Whisper Model Adaptation for FSR-2023 Hakka Speech Recognition Challenge
運用 Whisper 模型調適於 FSR-2023 客語語音辨識競賽

Yi-Chin Huang
Department of Computer Science and Artificial Intelligence
National Pingtung University
Pingtung city, Taiwan
ychuangnptu@nptu.edu.tw

Ji-Qian Tsai
Department of Computer Science and Artificial Intelligence
National Pingtung University
Pingtung city, Taiwan
aura01221@gmail.com

摘要

本文主要針對 FSR-2023 客語語音辨識比賽所提供的語料以及文字標記進行 Whisper 模型的調適訓練。Whisper 模型為基於自注意力機制的 Transformer 模型，其多樣性的語料以及弱標記的多任務訓練使其在多種語音相關任務中具有很好的強健性，尤其是在語音辨識的任務。本研究透過加入自行收集的客語語音語料以及透過不同大小的批次以及迭代式的模型訓練方法，嘗試獲得較穩健的客語語音辨識模型，以期在本次競賽中獲得不錯的結果。最終辨識結果在客文文字的辨識的任務中，在練習賽獲得學生組第三的成績。

Abstract

This study focuses on the adaptation and training of the Whisper model using the provided data and text annotations from the FSR-2023 Hakka Speech Recognition Challenge. The Whisper model is based on the Transformer architecture with a self-attention mechanism. Its diverse data and weakly labeled multitask training make it robust across various speech-related tasks, especially in speech recognition. In this research, we attempt to achieve a robust Hakka speech recognition model by incorporating self-collected Hakka speech data and employing different batch sizes and iterative model training methods. Finally, the recognition results of the Hakka speech-to-character task achieved a third-place ranking in the student division.

關鍵字：客家語音辨識、端到端語音辨識、Whisper 模型調適

Keywords: Hakka Speech Recognition, End-to-End Speech Recognition, Whisper Model Adaptation

1 Introduction

深度學習對於自動語音辨識（ASR）帶來了巨大的影響和優勢。傳統的 ASR 系統(Yu, 2017)以 HMM 為基礎，但在近年來的研究發展上，深度神經網路（DNNs），已經在語音辨識方面展現了相當大的進步。許多研究(Chan, 2016)已經證明，DNNs 相對於傳統方法在語音辨識的準確性上取得了巨大提升。此外，卷積神經網路（CNNs）和循環神經網路（RNNs）也被廣泛應用於語音辨識和合成(Shen, 2018)(Zen, 2016)，它們在處理聲音數據方面表現出色。由於深度學習領域的成功，引入了許多新的技術，像是注意力機制(attention mechanism)與延伸的 transformer 模型(Wolf, 2020)。注意力機制允許模型集中注意力於輸入資料中較為關鍵的部分；而 transformer 模型則有助於處理較長的序列資料。這些發展使語音處理系統的性能和多功能性更進一步提高，為不同領域的應用提供了更多可能性。

至於 Whisper 模型(Radford, 2022)，它是一種基於 transformer 的多任務模型，在語音辨識領域中有非常出色的表現，尤其是在沒有微調的狀況下便能有相當好的辨識結果。Whisper 在多種自然語言相關任務中表現也相當出色，即使在嘈雜的背景環境或多語言情境下也能保持高準確性。此外，Whisper 模型對於資料預處理和標記的要求相對簡單，這使它在語音處理中非常實用。它能夠處理多語言語音辨識、翻譯和語言識別，這是由於訓練時採用了多樣性的訓練數據和相應的標記。Whisper 降低了語料前處理的需求，簡化了語

音辨識流程，並且在無需微調的情況下，在許多測試語料上都表現得相當不錯。

在本篇研究報告中，主要目的是處理客語的語音辨識。客語的重要性在於根據客委會 105 年度全國客家人口暨語言基礎資料調查研究 (客家委員會, 2015) 指出，全國客家人推估約有 453.7 萬人，占全國 2349.2 萬人的 19.3%，是台灣第二大主要族群。但隨著時間的推移，年輕人對客語的使用和理解正在逐漸減少，這導致客語文化可能逐漸失傳。為了保護和保存這一重要的語言文化，為了保存客語的文化，十分感謝主辦方舉辦此次比賽，我們團隊也嘗試在加入自行收集的語料以及主辦方所提供的語料，基於 Whisper 模型的架構下，訓練出一套客語的語音轉文字的系統，希望能夠助於保存和傳承客語文化。

2 Whisper model 介紹

Whisper 是一個由 OpenAI 公司所開源的通用的語音辨識模型，它經過大量多樣化的語音資料訓練而成，可以應用於多種語言處理任務，包括多語言語音辨識、語音翻譯和語言識別。

此模型是基於 transformer 的序列轉序列 (Seq2Seq)架構，進行了多種語音處理任務的訓練，包括語音辨識、機器翻譯 (Johnson, 2017)、語音的語言識別 (Lopez-Moreno, 2014) 和語音活動檢測 (Zhang, 2012)。其中，為了達到一套模型處理上述的多種不同任務，研究人員對於收集的語料有以下幾個處理的方式，在此簡述。首先，針對於訓練數據，其中包含了不同環境、不同語言和不同說話者的多樣語音數據，這有助於訓練出更全面的語音辨識系統。

由於不同任務會有不同的標記方式，為了解決文本標記可能出錯的問題(研究團隊透過其他現有的辨識器來獲得可能的語音標記)，所有的語料使用了自動過濾方法，像是若辨識的結果較差，則不使用該語音、對於音檔的語言也有一定程度上的限制，如果文字是非英語的其他語言，則要求語音也必須是同一種語言；如果文字是英語，則語音可以是任何語言；另外，為了確保模型的強健性，若使用語音的文本跟常見的語音辨識語料所提供的相同，亦會濾除。

最後，將所有音檔切分為 30 秒 的片段，配上對應文本，作為模型的訓練資料，總共 68 萬小時的音檔，以及相對應的標記；其中，65%的資料（約 440k 小時）是英語語音及對應的英語標記；約 18%（126k 小時）的資料是非英語語音和英語文字標記；最後 17%（117k 小時）為非英語語音和相應的文字標記，這些非英語資料包括了 98 種不同的語言。

Whisper 使用了 Transformer 模型，也就是常見的 encoder-decoder 架構。聲音訊號的預處理包括將音訊檔案重採樣到 16000 Hz，並計算出 80 個頻道的梅爾頻譜，計算時視窗大小為 25ms，步長為 10ms。計算完梅爾頻譜後，將數值正規化到介於 -1 到 1 之間，作為輸入資料。對於每個音訊的 30 秒片段，因為每個區段為 10ms，所以有 3000 個時間點，形成 3000x80 的特徵。經由資料處理後，將 3000x80 的輸入資料通過兩個 1D 卷積層，得到 1500x80 的特徵。編碼器(encoder)部分包含 2 層 1D 卷積層，濾波器大小為 3，啟動函數為 GELU，第二層卷積的步長為 2；解碼器(decoder)部分則採用標準的 transformer decoder 結構，預期輸出不同任務的標記以及其對應預設的 prompt。

根據其論文中實驗的結果，研究團隊發現在不同語言之間的性能表現不均，特別是在資源有限或發現性較低的語言以及訓練資料較少的語言上，準確性較低。此外，在相同訓練資料量的情況下，某些語言（例如中文和韓文）的錯誤率相對較高(中文 WER 約為 20%)，可能因為這些東亞語言與主流語言差異較大，難以從多語言聯合訓練中受益，或是因為 Whisper 的 tokenizer 對這些語言不夠友好，因為其實驗是採用 BPE (Kudo, 2018) 作為標記方式。因此，本篇報告將著重於透過額外加入自行收集的四縣腔客語語料以及人工標記，針對不同的實驗方式來分析 whisper 在客語的辨識結果。

語料	總時長 (h)	語者個數	字數	字數(不重複)
主辦方提供語料(FSR)				
訓練集	47.45	60	274,750	2,781
驗證集	6.15	8	36,265	1,951
測試集	5.88	8	37,473	2,060
自行收集語料(NPTU)				
訓練集	20.94	3	184,330	3,085
測試集	3.32	1	26,747	1,952
FSR 練習賽				
測試集	10.01	11	57,101	2,369
FSR 正式賽				
測試集	17.03	未知	187,430	3,018

Table 1 本次所使用之相關語料統計

3 語料介紹

本次競賽分為練習賽以及正賽，主辦方也提供了訓練集給參賽者進行模型的訓練。另外也有提供兩種腳本來訓練模型，一種是基於端到端的 transformer 模型，採用 espnet 的工具(Watanabe, 2018)來實作，另一種則是前述的 whisper 為基礎的腳本，提供參賽隊伍使用語料來對原始模型進行調適。主辦方所提供的語料主要為四縣腔的語料，包含了 76 位語者的音檔，其中作為訓練集的有 60 位語者，其總時長約為 48 小時；驗證集的部分，有 8 位語者，約 6.15 小時；測試集的部分，同樣也是 8 位語者，約 8 個小時的時長。我們發現，部分訓練集與測試集中，會出現相同句子但是由不同語者所念的內容。

3.1 自行收集語料

由於本校位在屏東麟洛地區，離鄰近的六堆、美濃、高樹等地區相對接近，故有許多本校師長針對於本地的南四縣腔(六堆)客語的保存多有研究，像是由本校中文系的劉明宗教授所編的“美濃客家語寶典”，便收入了本地南四縣腔常用客家詞彙語對應之例句共 7,641 句，且具有中客文對應字句以及六堆客音標音，再加上客委會於其網站所公開的客語常用一百句跟客語認證(初級、中級、中高級)，共有約 6,640 句客語例句跟對應之中文句。

經由簡單的人工標記後，最終收集到的語料有兩位女語者跟一位男語者。這裡要特別指出，我們自行收集的語料為邀請客語薪傳師自行在家透過筆記型電腦的麥克風收集，因此可能會有背景雜音以及不清晰的狀況發生。Table 1 中呈現相關的語料統計。

3.2 語料相關統計與分析

從 Table 1 可看出，主辦方所提供之語料，語者個數相較起來較多，總時長約 60 小時。而我們自行收集的語料，由於多為詞語的範例句子或是客語片語或俚語，因此若單看字數，數量與主辦方語料相近甚至多一些(在不重複的狀況之下)，但時長約僅為其 1/3 的大小，可以了解我們額外所收集的語料特性較為單純而非日常的對話語料。練習賽的測試語料雖然時長有約 10 小時，大概是我們所收集語料時長的一半，但其所有字數卻不到我們所收集的語料的 1/3，再加上簡單聽音檔的內容可以發現語料是較為乾淨無雜訊的特性，對於語音辨識這個任務來說是較為簡單的狀況。

最後，正式賽的語料總時長有接近 17 小時，而字數甚至比我們自行收集的語料還多，代表其複雜度相較於練習賽來說高出許多。實際去聽語料可以發現，這些語料是有經過刻意的剪接，使得聲音聽起來較不平順；此外，語者的音調也跟訓練的語料不同，語調較為激昂，這些狀況也使得正式賽的語音辨識難度提升許多。

4 實驗設計與分析

4.1 實驗設計

剛開始比賽的時候，我們團隊嘗試透過基本的 Seq2Seq 的方式直接訓練客語的語音辨識模

	混合訓練			迭代訓練	
	FSR (batch: 16)	FSR (batch: 32)	FSR + NPTU	1st FSR 2nd NPTU	1st FSR 2nd NPTU*
FSR	7.73	**2.95**	9.56	14.32	10.98
NPTU	43.06	29.47	28.25	**25.98**	26.91
練習賽	12.00	**10.27**	24.06	16.47	15.47
正賽	82.22	**70.93**	71.17	72.84	75.94

Table 2 不同訓練設定在不同測試集的客語辨識 CER(%)結果

型，在相同的訓練語料所獲得的基底模型的狀況下，就相較使用 Whisper 模型的中文模型進行微調(finetune)以適應客語語音。結果發現 Whisper 模型的初始效果明顯好許多，驗證了原始 Whisper 論文的觀點，也就是在多樣性的語料下，的確在語音辨識的任務上，有較佳的強健性。因此，接續的實驗都採用 Whisper 模型來進行實驗。

在確定使用 Whisper 的框架下，為了調整出主辦方給的語料的最佳辨識效果，我們透過測試不同的批次大小(batch size)，去分析在不同測試集的效果，在此我們會拿主辦方提供語料的測試集部分，以及我們自行收集的語料的其中一部份來當作調適模型的依據。在此我們自行收集的測試集，並沒有拿進來作為訓練，僅單純測試，且故意挑選沒有出現在訓練語料的文字，希望可以讓調適的模型更加具有一般性，測試語料的數據可參考 Table 1。此外，由於 Whisper 模型較大且競賽時間有限，我們在本次競賽中，都是以 Medium 的模型(參數大小為 769M)來進行調整，主要採用 Nvidia A6000 顯示卡進行模型的調適。

由於 Whisper 為一個較大的模型，在以往的經驗中，要一次調適就獲得好的模型不容易成功，因此我們也嘗試將可以拿來調適的與料庫，分批次對於訓練好的模型進行迭代訓練(iterative training)，此外，由於我們自行收集的語料特性，可能會混淆已經訓練好的模型，故我們亦進一步地測試在調適時，只使用沒有出現在主辦方提供的語料中所出現的文字，希望是以補足原始客語模型不足之處。至於超參數的部分，我們主要調整批次大小，其餘則與原始模型保持相同。

4.2 結果分析

Table 2 顯示在不同訓練的設定之下，在不同的測試集所得到的 CER (Character Error Rate)。由於比賽時的測試語料並沒有正確的答案提供給參賽隊伍，在此我們便用主辦方提供的語料(FSR)以及自行收集的語料(NPTU)進行測試，並將表現最好的模型拿來辨識練習賽與正式賽的測試語料。在 FSR 的測試語料中，在使用批次大小為 32 時能獲得最佳的辨識率(2.95)，甚至還比加入我們自行收集的語料(NPTU)的效果還更好，這代表了兩邊的語料的特性差異甚大，混合訓練可能造成模型不穩定而混淆。

因此，我們在迭代訓練的設定時，皆是採用第一階段使用原本的 FSR 語料調整出來最好的模型，再用我們的語料進行第二階段的調適。其中 Table 2 中的迭代訓練，NPTU*表示在第二階段調適時，我們濾掉了跟 FSR 相同的文字音檔，希望降低混淆的效果。結果顯示，在 FSR 的測試集中，辨識效能的確有

所提升(14.32%變成 10.98%)。在我們自行收集的語料測試集中，迭代訓練可獲得最佳的結果(25.98%)，再次驗證我們自行收集的語料特性跟主辦方差異較大。由於練習賽測試語料與 FSR 語料相近，綜合以上分析，我們決定在測試賽時繳交批次大小為 32 的單純使用 FSR 語料訓練的模型，最終獲得了 10.27%，在學生組獲得第三名的成績。

正式賽的時候，由於時間來不及進一步調適我們的模型，故最後繳交的成績約為71%，具有非常多的錯誤，從正式賽的測試語料中分析，由於正式賽語料難度大幅提升，就算使用其他設定也沒有比較好，可能需要更為複雜的優化方式才能夠有好成績，像是背景雜音的擴充、抑或是加入一個較強的客文字的語言模型來幫助調整文字輸出、或是必須透過 VAD 的方式，將語音分段後再辨識等，有許多可能測試的方向。

5 結論

感謝主辦方辦理此次的客語語音辨識競賽，由於時程較趕，故無法精緻地調整模型來獲得較佳的辨識結果，希望明年度還有機會參加，並在本次的基礎上，調整出一套具有更加強健性的客語語音辨識模型。

References

Yu, D., & Deng, L. (2016). *Automatic speech recognition* (Vol. 1). Berlin: Springer.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4960-4964). IEEE.

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., & Szczepaniak, P. (2016). Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. *arXiv preprint arXiv:1606.06061*.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4779-4783). IEEE.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339-351.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., & Moreno, P. (2014, May). Automatic language identification using deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5337-5341). IEEE.

Zhang, X. L., & Wu, J. (2012). Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(4), 697-710.

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., ... & Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

客家委員會. 105 年度全國客家人口暨語言基礎資料調查研究. 典通股份有限公司, 2017.