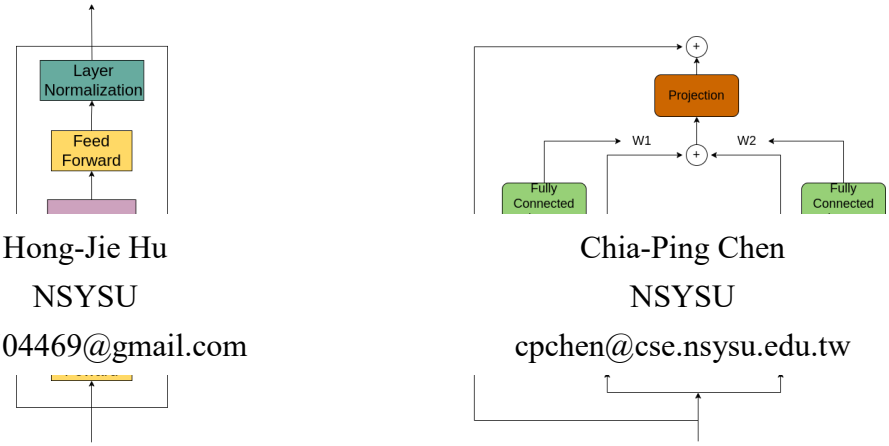
語音識別是一種將輸入的音檔轉成對應文字的任務。模型可以從訓練資料集學習與分析特徵，從而解析出最有可能的文字。近年來，在語音識別任務的領域裡，利用端到端的架構可以使該任務獲得很不錯的成績，因此這樣的架構逐漸變得熱門，例如：卷積神經網路 (Convolution Neural Networks,CNNs) 以及 Transformers。卷積神經網路的核心概念爲藉由卷積核，對輸入特徵進行卷積，模型能藉由參數學習如何最佳化卷積核，從而擷取 local feature 又或者說是高頻特徵。Transformers 使用自注意力機制，藉由計算在時間或序列上特徵間的相互關係，擷取 global features 又或者說是低頻特徵。

近年來，在台灣會說客家語的人逐漸的減少，客家語逐漸成爲一種失落的語言，客家文化也逐漸沒落。因此，客語語音識別系統會是是維護客家文化的一大關鍵。我們很榮幸參與了這次 2023 年的 Formosa Speech in the Wild (FSW) 計畫中的客語語音識別競賽 Formosa Speech Recognition Challenge 2023 (FSR-2023)。在本次競賽中，我們將會使用 Enhanced Branchformer (E-Branchformer) 以及對 E-Branchformer 進行改進的 Dynamic Convolution Enhanced Branchformer (DCE-Branchformer) 參與客語語音識別競賽中 Track-1 客語辨識漢字任務以及 Track-2 客語辨識拼音任務。漢字任務會是從客語語音辨識出其漢字結果，例如：今晡日係拜二。而拼音任務則是從客語語音辨識出其對應的拼音結果，例如：gim24 bu24 ngid2 he55 bai55 ngi55。

E-branchformer 是由 Branchformer (Peng et al., 2022) 改進而來，Branchformer 的結構

Hong-Jie Hu
NSYSU
jie12504469@gmail.com

Chia-Ping Chen
NSYSU
cpchen@cse.nsysu.edu.tw



(a) Conformer encoder (b) Branchformer encoder

- 圖 1. 上方爲 Conformer encoder layer 與 Branchformer encoder layer 結構示意圖。

如圖 1. 所示。Branchformer 的核心概念爲爲了研究 local context 與 global context 在各個 encoder layer 間重要程度是否應該像 Conformer (Gulati et al., 2020) 同等的重要，對其結構做了較彈性的調整，並觀察 w_1 與 w_2 在各層 encoder 中的狀況。Conformer 在語音辨識任務上普遍獲得很不錯的成績，其 encoder 的核心結構爲在自注意力層的後面再加入卷積層，能夠使模型同時考慮 local 資訊與 global 資訊。但由於自注意力層與卷積層是串接的的關係，如果 global 資訊是由自注意力層擷取，而 local 資訊是從卷積層得來，這可能會使模型對於 local 資訊與 global 資訊上的考量不夠彈性，因爲對每層的 encoder 來說，local 資訊與 global 資訊是同等重要的。因此，爲了能較彈性的調配 local 資訊與 global 資訊的重要性，Branchformer 自注意力層與卷積層改爲並接的方式結合，並給予兩個 branch 輸出權重後再相加，再經過一層全連接層，對兩個 branch 的輸出做投影。權重的計算由兩層可訓練的全連接層而來。E-branchformer 則是類似於 Conformer 與 Branchformer 兩者間的折衷。

在本篇論文中將會比較 E-branchformer 與 DCE-Branchformer 的在客語辨識競賽的表現。總共會有六個部份：第一個部份是 Introduction; 第二個部份是 Electronically-available resources，會簡單描述實驗的硬體規格; 第三個部份是 Method，會在此章節介紹本篇論文實驗中會使用到的方法，如資料處理以及模型架構; 第四個部份是 Experiment，會在此章節介紹使用的資料集，以及模型的一些參數設定; 第五個部份是 Result，會講述實驗的結果與發現; 第六個部份是 Conclusion，爲本篇論文的總結。

2 Electronically-available resources

CPU: Intel(R) Core i5-10400 2.90 GHz
CPU cache size: 12 MB
RAM: 40 GB
GPU: 2xNVIDIA GTX 1080 Ti
GPU VRAM: 2x11GB

3 Method

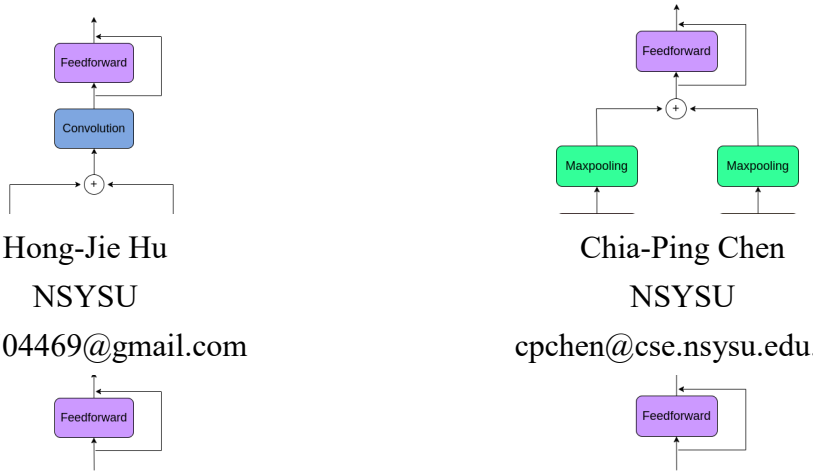
在這個章節將會介紹本次競賽所運用到的各種方法，如前處理模組、主模型架構和訓練方法。在本次競賽中主要使用的模型爲 Enhanced branchformer (E-branchformer) (Kim et al., 2023) 以及對其 Multilayer perceptron(MLP) (Sakuma et al., 2021) 改進的 Dynamic convolution enhanced branchformer (DCE-branchformer) 。此外，考慮到訓練資料可能不夠，參考了此篇論文 Zhang et al. (2020) 的作法，使用預訓練模型 wav2vec2 (Baevski et al., 2020) 做爲模型的前處理模組，並用一層全連接層做特徵轉換，最後再將這些特徵送入 E-branchformer 與 DCE-branchformer ，進行模型 finetune，來達到更低的 CER 與 WER。

3.1 前處理模組

在聲音特徵送入 Encoder 前，一般會先需要先經過資料前處理，如濾波、頻譜轉換和位置資訊編碼..... 等等，待資料前處理後，才會送入模型進行訓練。由於這些步驟在預訓練模

Hong-Jie Hu
NSYSU
jie12504469@gmail.com

Chia-Ping Chen
NSYSU
cpchen@cse.nsysu.edu.tw



(a) E-Branchformer encoder 的結構圖　　(b) DCE-Branchformer encoder 的結構圖

- 圖 2. Enhanced Branchformer layer 與 Dynamic convolution enhanced branchformer layer。上方爲 E-Branchformer 與 DCE-Branchformer encoder 的結構: (a) Branchformer encoder 主要分爲五種區塊: 兩個 feedforward 區塊、兩個 layer normaliztion 區塊、一個多頭自注意力區塊、和一個卷積區塊。E-Branchformer 的運算步驟如下：1. 輸入會先經過一層 feedforward 層。2. local 資訊與 global 資訊分爲兩個 branch 來做處理，，先各自做 layer normaliztion ，再送入自注意力層與 CGMLP 層。3. 將特徵串接爲原來 feature dimension 的兩倍。4. 透過 pointwise convolution 將特徵維度降維 5. 最後，在經過一層 feedforward 層 (b) DCE-Branchformer 的運算步驟大致與 E-branchformer 相同，差在 local branch 使用動態卷積處理 local 資訊，且合併兩個 branch 的方式改爲 maxpooling。

型中已經完成或是有了替代的方式，我們可以省略上述資料預處理的步驟，直接使用預訓練模型的輸出作爲整體 E-branchformer 與 DCE-branchformer 中的輸入。本次實驗中使用的預訓練模型 wav2vec 2.0 由約 960 小時的英文資料集 Librispeech (Panayotov et al., 2015) 訓練而成。利用 wav2vec 2.0 作爲模型前端將原本連續的聲學特徵，轉爲離散的聲學特徵，在資料量較少的情況下對模型 finetune 也有不錯的效果。

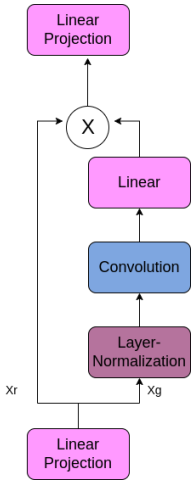
3.2 模型架構

在本篇論文中主要使用的模型爲 E-branchformer 和對其進行改進的 DCE-branchformer。

3.2.1 Enhanced branchformer encoder

E-branchformer 對於 local context 與 global context 的處理方式和 Branchformer 相同，如圖 2. 所示。都是用 Convolutional Gated MultiLayer Perceptron (cgMLP) (Rajagopal and Nirmala, 2021) 與 self-attention (Vaswani et al., 2017) 來處理。

cgMLP 的運算方式如圖 3. 所示。cgMLP 的結構爲在兩層用來做特徵投影的全連接層中



- 圖 3. cgMLP 結構圖。Linear projection 皆爲一層的全連接層所構成。第一層 Linear projection 的輸出會沿著特徵維度平分爲 X_r 與 X_g。X_g 經過 Layer-normalization、卷積層以及一層全連接層後，X_g 與 X_r 內積，最後經過最後一層 Linear projection ，將特徵維度投影回原特徵維度。

間，再夾了一層 Convolutional Spatial GatingUnit (CSGU)。cgMLP 使用 CSGU 對局部特徵做擷取。

self-attention 的運算方式如下

$$Matrix = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Hong-Jie Hu
NSYSU
jie12504469@gmail.com

在，d_k 是 multi-head 的 level 數，用來避免某段關係相關性過高，主宰 *Softmax* 中的結果。

Branchformer 與 E-branchformer 兩者間最大的不同在於，合併兩個 branch 輸出的方式並非像 Branchformer 使用全連接層，而是直接將兩個 branch 的輸出並接後，再用卷積將其沿著特徵維度降維合併。

3.2.2 Dynamic convolution enhanced branchformer encoder

E-branchformer 在 local branch 上使用 cgMLP 擷取 local 的特徵，其表現雖然十分亮眼，但合併兩個 branch 的方式說不定還有改善的空間。當在不使用預訓練模型時，不確定是否因爲 batch size 不夠多的關係，直接使用 Branchformer 與 E-branchformer 進行訓練時，在拼音任務上訓練出現一些困難，準確度的收斂狀況不佳。原因可能在用卷積的方式合併兩個 branch 的輸出，雖然對比 Branchformer 的合併方式，兩個 branch 間在特徵維度上的關注度能有更彈性的考量，但或許因爲卷積的參數不夠靈活，可能間接使合併的過程中對於 local 資訊與 global 資訊較特別的資料會有錯誤的偏重。

因此，我們嘗試將 cgMLP 更換爲動態卷積 (Wu et al., 2019)，並簡化合併兩個 branch 的方式。簡化後的合併方式爲對兩個 branch 沿著特徵維度直接 maxpooling 成原本的一半後，再相接起來，這樣可以避免合併的方式對特徵擷取做了過多的干涉，確保 local branch 與 global branch 關注度不會有過於極端的偏重。

3.2.3 Decoder

Decoder 將會藉由 Encoder 的 K 與 V，利用 masked self-attention 計算從 encoder 輸出的向量與目前已經輸出的文字間的相互關係。masked self-attention 的核心概念爲了讓解碼的過程是由前往後的，需要避免相關性的計算會考慮到尚未解出的字，因此需要對相關性矩陣與一個上三角矩陣內積，屏蔽不需要的相關性計算。

4 Experiment

在這個章節將會描述我們在本次客語競賽中使用的資料集以及模型建置的相關參數。

4.1 Dataset

Chia-Ping Chen
NSYSU
cpchen@cse.nsysu.edu.tw

外還共計 8 次的語料作爲測試集，剩餘的語料則爲訓練集。所有音檔皆爲單聲道，取樣率爲 16kHz。

4.2 Data Augmentation

在本次實驗中，使用變速擾動 (Speed-Perturbation) (Ko et al., 2015) 對我們的訓練資料做資料增強，使我們的模型更強健，降低模型過學習發生的機會。變速擾動會將訓練音檔在不破壞原始音頻的情況下，以特定比例調整音訊速度。

4.3 Model setup

在本次實驗中，我們使用了工具包 ESPnet (Watanabe et al., 2018)，建構我們的實驗模型。E-branchformer 與 DCE-branchformer 的參數設定大致相同。Encoder 的層數皆爲 16 層，decoder 的層數皆爲 6 層。兩者的 batch size 皆爲相同。E-branchformer 的 cgMLP 卷積核大小爲 16。DCE-branchformer 動態卷積卷積範圍爲 101。優化器皆使用 Adam Optimizer (Kingma and Ba, 2014)。Learning rate 皆爲 0.0015。

5 Result

在 Track-1 客語辨識漢字任務中，wav2vec2.0 ＋ E-Branchformer 在 FSR-2023-Hakka-Lavalier-Train 測試集下的 CER 爲 4.1%。由於沒辦法在期限內訓練好 Track-1 的 wav2vec2.0 ＋ DCE-branchformer，我們將以 wav2vec2.0 ＋ E-Branchformer 的輸出結果作爲我們 Track-1 的最終輸出結果。

在 Track-2 客語辨識拼音任務中，我們有嘗試不用預訓練模型進行訓練，結果如表 1. 所示。E-Branchformer 在 FSR-2023-Hakka-Lavalier-Train 測試集下的 SER 爲 22%, 而 DCE-Branchformer 的 SER 爲 14.5%。從此結果能得知，在沒有使用預訓練模型的情況下，這樣的改進能降低 7.5% 的 SER，同時參數量也從 42.94M 降低爲 37.67M，似乎是還能接受的改進。

- 表 1. 不使用預訓練模型下在 Track-2 的訓練結果。

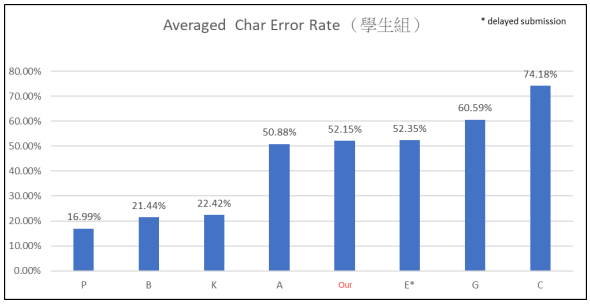
Model Type	Model size	SER
E-Branchformer	42.94M	22.0
DCE-Branchformer	37.67M	14.5

Hong-Jie Hu
NSYSU
jie12504469@gmail.com

wav2vec2.0 ＋ E-Branchformer	7.0
wav2vec2.0 ＋ DCE-Branchformer	6.9

- 表 3. Track-1 客語辨識漢字任務在朗讀資料與自發性語音資料上的測試結果。

Data Type	CER
Reading	22.6
Spontane	56.7
Avg	52.2



- 圖 4. Track-1 客語辨識漢字任務的決賽排名結果。

使用 wav2vec2.0 作爲預訓練模型後，訓練的結果如表 2. 所示。wav2vec2.0 ＋ E-Branchformer 與 wav2vec2.0 ＋ DCE-Branchformer 在 SER 上最高降了 15%，兩者間的差距降爲只有 0.1% SER 的差距。

Track-1 客語辨識漢字任務的決賽排名與結果如圖 4. 與所示。決賽辨識資料由兩種客語資料組成，一是朗讀資料 Reading，另一個是自發性語音 Spontane，各自的結果如表 3. 所示。

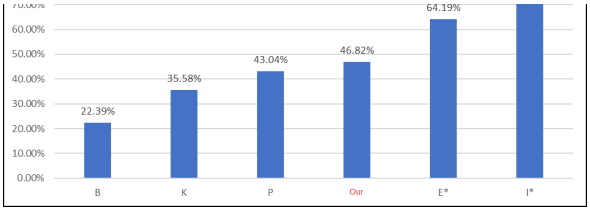
Track-2 客語辨識拼音任務的決賽排名與結果如圖 5. 與所示。決賽辨識資料與 Track-1 一樣有 Reading 與 Spontane 兩種語料，各自的結果如表 4. 所示。

從決賽結果能發現與 FSR-2023-Hakka-Lavalier-Train 測試集有很大的差異。FSR-2023-Hakka-Lavalier-Train 測試集下，Track-1 的 CER 爲 4.1%，而決賽的 CER 爲 52.2%。Track-2 的 SER 爲 6.9%，但決賽的 CER 爲 46.8%。造成如此差異的原因可能主要在訓練資料量不足，以及訓練資料與決賽辨識資料的差異過大，使得模型不善於辨識資料集以外的資料，其差異如表 5. 所示。

- 表 4. Track-2 客語辨識拼音任務在朗讀資料與自發性語音資料上的測試結果。

Data Type	SER
Reading	16.8
Spontane	51.2

Chia-Ping Chen
NSYSU
cpchen@cse.nsysu.edu.tw



- 圖 5. Track-2 客語辨識拼音任務的決賽排名結果。

- 表 5. FSR-2023-Hakka- Lavalier-Train 測試集與決賽集平均每段音檔對照。

Data Type	平均字數	平均秒數	字數/秒數
FSR	19.86	9.68	2.05
決賽集	32.6979	10.368	3.15

6 Conclusion

從實驗結果能發現，在 Track-2 客語辨識拼音任務中，使用了預訓練模型 wav2vec2.0 後，E-Branchformer 與 DCE-Branchformer 間 SER 的差距縮小很多，原因很可能在於訓練資料不足。同時，也能發現到，這樣的改進適合用在訓練資料較不足的情況，原因推測可能是動態卷積相對 cgMLP 更能多考慮一些頻率較低的高頻特徵。因此在資料量較少時，低頻特徵相對高頻特徵不易學習的情況下，對於高頻特徵的考慮範圍更廣更彈性的動態卷積因此得到優勢。而 wav2vec2.0 因爲會將原本連續的特徵化簡爲離散特徵，因此對於 local 資訊學習能力相對較強 E-Branchformer 能有較大的效能提昇。

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460.

preprint arXiv:2005.08100.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe. 2023. E-branchformer: Branchformer with enhanced merging for speech recognition. In 2022 IEEE Spoken Language Technology Workshop (SLT), pages 84–91. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

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.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In International Conference on Machine Learning, pages 17627–17643. PMLR.

A Rajagopal and V Nirmala. 2021. Convolutional gated mlp: Combining convolutions & gmlp. arXiv preprint arXiv:2111.03940.

Jin Sakuma, Tatsuya Komatsu, and Robin Scheibler. 2021. Mlp-based architecture with variable length input for automatic speech recognition.

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.

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.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430.

Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le,

Hong-Jie Hu
NSYSU
jie12504469@gmail.com

Chia-Ping Chen
NSYSU
cpchen@cse.nsysu.edu.tw