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 NTNU Super Monster Team (SPMT) system for the Formosa Speech Recognition Challenge 2023 - Hakka ASR

Tzu-Ting Yang Hsin-Wei Wang Meng-Ting Tsai Berlin Chen
National Taiwan Normal University
{tzutingyang, hsinweiwang, mengting7tw, berlin}@ntnu.edu.tw

摘要

本篇論文旨在紀錄隊伍 Super Monster Team (SPMT)參加 2023 福爾摩沙客家語語音辨識競賽的臺文賽道之歷程。全程主要利用臺灣四縣腔的客家語語料庫進行語音辨識。近期國內使用客家語的人口佔據全國人口百分比僅約 5.5，且仍逐年下降，因而造成語料取得上的阻力；由於在客家文化群演變的歷史中，族群認同感較強，因此具有較強的語言獨立性。綜上所述，客語語料稀少，並且難以借鑒其他方言相互增益訓練，無疑加劇了本次客家語語音辨識競賽的難度。本次競賽我們結合了資料擴增、自監督學習特徵、語言模型以及語音活性檢測等方法，並與近期倍受關注的大型語音辨識模型 Whisper 比較。本文旨在了解各模組對客家語辨識效能的影響，並在最終取得了豐碩的成果，期望可以為我國之瀕危語言存續盡一份心力。

Abstract

This paper aims to record the progress of the NTNU Super Monster Team (SMPT) in the Formosa Speech Recognition Challenge 2023 (FSR-2023), which is the third event of the Formosa Speech in the Wild (FSW) project. The primary task was to recognize Hakka speech using a corpus of Hakka speakers in Taiwan. We present our participation results in Track 1: Taiwanese Hakka recommended characters speech recognition. Recently, the percentage of Hakka speakers in Taiwan is only about 5.5 percent of the total population, and is still decreasing year by year, which causes resistance in acquiring the corpus; due to the strong ethnic identity of the Hakka cultural group, it has a strong linguistic independence and exclusivity. In summary, the scarcity of Hakka paired-corpus and the difficulty of learning other dialects for mutual benefit have undoubtedly aggravated the difficulty of the FSR-2023. In this study, we try to investigate the interleaving effects of various components by integrating data augmentation, self-supervised learning features, large-scale speech recognition models, and language models to improve the performance of Hakka speech recognition. This article aims to explore the impact of various modules on Hakka speech recognition performance and has ultimately achieved fruitful results. We hoped that this effort can contribute to the preservation of endangered languages in our country.

關鍵字：客家語、語音辨識、FSR-2023
Keywords: Hakka, Speech Recognition, FSR-2023

1 簡介

本次語音競賽辨認目標為臺灣客家語系中的四縣腔。作為漢文的一個方言分支，客家語在 109 年《運用聯合國教科文組織(UNESCO)語言活力指標評估臺灣客語活力之研究》的調查中指出(張 et al., 2020) ，現今使用客家語的人口百分比僅約 5.5 ，且逐年下降，被評定為「嚴重瀕危」的等級。鑒於母語為文化發展的根本，為確保文化多樣性，母語的存續與否至關重要。

各種技術蓬勃發展的現今，許多語言學習的相關研究都能夠有效地幫助大眾理解和學習這些「瀕危母語」(Chen et al., 2016; Wang et al., 2022; Kheir et al., 2023)。在許多語言學習方法都依賴於優良的自動語音辨識模型作為基石的前提下，這項技術在語言保護上顯得更為重要(Zhang et al.,2021; Al-Ghezi et al., 2022)。儘管現今的語音辨識技術在諸如英語、中文

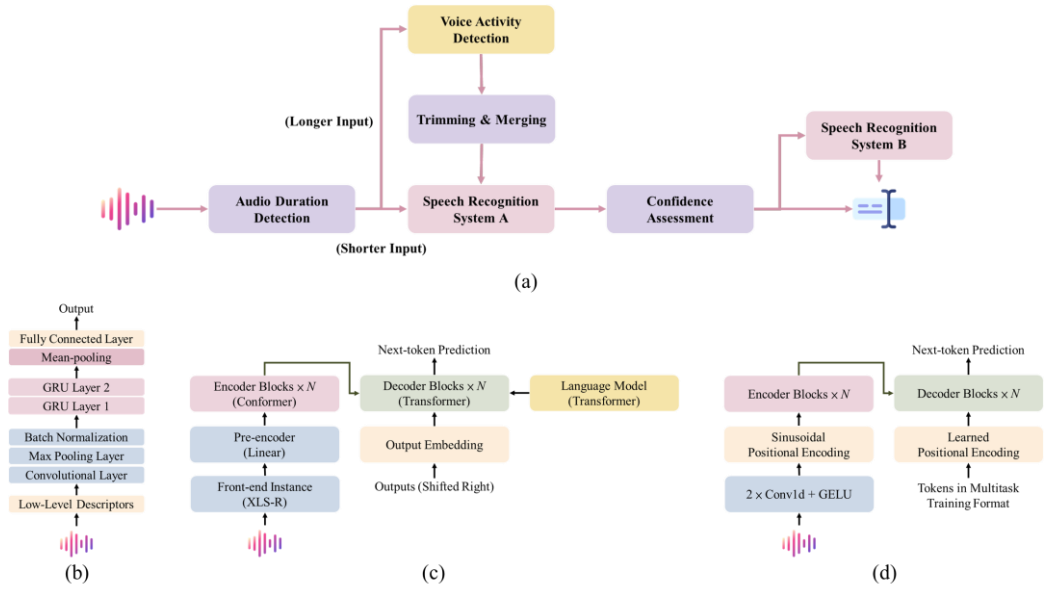

圖 1. 決賽提交模型整體架構。圖(a)為整體架構，其餘則分別為模組的詳細架構圖；圖(b)為語音活性檢測模型架構；圖(c)為自動語音辨識系統 A；圖(d)為自動語音辨識系統 B。

等主流語言的表現已可匹敵人類(Zhang et al., 2020; Xu et al., 2021; Gao et al., 2023; Zhou et al., 2023)，但在處理低資源語言時卻仍不盡理想(Bhogale et al., 2023; Sukhadia et al., 2022; Singh et al., 2023; Thomas et al., 2013; Chen et al., 2015; Dalmia et al., 2018; Xu et al., 2015; Miao et al., 2013;Müller et al., 2016)。不同於傳統 DNN-HMM(Hinton et al., 2012)模型的語音辨識方法，訓練深度神經網路模型需要大量的人工標記資料。然而，由於客家語並非台灣的主流方言，這使得取得客家語的語料變得極其困難。

在低資源語言的辨識任務中，經常利用主流語言協助低資源的語言種類進行辨識，這是為了藉由尋找相似的語法結構和單詞來協助低資源的語言進行語音辨識(Chen et al., 2023; Dalmia et al., 2018; Xu et al., 2015; Krishna, 2021; Chuangsuwanich, 2016; Chung et al., 2019)。然而，在客家文化群與其他文化群接觸的過程中，由於人我殊異及資源競爭，族群的認同感較強，相比其他漢文方言，客家語具有較強的語言獨立性(林, 2013)。這些因素都表明了客語語音辨識是一富有挑戰性的任務。

近期的研究(Zhao et al., 2022)指出，自監督模型所抽取的特徵在面臨低資源的情境時表現優越。許多研究利用少量資料微調預訓練的大型語音辨識模型也取得了豐碩的成果(Chung et al., 2019; Chung et al., 2020; Liu et al., 2020b; Liu et al., 2020a; Liu et al., 2021; Riviere et al., 2020; Schneider et al., 2019; Baevski et al., 2019; Baevski et al., 2020; Hsu et al., 2021)；語音活性檢測 (voice activate detection, VAD)的目的是分辨出音訊訊號中所包含的語音片段。因此，VAD 是語音辨識系統面臨真實應用場景時的必要模組(Alisamir et al., 2022)；先前的研究已證實語言模型(Hannun et al., 2014; Gulcehre et al., 2015; Chorowski et al., 2016; Kannan et al., 2018)及資料擴增(Ko et al., 2017; Park et al., 2019; Snyder et al., 2018)可以被用來改進語音辨識效能。

因此本次競賽中我們結合了自監督學習特徵、大型語音辨識模型、語音活性檢測、語言模型及資料擴增等方法，旨在探討各元件對辨識效果的影響。期望可以為我國瀕危語言的保護工作盡一份心力。

圖 1. 展示了我們在決賽提交模型的整體架構圖、語音活性檢測模型以及兩個自動語音識別系統。為了確保我們在各種情況下都能夠生成有效的辨識結果，在語音識別系統A未能產生結果時，將利用語音辨識系統B的輸出備援。語音識別系統 A 和語音識別系統 B 分別使用第三章所提及[H]和[G]的實驗設定。在後續的論文中，我們將在第二章詳細介紹比賽策略以及所採用的各類方法，並在第三章詳細介紹不同的實驗設定下的實驗結果。最後，第四章展示了我們的最終競賽成果。

2 策略及方法

我們在對輸入作各類型的資料擴增後，嘗試結合各類型的自監督學習模型來生成模型前端的輸入特徵，並在最後運用語言模型來融合外部資訊，以增加模型的泛化能力。考量到最終競賽時即興交談測資的不確定性，在測試時我們運用擁有一定降噪能力的語音活性檢測對輸入進行前處理，將變因控制在一定範圍內。最後將我們的模型與近期備受關注的大型語言模型－Whisper 進行比較。

2.1 資料擴增

由於訓練資料多為清晰的朗讀式錄音，因此我們在原音檔上疊加其他雜訊以進行資料擴增，用以增加現有訓練語料的多樣性和總量。我們同時採用混響和疊加其他音訊的方法對資料進行擴增。首先，我們使用 Kaldi(Povey et al., 2011)語音工具及公開語料[1]來生成模擬混響後的音檔。接著在疊加其他音訊時，則是使用 MUSAN 資料集(Snyder et al., 2015)，該資料集包含了超過 900 種噪音(不包括明顯可辨識說話內容的人聲)、42 小時的音樂以及 60 小時的多語言語音。

我們透過以下方法模擬各類雜訊，最終將訓練資料額外擴增了四倍：

- 混響(reverb)：通過卷積模擬房間脈衝響應(Room Impulse Response，RIR)，為乾淨的訓練集附加混響效果。
- 雜踏式噪音(babble)：從 MUSAN 語音中隨機選擇三到七位語者的音檔，將其合併後，再疊加到乾淨的訓練集音檔中，其信噪比為 13-20dB。
- 背景音樂(music)：從 MUSAN 中隨機選擇一個包含音樂音檔，根據需要進行修剪或重複，使其與原始信號的時長相匹配，然後疊加到乾淨的訓練集音檔中。信噪比為 5 至 15dB。
- 背景噪音(noise)：在整個乾淨的訓練集音檔過程中，每隔一秒鐘加入一次 MUSAN 噪音音檔。信噪比落在 0 至 15Db 之間。

2.2 自監督學習模型

自監督學習(Self-Supervised Learning, SSL)是一從輸入資料中萃取資訊，將其作為學習目標進行自我訓練的方法。利用大量訓練資料預訓練後輸出的自監督學習特徵，其對於包括了語音辨識在內的多項任務，皆可取得明顯的進展。SSL 方法與下游任務的結合主要分為兩個階段：首先，在第一階段會利用大量資料並以 SSL 方法建立自監督模型。自監督模型所生成的自監督特徵需盡可能地保留輸入資料的完整資訊。接著，在第二階段中，下游任務可以將這些自監督特徵作為模型輸入，根據任務需求進一步從中擷取有意義的資訊供後續的任務使用。

由於自監督模型經過大量語料的預訓練，我們認為它可以在一定程度上彌補低資源語言語料的不足(Zhao et al., 2022)。因此我們嘗試了以 wav2vec (Baevski et al., 2020)為基礎，經由 128 種不同語言的語料預訓練出的 XLS-R (Babu et al., 2021)，以及使用 MFCC 作為訓練目標並加入遮罩機制的 Hubert (Hsu et al., 2021)作為上游模型。儘管客家語擁有較強的語言獨立性，但我們仍然想了解相近語系的語料對客家語辨識效果的提升程度，為此我們試用了以中文為預訓練資料的 Chinese HuBert[2]。最後我們進一步嘗試結合基於 HuBERT 改良的 WavLM (Chen et al., 2022)。WaveLM 利用門控相對位置偏置 (gated relative position bias)並將相異的語句作為噪音混入訓練資料，促使模型可以在考量幀與幀之間位置關係的同時，增強自監督特徵的穩健性。

2.3 語言模型

由於純文字資料比標記的語音語料更容易取得，語言模型可以使語音辨識模型快速適配至跨域的測試場景中。淺層融合(Shallow Fusion) (Hannun et al., 2014; Gulcehre et al., 2015; Chorowski et al., 2016; Kannan et al., 2018)是一種直覺而有效的方法。Shallow Fusion 在解碼過程中，會在語音辨識模型和語言模型預測

[1] http://www.openslr.org/28

[2]https://github.com/TencentGameMate/chinese_speech_pretrain

的假設分數之間進行對數線性插值，彙整後即可同時利用兩個模型的知識，進行輸出令牌的預測。

客委會釋出的語料集中主要包含了台灣常見的六種客家語腔調。為了判別這些腔調之間的影響，我們為四縣腔獨立訓練一個語言模型(Four_LM)，將其與全部文本作為訓練資料的語言模型(Full_LM)進行比較。

語言模型除了可以與語音辨識模型進行淺層融合外，也可以被運用在語音辨識的前的前 N 個最佳假設重新排序上。追隨 PBERT (Chiu et al., 2021) – 基於 BERT (Devlin et al., 2019)演進的模組，用來對語音辨識的前 N 個最佳假設重新排序。PBERT 利用 BERT 與簡單的全連結前饋網路(Feed Forward Network, FFN)來挑選前N個最佳假設中，擁有最低字符錯誤率的黃金假設(oracle hypothesis)。

2.4 語音活性檢測

語音活性檢測的目的是偵測出音訊訊號中所包含的語音片段，主要包含了以下步驟：

1. 對於幀級別(frame level)的輸入各自計算其包含語音的後驗機率。
2. 應用閾值對後驗機率進行篩選，確立出包含語音的候選片段。
3. 在合併位置上相近的候選片段後，移除過短的語音片段。

我們選用 speechbrain 中所包含的一個以 LibriParty[3]訓練於 CRDNN (Ullah et al., 2022)架構的 vad-crdnn-libriparty[4]。

在對模型輸入資料進行前處理時，我們會預先偵測音檔的時長，將大於 30 秒的長音檔輸入至語音活性檢測模型，並在最後合併模型所偵測出包含語音的數個片段，作為語音辨識模型的輸入。

2.5 大型語音辨識模型

自監督模型透過固定其自身模型參數，或在微調後與下游任務進行整合。由於在訓練過程中存在不一致性，下游模型將難以充分發揮自監督模型的所有性能優勢。鑒於測試語料可能包含了日常生活中的常見噪音，我們比較了由 OpenAI 所開發的大型語音辨識模型 Whisper (Radford et al., 2023)。Whisper 模型經由 680,000 小時的語料訓練，並且其輸入為具有連續性的時頻譜(Spectrogram)，大幅提升了辨識過程中的穩健性。

先前的研究發現 (Li et al., 2018; Aghajanyan et al., 2020)，過度參數化(over-parametrized)的模型在學習時，權重的變化主要存在於較低的本徵維度 (Intrinsic dimension)內。因此，後來的學者們認為，在將模型微調以套用到新的特定領域時，也不外乎於本徵維度中調整參數，從而提出了 Low-Rank Adaptation (LoRA)方法(Hu et al., 2021)。有別於原始的模型骨幹，LoRA 額外新增了包含兩個旁支矩陣的分支，用於學習原有模型適應新領域所需的參數變化。由於凍結了原有模型的參數，LoRA 在不減損效能的同時還能顯著的降低模型訓練時的記憶體需求。

3 實驗結果與競賽策略

3.1 實驗設定

本次 2023 福爾摩沙客家語語音辨識競賽總共釋出了兩組語料，分別是名為 Lavalier 的初期訓練語料集以及名為 XYH-8-X 的熱身賽測試集(Pilot-Test)。我們根據賽程分別設定了兩組實驗，第一組實驗目的在於探索實驗的框架，第二組則用於評估決賽的提交模型。

Lavalier 與 XYH-8-X 語料庫皆為四縣腔的語料，總共由 87 位語者錄製而成，每位語者的錄音皆在半小時至一小時之間，總時長約 70 小時。統計資料如**表 1**.所示。

表 1. FSR-2023-Hakka 的統計資料

	Spks	Sent	Chars	Hrs
Lavalier	76	20,613	348,488	59.49
XYH-8-X	11	3,598	50,504	10.02

第一組實驗遵循官方的基線設定來分割 Lavalier 語料集，分別挑選兩組 4 男 4 女的語料作為驗證集(Init-Dev)和測試集(Init-Test)，其餘全部用於訓練(Init-Train)。詳細的切分資訊如**表 2.**：

[3]https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriParty/generate_dataset

[4] https://huggingface.co/speechbrain/vad-crdnn-libriparty

表 2. 初始資料集實驗設定(Initial Sets)

	Spks	Sent	Chars	Hrs
Init-Train	60	16,299	274,750	47.45
Init-Dev	8	2,126	36,265	6.16
Init-Test	8	2,187	37,473	5.88

第二組實驗設定則是將 Init-Train 和 Init-Dev 合併為新的訓練集(Fin-Train)，並將 Init-Test 視為本次實驗的驗證集(Fin-Dev)。XYH-8-X 則作為測試集(Fin-Test)參與評估。詳細的切分資訊如**表 3.**所示：

表 3. 最終資料集實驗設定(Final Sets)

	Spks	Sent	Chars	Hrs
Fin-Train	68	18,425	311,015	53.61
Fin-Dev	8	2,187	37,473	5.88
Fin-Test	11	3,598	50,504	10.02

3.2 實驗結果

如**表 4.**所示，我們首先比較了幾種基礎模型的辨識效能。透過大量語料預訓練過後，Whisper (Medium)在各類語言的領域知識潛移默化之下，取得了領先的辨識效果。在將 Branchformer (Peng et al., 2022)應用於客語的語音辨識時，其表現似乎不甚理想。由於 Whisper 限定以時頻譜作為輸入，因此在隨後的一系列以自監督模型作為前端編碼器的實驗中，僅包含了 Conformer (Gulati et al., 2020) 以及 E-Branchformer (Kim et al. 2022)。

表 4. 各類基礎模型之字元錯誤率(CER)比較

	Model	CER (%)
[A]	Conformer	4.11
[B]	Branchformer	4.63
[C]	E-Branchformer	4.07
[D]	Whisper (Medium)	**2.96**

實驗過程中採用的各類自監督模型詳細資訊紀錄在**表 5.**中。從**表 6.**可以觀察出，各種自監督特徵模型作為前端編碼器的實驗中，以 XLS-R 的效果最為突出，我們認為這是在各種語言的交互增益之下產生的結果，其概念類似於利用多語種語料預訓練的 Whisper。在[A-1]和[C-1]的對比中，可以觀察到 Conformer 在利用自監督特徵時的效率優於 E-Branchformer。此外在僅憑藉中文作為預訓練資料的情況下，Chinese Hubert 對比 Hubert 有較佳的結果表現，這顯示了客家語仍然可以從相近的語言中獲得一些對效能有益的訊息。由於 WavLM 改良了 Hubert 的穩健性，在同為英語預訓練的場景下，WavLM 的辨識性能優於 Hubert。

表 5. 自監督模型資訊

	ckpts
Hubert	hubert_large_ll60k
WavLM	wavlm_large
Chinese Hubert	chinese-hubert-large
XLS-R	xlsr2_960m_1000k (s3prl)

表 6. 各類自監督特徵之字元錯誤率(CER)比較

	Model	CER (%) Init-Test
[A]	Conformer	4.11
[A-1]	+ XLS-R	**1.95**
[A-2]	+ Hubert	2.15
[A-3]	+ WavLM	2.06
[A-4]	+ Chinese Hubert	2.00
[C]	E-Branchformer	4.07
[C-1]	+ XLS-R	2.12

在語言模型對比的實驗中，我們可以觀察到，儘管不過濾腔調的訓練文本的字數是四縣腔的兩倍，但在效能上仍以 Four_LM 模型較優。這顯示了不同腔調的客家語可能因地緣關係而自行演變出各自的獨特用語。這些相異的特徵在辨識四縣腔語料時，可能導致語言模型在語義理解時出現混淆，進而導致效能輕微下降。因此我們將以 Four_LM 作為後續實驗套用的語言模型。

表 7. 結合語言模型之字元錯誤率(CER)比較

Models	LM	LM Chars	CER (%) Init-Test
[A-1]	-	-	1.95
[E]	Four_LM	580k	**1.69**
[F]	Full_LM	1225k	1.77

在前*N*個最佳假設重新排序的實驗中，我們凍結預訓練語言模型 BERT(ckiplab 所釋出的 bert-base-chinese)，並透過全連結前饋網路來挑選最佳假設。如**表 8.**所示，可以看到使用 PBERT 的重新排序模型在客語語料上沒有顯著性效果，推估是因為預訓練語言模型沒有看過客語語料。由於我們也沒有足夠的訓練資料可以重新訓練 BERT，因此我們為了避免不確定性因素，在後續的實驗決定不採用重新排序模組。

表 8. 語音辨識假設重新排序成效

Models	CER (%)
	Init-Test
[B]	4.63
Oracle	3.91
[B-1]	4.46

為了測試目前模型在遭遇環境雜訊時的辨識能力，我們任意選擇兩位 Fin-Test 中的語者語料，並從中隨機挑選了部分音頻來疊加雜訊，以此構成了 Fin-Test-sub 測試集。從**表 9.**中可以明顯看出模型在 Fin-Test-sub 上的辨識表現大幅下降，其中括號內的數據為官方公告中熱身賽的錯誤率。

表 9. 分析最佳實驗配置模型的抗雜訊能力

Models	LM	CER (%)	
		Fin-Test	Fin-Test-sub
[A-1]	-	8.18 (8.16)	36.52
[E]	Four_LM	7.60	31.88

隨後為了進一步分析模型無法抵抗的雜訊類型，我們預先從訓練集挑選三句長度不一的話語並對其添加不同雜訊(模擬雜踏式噪音/模擬背景音樂/模擬背景噪音/模擬混響)。最後為無添加雜訊的五種情況都進行不同程度的速度擾動(0.8/1.0/1.2)。

表 10.中的實驗表現與我們的假想相符。由於訓練資料的收音場所為錄音室，因此混響類型的雜訊為[A-1]模型的已知情況。同時 XLS-R 所輸出的自監督學習特徵在面臨背景噪音時，仍保有一定程度的穩健性。但面對包含雜踏式噪音與背景音樂的測試資料時，[A-1]模型的辨識效能卻明顯下降。

表 10. 分析模型抵抗不同雜訊的能力

	CER (%)
原始	0.0
雜踏式噪音	26.67
背景音樂	4.76
背景噪音	0.0
混響	0.0

根據以上觀察，我們在最終的競賽階段特別針對雜踏式噪音與背景音樂等噪聲模式作資料擴增。生成的所有的副本都將利用速度擾動和頻譜擴增(Park et al., 2019)以構成 Fin-Train。我們運用 Fin-Train 訓練了兩種模型，一種是使用 LoRA 架構的適配器對 Whisper（Large）模型進行微調。另一種則是基於[E]使用 XLS-R 作為前端特徵的 Conformer 架構訓練而成的[H]。從**表 11.**可以發現[H]在 Fin-Test 上的效果對比[E]出現輕微減損。我們認為這是因為包含噪音的訓練資料導致模型在學習時的複雜度增加。儘管如此，[H]在 Fin-Test-sub 子集上的表現相比[E]相對減少了 73%。這表明了我們針對噪音的訓練方法，能夠有效的泛化模型的抗噪效能。

表 11. 最終競賽階段所採用的模型

	Model	CER (%)	
		Fin-Test	Fin-Test-sub
[G]	Whisper (Large)	12.59	13.57
[H]	Conformer + XLS-R	7.99	8.60

4. 提交模型與成果

我們從**表 6.**的實驗中可以得知，Conformer 架構結合作為前端的自監督模型 XLS-R 的[A-1]在最初的官方測試集 Init-Test 上，取得了字符錯誤率(CER) 1.95%的最佳成果。對比 Whisper (Medium)模型[D]，[A-1]仍然相對進步了 34.1%。因此我們以[A-1]做為熱身賽的提交模型，並取得了學生組第二名的佳績(參見**圖 2.**)。

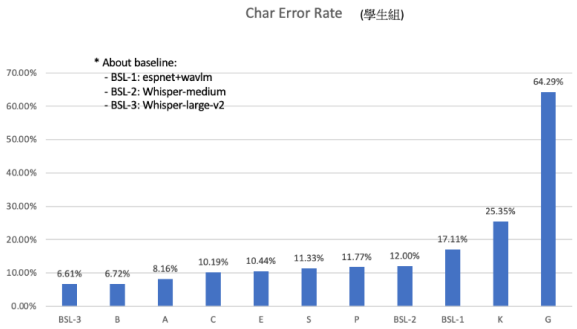

圖 2. 官方公告熱身賽隊伍競賽結果。本隊所對應的是學生組 A 組。

在最終競賽時我們沿用了[A-1]的模型設定，隨後融合了以客委會提供的四縣腔文本訓練而成的(Four_LM)，以此作為主要的模型架構。在經由各類方法擴增並且透過一系列實驗驗證實能夠強化模型對於噪聲抗性的 Fin-Train 訓練後，[H]在 Fin-Test 及 Fin-Test-sub 的表現相比 Whisper (Large) [G]，皆至少相對進步了 35.1%。我們將此模型作為本競賽的最終提交成果，並於最終取得了第四名的殊榮。

5. 參考資料

林正慧。2013。華南客家形塑歷程之探究。

張學謙、蘇鳳蘭、劉彩秀。2020。運用聯合國教科文組織（UNESCO）語言活力指標評估臺灣客語活力之研究期末報告。

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. *arXiv preprint arXiv:2012.13255.*

Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik, Mittul Singh and Mikko Kurimo. 2022. Automatic rating of spontaneous speech for low-resource languages. In *SLT*.

Sina Alisamir, Fabien Ringeval and Francois Portet. 2022. Cross-domain voice activity detection with self-supervised representations. *arXiv preprint arXiv:2209.11061*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Steffen Schneider and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed and Michael Auli. 2020. wav2vec 2.0: A Framework for self-supervised learning of speech representations. In *NeurIPS 2020*.

Kaushal S. Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar and Mitesh M. Khapra. 2023. Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages. In *ICASSP*.

Dongpeng Chen and Brian K.-W. Mak. 2015. Multitask learning of deep neural networks for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Nancy F. Chen and Haizhou Li. 2016. Computer-assisted pronunciation training: from pronunciation scoring towards spoken language learning. In *APSIPA*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, Shinji Watanabe. 2023. Improving Massively Multilingual ASR With Auxiliary CTC Objectives. In *ICASSP*.

Shih-Hsuan Chiu and Berlin Chen. 2021. Innovative bert-based reranking language models for speech recognition. In *SLT*.

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Ekapol Chuangsuwanich. 2016. Multilingual techniques for low resource automatic speech recognition. *Massachusetts Institute of Technology Cambridge United States*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. In *INTERSPEECH*.

Yu-An Chung, Hao Tang and James Glass. 2020. Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*.

Siddharth Dalmia, Ramon Sanabria, Florian Metze and Alan W. Black. 2018. Sequence-based multi-lingual low resource speech recognition. In *ICASSP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *NAACL*.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. 2023. FunASR: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo, Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. *arXiv preprint arXiv:2005.08100*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685.*

A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen and R. Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *ICASSP*.

Yassine E. Kheir, Shammur Absar Chowdhury and Ahmed Ali. 2023. Multi-View multi-task representation learning for mispronunciation detection. In *SLaTE*.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, Shinji Watanabe. 2022. E-Branchformer: Branchformer with Enhanced Merging for Speech Recognition. In *SLT*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *ICASSP*.

D. N. Krishna. 2021. Multilingual speech recognition for low-resource Indian languages using multi-task conformer. *arXiv preprint arXiv:2109.03969*.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the Intrinsic Dimension of Objective Landscapes. *arXiv preprint arXiv:1804.08838*.

Andy T. Liu, Shu-Wen Yang, Po-Han Chi, Po-Chun Hsu and Hung-Yi Lee. 2020a. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP*.

Andy T. Liu, Shang-Wen Li and Hung-yi Lee. 2021. TERA: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Alexander H. Liu, Yu-An Chung and James Glass. 2020b. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406.*

Yajie Miao, Florian Metze and Shourabh Rawat. 2013. Deep maxout networks for low-resource speech recognition. In *ASRU*.

Markus Müller, Sebastian Stüker, and Alex Waibel. 2016. Towards improving low-resource speech recognition using articulatory and language features. In *IWSLT*.

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.*

Yifan Peng, Siddharth Dalmia, Ian Lane, Shinji Watanabe. 2022. Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding. In *ICML*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Luka´s Burget, Ond˘rej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motl´ıcek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *ASRU*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *PMLR*.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré and Emmanuel Dupoux. 2020.

Unsupervised pretraining transfers well across languages. In *ICASSP*.

Steffen Schneider, Alexei Baevski, Ronan Collobert and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862.*

Abhayjeet Singh, Arjun Singh Mehta, Ashish Khuraishi K S, Deekshitha G, Gauri Date, Jai Nanavati, Jesuraja Bandekar, Karnalius Basumatary, Karthika P, Sandhya Badiger, et al. 2023. Model adaptation for ASR in low-resource Indian languages. *arXiv preprint arXiv:2307.07948.*

David Snyder, Guoguo Chen and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484.*

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey and Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *ICASSP*.

Vrunda N. Sukhadia and S. Umesh. 2022. Domain adaptation of low-resource target-domain models using well-trained ASR conformer models. In *SLT*.

Samuel Thomas, Michael L. Seltzer, Kenneth Church and Hynek Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *ICASSP*.

Rizwan Ullah, Lunchakorn Wuttisittikulkij, Sushank Chaudhary, Amir Parnianifard, Shashi Shah, Muhammad Ibrar and Fazal-E Wahab. 2022. End-to-End Deep Convolutional Recurrent Models for Noise Robust Waveform Speech Enhancement. *Sensors*.

Hsin-Wei Wang, Bi-Cheng Yan, Hsuan-Sheng Chiu, Yung-Chang Hsu and Berlin Chen. 2022. Exploring non-autoregressive end-to-end neural modeling for English mispronunciation detection and diagnosis. In *ICASSP*.

Haihua Xu, Van Hai Do, Xiong Xiao and Eng Siong Chng. 2015. A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition. In *INTERSPEECH*.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP*.

Huayun Zhang, Ke Shi and Nancy F. Chen. 2021. Multilingual speech evaluation: Case studies on english, malay and tamil. In *INTERSPEECH*.

Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. In *NeurIPS SAS.*

Jing Zhao and Wei-Qiang Zhang. 2022. Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing.*

Xiaohuan Zhou, Jiaming Wang, Zeyu Cui, Shiliang Zhang, Zhijie Yan, Jingren Zhou and Chang Zhou. 2023. Mmspeech: Multi-modal multi-task encoder-decoder pre-training for speech recognition. *arXiv preprint arXiv:2212.00500.*