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

The DMS-ASR System for the Formosa Speech Recognition Challenge 2023

Hsiu-Jui Chang, Wei-Yuan Chen
Delta Management System, Delta Electronics, Inc., Taiwan
{ryan.hj.chang weiyuan.wy.chen}@deltaww.com

Abstract

This report primarily describes the techniques employed in the Formosa Speech Recognition Challenge 2023 (FSR-2023). In the context of this Hakka language speech recognition competition, we compared two methods for training speech recognition models. Specifically, we employed both fine-tuning of pretrained speech recognition models and direct training of end-to-end (E2E) models. Furthermore, we utilized data augmentation techniques, such as Multi-style Training (MTR) and spectrum augmentation (SpecAugment), to mitigate the impact of noise on recognition accuracy. Additionally, model weight averaging was employed to achieve improved results.

Keywords—Hakka automatic speech recognition, end-to-end speech recognition, pretrained asr model, finetune, model averaging

摘要

這篇報告主要描述我們在 Formosa Speech Recognition Challenge 2023(FSR-2023)使用到的技術。針對這次的客語語音辨識比賽，我們比較了兩種語音辨識模型訓練方式。我們分別使用了微調預訓練語音辨識模型的方法以及直接訓練端對端(End-to-End, E2E)的方法來訓練語音辨識模型。此外，我們使用資料擴增(Data Augmentation)，例如多型態訓練(Multi-style Training, MTR)和頻譜擴增法(SpecAugment) 來降低噪音對辨識的干擾，也使用了模型權重平均的方式使達到更好的結果。

關鍵字—客語語音辨識, end-to-end 語音辨識,預訓練模型 , 模型權重平均

1 Introduction

隨著深度學習模型的技術不斷突破，簡化了許多傳統語音辨識模型所需要的流程的端對端語音辨識模型已成為近年來研究主流。端對端語音常見的模型有注意力模型(Attention model) [1], 連結時序分類模型(Connectionist Temporal Classification, CTC) [2], 連結時序分類注意力混合模型(Hybrid CTC-Attention) [3,4]和序列轉換遞迴式神經網路(Recurrent Neural Network Transducer, RNN-T) [5]等模型。而基於自監督技術(SSL, self-supervised learning)的大型預訓練語音模型如[6,7]也改變了以往語音辨識模型的訓練方式。通過預訓練模型，只需要少量資料微調或是將預訓練模型的輸出作為特徵訓練模型，便能夠在各種語音相關下游任務如:語音辨識，語音翻譯，語音降噪等，得到顯著的效果。

這次 Formosa Speech Recognition Challenge 2023 是一個客語的語音辨識任務，總共分為兩種類型，如 Table I 所示。 輸入音檔的語言為客語，Track1 的輸出是要輸出文字、Track2 是要輸出拼音。

TABLE I. FSR-2023 OUTPUT FORMAT

Type	FSR-2023 客家語音辨識任務	
	輸出類型	例句
Track1	漢字	今晡日係拜二
Track2	拼音	gim24 bu24 ngid2 he55 bai55 ngi55

我們在此比賽中分別使用了 zipformer transducer[8]，wavlm-large[6]+conformer[9] ， branchformer[10] ，Whisper[11]進行實驗，而其中 zipformer transducer，wavlm-large+conformer ，Branchformer 訓練的是拼音模型，而 Whisper 則是微調成為客語文字辨識的模型。

2 Methods

2.1 Model Architecture

我們作法分為直接訓練端對端語音辨識模型以及微調預訓練模型兩種方式。直接訓練端對端語音辨識模型的部分，我們使用了多種模型架構，包括 Zipformer transducer 模型、Branchfromer CTC-Attention 模型以及使用預訓練模型 Wavlm-large 作為輸入特徵，再訓練 Conformer CTC-Attention 模型。微調預訓練模型任務我們則使用 AdaLoRA[12]的方式微調 Whisper-small 模型。

2.1.1 Connectionist Temporal Classification (CTC)

基於隱藏式馬可夫模型(Hidden Markov Model, HMM)的語音辨識系統在訓練聲學模型時通常需要額外處理對齊資訊，使用強制對齊(Forced-Alignment)的方式進行；而 CTC 是一種可以避開 Forced-Alignment 的一種方式，它列出所有可能的對齊輸出機率分布，最後輸出最可能的結果。

2.1.2 Sequence to Sequence Attention

Seq2seq 最初是為了機器翻譯任務而開發，而後來被用於語音辨識等任務，他由兩部分組成:編碼器(Encoder)和解碼器(Decoder)。輸入語音序列，編碼器將其轉成固定長度的隱藏層向量，解碼器接收編碼器隱藏層的向量生成輸出序列。編碼器通常使用循環神經網路(RNN) 或是卷積神經網路(CNN)處理輸入序列。然而這樣的模型在處理長序列時通常會忽略較早的訊息。因此注意力機

制的引入，對於每個解碼步驟，注意力機制計算一個權重分布告訴編碼器那些部分是當前最重要的，使解碼器在生成每個輸出標記時可以關注序列的不同部分。

儘管加入注意力機制在語音辨識中表現良好，他仍有一些限制，例如處理速度較慢以及並行性不佳。為了克服此問題，研究員引入了 Transformer 模型[13]。Transformer 是一種基於自注意力機制（Self-Attention）的神經網路架構，它可以併行處理輸入序列的不同部分，因此在處理速度上具有顯著優勢。此外，Transformer 的模型架構能夠捕獲更長距離的依賴關係，在語音辨識中也帶來了優異的辨識效果，後續更有不少基於 Transformer 改進的模型，如:將 CNN 結合 Transformer 的 Conformer 、具有 Temporal U-Net 結構的 Squeezeformer[14]、分為兩個平行分支去分別提取全域特徵和局部特徵的 Branchformer 和基於 Conformer 以及 Squeezeformer 的架構修改的 Zipformer。近期 OpenAI 發布的 Whisper 模型也是基於 Transformer 的架構訓練多任務的語音模型。

2.1.3 Transducer

RNN-T [15]是基於 CTC 的一種改進方式，解決了 CTC 輸出之間條件獨立以及缺少語言模型能力的不足，讓語言模型和聲學模型可以同時在訓練時優化。而以往 RNN-T 會使用 RNN，我們則是使用了 Zipformer 作為編碼器，解碼器則使用了 stateless model[16]。

2.1.4 Hybrid CTC/Attention

通過結合 CTC 和 Attention，Hybrid CTC/Attention 模型在語音辨識中可以更好地處理不同長度的輸入序列，同時也能考慮到語音訊號的時序性。使得它的效果更優於單獨使用 CTC 或是 Attention 模型。

2.2 pretrained model

2.2.1 WavLM

近年來通用型模型受到學術及工業界關注，這種大型模預訓練模型通常可以在各式下游任務中取得優異的表現，在微調時也不需要太大量的語料。WavLM 是微軟提出的大型預訓練模型。不僅透過語音上進行遮罩預測任務學習了語音辨識相關的信息，還透過語音去噪提高了非 ASR 任務的潛力。我們認為將 WavLM 模型的特徵表示訓練新的 ASR 模型能夠更進一步提升現有模型效果。

2.2.2 Whisper

Whisper 是 OpenAI 基於 Transformer 架構訓練的語音辨識語及語音翻譯模型，能夠將多國語言轉成文字。由於運算資源有限，直接對所有參數進行微調訓練成本太高了。

近年來出現了更有效率的微調方式稱作 Parameter-Efficient Fine-Tuning (PEFT)。微軟提出的 Low-Rank Adaptation（LoRA）[17]便是其中一種方式。其原理是凍結原始的預訓練模型權重並搭配一個小模型微調。使用 LoRA 微調可以大幅降低訓練所需要的記憶體使用量。而我們使用了 LoRA 的改進方法 AdaLoRA[12]，這種方法使用了奇異值分解的形式對權重矩陣的增量更新進行參數化。然後根據新的重要性指標，透過操縱奇異值，在增量矩陣之間動態地分配參數預算。而在[12]中提到通常 AdaLoRA 的效果會優於 LoRA。

3 Experiments

3.1 Experiment Settings

本篇使用到的語料 FSR-2023-Hakka 為錄製語料，收集來自台灣各地的腔調。我們總共使用約 60 小時的語料共 20613 句，我們將隨機抽取其中 600 句作為驗證集。

聲學模型：我們分別在拼音的任務中訓練 Zipformer transducer 模型、Branchfromer CTC-Attention 模型以及使用預訓練模型 Wavlm-large 作為輸入特徵訓練 Conformer CTC-Attention 模型，其中 Zipformer transducer 有額外做 Multi-Style Training，其餘的模型只有速度擾動再加上 SpecAugment。而文字的任務中我們則使用了 Whisper-small 模型加上 AdaLoRA 微調的方式。全部都沒有額外使用任何語言模型進行解碼，我們在這個任務沒有額外做任何資料擴增。

在資料擴增方面，我們進行速度擾動，再分別進行以下 2 個方面的處理：

1．使用 SpecAugment [18]直接在神經網路的特徵進行強化，設定如下 Time_warp： max_time_warp=5、Freq_mask： F=30, n_mask=2、Time_mask：T=40, n_mask=2。

2．使用傳統的 Multi-Style Training 的方式，將噪音和迴響和訓練語料的音檔中，噪音和迴響的語料是使用 MUSAN[1]、RIRS[2]。

3.2 Experiment Results

由於我們僅有兩張 GTX TITAN X 進行實驗，較難以進行更多不同實驗。我們做比較多實驗在 Track2 上。以下我們將分別討論文字及拼音的辨識結果。

3.2.1 Track1 Results

由於運算資源限制我們只能夠微調訓練 Whisper small 模型，在熱身賽的表現為字錯誤率 16.78%略優於 espnet+wavlm 的結果。 而在決賽中我們的平均結果 CER 只有 39.79，決賽語料可以分為朗讀與口語，其中我們在朗讀的部分字錯誤率為 21.15 而口語的部分由於訓練與測試資料差異的關係字錯誤率 42.67。相關數據如下：

TABLE II. RESULT – TRACK1

	Setting	CER
A	Whisper-small+AdaLoRA	16.78
BSL-1	Espnet+wavlm	17.11

TABLE III. RESULT – TRACK1 FINAL

Averaged	Reading	Spontaneous
39.79	21.15	42.67

3.2.2 Track2 Results

此次測試的輸出為拼音，結果如下表所示。在熱身賽中，我們提交了 Branchformer+CTC-Attention 模型的結果。隨後，我們訓練了 Zipformer transducer 以及 Wavlm+conformer CTC-Attention 模型。在 Wavlm+conformer CTC-Attention 模型中，我們取得了最佳結果，TER 為 5.9。至於決賽語料，我們使用了 Wavlm+conformer CTC-Attention 模型作為最終辨識結果提交。朗讀部分的識別效果達到了 TER 10.9。然而，在

MUSAN[1]: A Music, Speech, and Noise Corpus
RIRS[2]: Room Impulse Response and Noise Database[2]:

口語部分的表現不佳，TER 達到 39.68，平均錯誤率為 35.99。

TABLE IV. RESULT – TRACK2

	Setting	**TER**
A	branchformer+ctc-attention	16.45
B	zipformer transducer	9.98
C	wavlm-large+conformer ctc-attention	5.9

TABLE V. RESULT – TRACK2 FINAL

Averaged	Reading	Spontaneous
35.99	10.9	39.68

從實驗可以得知，隨著 End-to-End 語音辨識技術的演進，突破了以往需要大量語料才能夠有較好結果的限制。並且藉助於預訓練模型更有助於少量語料的訓練，然而訓練與測試資料仍然需要類型相似才不會導致辨識效果很差。

4 CONCLUSION

在這次比賽中，我們比較了各種 End-to-End 語音辨識模型辨識效果。此外，我們也透過資料擴增來降低噪音對模型的干擾。在實驗中發現確實使用預訓練模型可以使訓練出來的模型效果更好。然而我們最好的模型 Wavlm-large+conformer CTC-Attention 的解碼速相對太慢。Zipformer transducer 較符合實務需求。未來我們將針對使用預訓練模型的架構去研究如何改善推論速度。

REFERENCES

[1] William Chan, et al. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[2] Alex Graves, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Proceedings of the 23rd international conference on Machine learning, 2006.

[3] Shinji Watanabe, Takaaki Hori, et al. "Hybrid CTC/attention architecture for end-to-end speech recognition." IEEE Journal of Selected Topics in Signal Processing 11.8 ,2017.

[4] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. "Joint CTC-attention based end-to-end speech recognition using multi-task learning." 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017.

[5] Alex Graves. "Sequence transduction with recurrent neural networks." arXiv preprint arXiv:1211.3711,2012.

[6] S Chen, Chengyi Wang, et al. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing” IEEE Journal of Selected Topics in Signal Processing,2022.

[7] A Baevski, Yuhao Zhou, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations, NeurIPS, 2020

[8] Zipformer, https://github.com/k2-fsa/icefall/tree/master/egs/librispeech/ASR/pruned_transducer_stateless7

[9] Anmol Gulati, James Qin,et al. “Conformer: Convolution-augmented Transformer for Speech Recognition”, InterSpeech 2020

[10] Yifan Peng, et al. “Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding”, ICML,2022

[11] Alec Radford, Jong Wook Kim, et al. “Robust Speech Recognition via Large-Scale Weak Supervision”, arXiv, 2022

[12] Qingru Zhang, Minshuo Chen, et al. “Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning”, ICLR, 2023

[13] Ashish Vaswani, Noam Shazeer, et al. “Attention Is All You Need”, NeurIPS, 2017

[14] Sehoon Kim, Amir Gholami, et al. “Squeezeformer: An Efficient Transformer for Automatic Speech Recognition”. NeurIPS, 2022

[15] Alex Graves, “Sequence Transduction with Recurrent Neural Networks”, ICML, 2012

[16] Mohammadreza Ghodsi, Xiaofeng Liu,“Rnn-Transducer with Stateless Prediction Network”, ICASSP, 2020

[17] Edward J. Hu, Yelong Shen, “LoRA: Low-Rank Adaptation of Large Language Models”, ICML, 2021

[18] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." arXiv preprint arXiv:1904.08779 (2019).