Application of Fine-Tuning XLSR-53 Acoustic Model for Speech Learning in Atayal

XLSR-53 聲學模型微調在泰雅語語音學習之應用

專題領域：系統領域　組別： B492　指導教授：劉奕汶　組員：林正硯、林宜蓁、許修睿

摘要

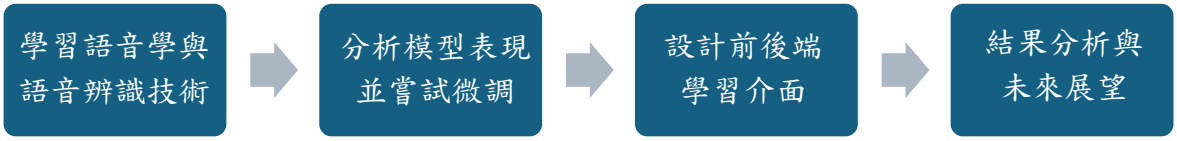
本研究介紹一套針對臺灣泰雅語的電腦輔助發音訓練系統(CAPT)。該系統利用語音辨識技術，自動偵測學習者的發音錯誤，促進發音學習。泰雅語是台灣獨特的原住民語之一，具有重要的文化與語言價值。本研究針利用少數語言的語音辨識系統建模技術，實作了一個電腦輔助發音訓練系統。該系統主要分為兩部分：一是診斷學習者的音韻表現，二是檢測語調變化。在音韻指標方面，本研究採用開源跨語種聲學模型 XLSR-53，對學習者的語音進行音位辨識。接著，透過序列比對演算法，將辨識出的音位序列與正確音位序列進行比較，以實作音韻指標。在語調指標方面，首先利用基音估測演算法，對學習者和教師的音檔進行基音抽取，計算基音差分序列。然後，通過動態時間規整技術，對齊兩差分基音序列，計算均方根誤差，以評估學習者的語調。。

除此之外，本研究開發了一個使用者學習介面網站並在電腦本機上運作。在網頁介面中，音韻指標方面，網站中會列出正確的 IPA 音標序列以及 XLSR-53 聲學模型偵測到學員音檔的音標序列；在語調指標方面，本研究則是將音檔視覺化，透過音檔的梅爾頻譜(Mel spectrogram)並在頻譜上標示出估計出來的語調追蹤線，以展示出學員及教師句中的語調起伏。

一、研究背景與動機

本題目源自於清大幼教系辛靜婷教授提出之「應用 AI 語音辨識於醫療醫療族語與原住民族醫療知識之學習」，宗旨是希望開發一套針對泰雅語的自學發音系統，給予學士後醫學系同學使用，利用這套泰雅語發音學習系統，能夠使得本校學士後醫學系的同學，甚至是有興趣接觸泰雅語的人，能夠更有效率地學習發音。然而支持語音辨識技術的語言都以主流語言為主，泰雅語屬於瀕危語言，甚至沒有大型語料庫可供使用。因此由碩士班莊裕嵐學長改良 Sheoran 提出之 CAPT 架構、由 Li 提出使用泛用聲學模型對少數語言進行語音辨識建模等技術，成功實作出基於跨語言模型 XLSR-53 的泰雅語電腦輔助發音訓練系統，而本研究推進其結果，進一步嘗試微調模型表現，並結合實際教學需求，實作出完整的前後端介面。

二、研究流程圖



三、研究方法

（一）賽考利克泰雅語

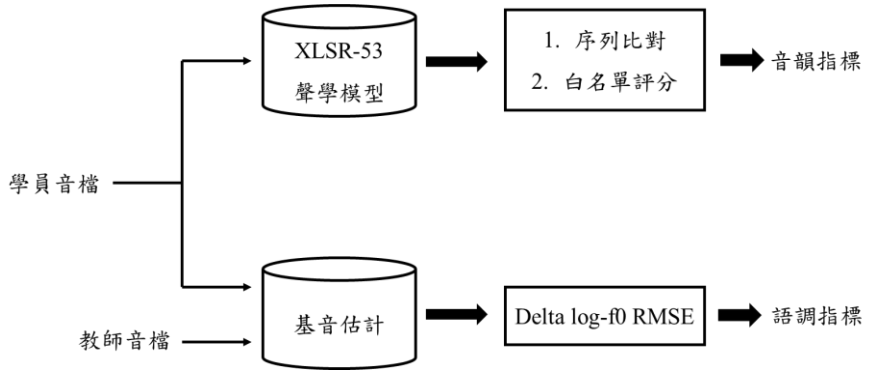
泰雅族有許多分支，其中又以賽考利克為分支最大，因此接下來我們將以賽考利克泰雅語做為本研究學習者主要的學習語言。泰雅語共有 19 個輔音和 5 個元音，如下表所示，雖然根據表格可以大致模擬泰雅語的發音，但在特殊情況下寫法與唸法仍有所差異。

發聲方式/部位		雙唇	齒齦	硬顎	軟顎	小舌	喉
塞音		p	t		k	q	‘ (ʔ)
塞擦音			c (ts)				
擦音	清音		s		x		h
	濁音	b (β)	z		g (ɣ)		
鼻音		m	n		ng (ŋ)		
邊音			l				
顫音			r				
滑音		w		y (j)			

舌位前後/ 高低	前	央	後
高	i		u
中	e		o
低		a	

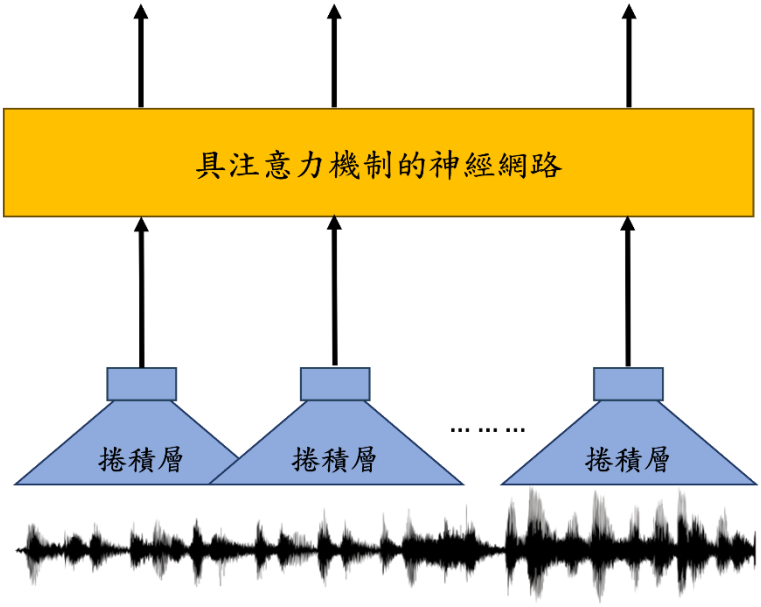
（二）系統架構

我們使用改良 Sheoran 之架構，主要包含音韻及語調兩個部分，音韻部分使用了跨語言模型 XLSR-53，亦是研究核心。語調部分則以教師音檔與學生音檔進行計算比對。



（三）XLSR-53 聲學模型

XLSR 的全名為 Cross-lingual Speech Recognition，該模型使用包含 53 種語言的語料庫作為訓練集，並將各語言的文字轉換為國際音標，用來辨識輸入音檔中所說的音。其包含抽取聲學意義的捲積神經網路 CNN，以及具有根據 Transformer 架構微調的神經網路兩個部分，並採用連結時序分類(CTC)作為模型的損失函數。根據 Meta 的原始論文，捲積層的部分已經訓練的足夠好，我們微調的對象為黃色區塊的神經網路，用來將發音映射到泰雅語的音標上。



（四）音韻指標

音韻的計算採用音標序列比對演算法，主要應用於生物基因序列比對。將辨識結果與正確序列進行比較配對，如下圖。以此根據配對結果計算分數，並根據將根據「白名單」機制進行後續給分，「白名單」基於族語老師教學經驗設計，主要記錄一些口語上可被容忍的發音組合。

```
Pronunciation Evaluation Summary
Reference : a ɣ - aɪ w a h n j - u x s a k u m ə - t a k u i l a
            |     | | |       |   | | | | |   | | | | | | | |
 Learner  : a - k aɪ w a - - - ɲ u - s a k u m - a t a k u i l a


Final Score : 78.26
=================================================================
Some Highlights
Vowel Pronunciation Errors : [('ə', 'a')]
Consonant Pronunciation Errors : [('ɣ', 'k'), ('n', 'ɲ')]
Pronunciation Missings : [('h', '-'), ('x', '-')]
```

（五）語調指標

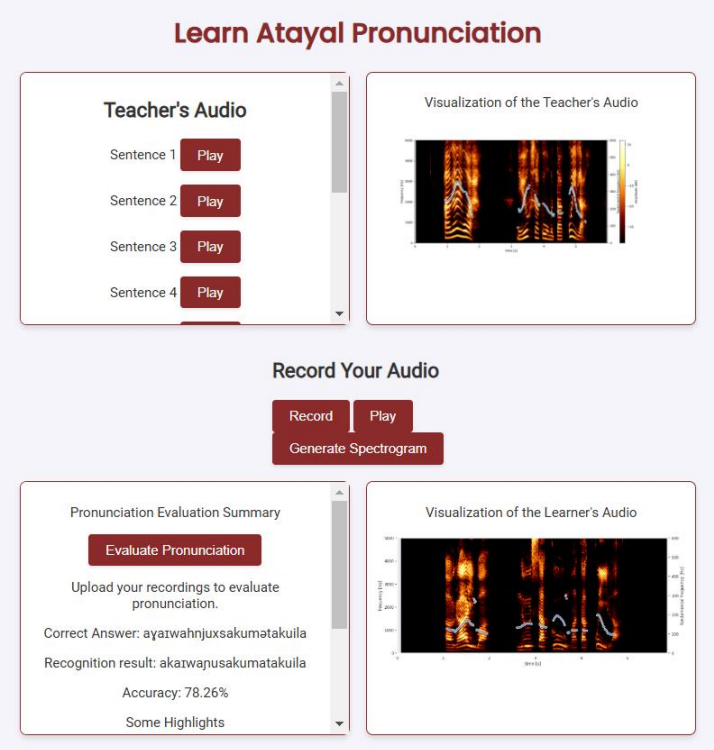
在語調指標的評估機制中，也使用基於 Sheoran 計算方式的改良方法，使用 delta log-f0 root-mean-square error (簡稱 delta log-f0 RMSE)計算語調指標，此方法著更重於時序上的差異，不隨著基音絕對數值而改變。我們使用 WORLD 聲碼器估計音檔之基音序列，取對數後以 f0 表示。抽取出基音序列之後，可以求得各自的時間差分序列，利用 DTW 演算法，對兩基音序列作對齊。核心概念是將兩序列對齊後，累計彼此間差異的距離，再代入公式得到相應的分數。

$$\Delta f_0(t) = f_0(t+1) - f_0(t) \qquad \Delta F_0 = [\Delta f_0(1) \dots \Delta f_0(k)]$$

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\Delta F'_0{}^{learn}(i) - \Delta F'_0{}^{instr}(i))^2}$$

（六）網頁設計

我們希望設計方便的介面讓學員使用，因此我們設計了網頁並盡可能視覺化語音學結果。參考 CoolEnglish 線上學習網站的功能，設計了學習音檔、錄音功能，以及發音診斷三個要素。能夠視覺化頻譜與語調估計、顯示發音辨識結果、提供分數與發音建議，並通過 local host 的方式呈現。



四、研究結果

（一）XLSR-53 準確率

我們使用常見的音錯誤率(phonetic error rate，或簡稱 PER) 來評估整體辨識率的好壞。計算方式如下所示，式中 N 為有所有的音，S 為被替換掉的音，I 為額外插入的音，而 D 為被刪除的音。在數值的判讀上，PER 數值越小，則辨識表現越佳。

$$PER = \frac{S + I + D}{N}$$

透過對整體泰雅語語料辨識，可以得到初始 PER 為 22.1%，此為尚未考慮同位異音(allophone)的情況所得出的結果。同位異音(allophone)指同一個音位(phoneme)可以表示多個音，以英文單字 pin[p^hm]與 spin[spɪn]舉例，[p^h]與[p]都代表音位/p/，因此為一對同位音，但是在中文中，[p^h]對應注音符號ㄆ，[p]對應注音符號ㄅ，因此在中文中並不是同位音，由此可知，同位音並不會改變原有詞意，且不同語言有各自的認定。經過族語老師篩選排除後，重新計算過後的 PER 為 9.6%。由此可見，目前 XLSR-53 聲學模型，套用在泰雅語語料上的表現並非完美。統計後發現錯誤源自兩大原因，一為泰雅語本身音韻構成比例，二為近音對應的元音:

1. 泰雅語的音韻構成

   以泰雅語中 g 舉例，字母 g 為[γ] ，但 XLSR-53 傾向於辨識為[g] 。原因是大多數語言並不存在[γ]，使得訓練後的 XLSR-53 更傾向於[g]，而[γ]則有第二高後驗機率。諮詢族語老師後，雖然同學與模型相同皆傾向發成[g]，但老師不認為兩者為同位音，仍納入 PER。

2. 進音對應元音

   在近音辨識上，可能會額外多辨識出近音所對應的元音。近音是介於元音和輔音之間的音，發音時長較元音短，口型也不明顯，如英文中的/w/。在泰雅語中，字母 w(/w/)、y(/j/)都是近音。輔音/w/對應元音為/u/，輔音/j/對應元音為/i/，所以模型容易在近音前後辨識出對應之元音。

錯誤類型	次數	百分比
/ㄚ/辨識成/g/	43	1.96%
近音多辨識出的對應元音	35	1.59%
未辨識出聲門塞音/h/	25	1.14%
單純辨識錯誤	102	4.66%

（二）XLSR-53 微調之結果

使用 21 位受訪者共 86 分鐘的訪談影片對神經網路進行微調。我們使用 Hugging face 配套之 Trainer 進行微調，並使用 WER(word error rate)做為損失函數，並以原有之語料庫做為測試資料，結果如下。

錯誤類型	次數(與微調前相比)	百分比
/ㄚ/辨識成/g/	37(-6)	1.68%
近音多辨識出的對應元音	42(+7)	1.91%
未辨識出聲門塞音/h/	25(0)	1.14%
單純辨識錯誤	99(-3)	4.51%

在多次調整參數後，發現模型並沒有較好的表現，推測是受訪者本身發音並不標準，聲音背景雜訊過多，轉錄稿與辭典有誤等，整體 PER 與微調前持平。

（三）網頁測試之結果

本研究之網頁在運作時，錄音的部分透過電腦本身的錄音功能進行錄音，由於錄音環境具有周遭的背景雜音影響，以及電腦內建的錄音設備並不優良，因此在圖 10 中右下角學員錄音檔之頻譜方面，語調估計線相較於右上角的學習音檔效果來說，較容易出現斷斷續續、語調估計明顯錯誤，或是偵測失敗的情況發生，但在句中抑揚頓挫較明顯的部分，仍可看出估計線的高低起伏。

五、結論

本研究完整實作了泰雅語發音學習之電腦輔助系統使用介面，並展示了前端操作介面、後端部署、微調模型在技術上的可行性。開發出能夠錯誤診斷功能、具擴充性之網頁，但仍有諸多不足與未來延伸研究空間。

1. 白名單機制

白名單的存在，本質上為語音辨識準確度與實務教學上的平衡，對於準確發音的定義以及外語學習者的標準存有曖昧空間，同時也彌補現階段模型的不足之處。

2. 模型準確度

截至目前為止，微調並沒有取得重大突破，整體 PER 維持在 9.6%，推測受制於高品質語料庫大小。通過諮詢辛靜婷教授得知，目前有蒐集約 30 小時的語料可供未來微調使用，為後續發展的首要任務。

3. 加強系統學習功能

現階段的系統僅能針對單一學生「診斷發音錯誤」，並沒有達成「訓練」的目標，僅作為學習系統之工程樣品，主因為沒有蒐集該學習者在其他題目的錯誤類型、其他學習者在此題目的錯誤類型，分析後得到個人化的改進建議。且成果尚未與專業族語老師的評分比對，亦未得到足夠的使用經驗回饋。後續將嘗試實現推薦系統，根據學習者們的錯誤類型推薦對應學習教材。

六、心得感想

林正硯

完成這次專題後，我對於少數民族語言的保存與發展有了更深的體悟。在教授、學長的帶領下，我學到了語音處理相關知識、泰雅語語音學、聲學模型的應用，特別是少數語言與應用情境下的適用性與挑戰。透過對音韻和語調指標的計算與比較，我們能夠更具體地了解學習者的發音問題，進而提供針對性的改進建議。也根據教師們的回饋開發出相關網頁及功能。感謝指導我們提供建議的劉奕汶教授，提供語料的辛靜婷教授與 Apang Bway、Sayun Yumin、Sugiy Tosi、Toyu Watan 四位族語老師、開發改良架構的莊裕嵐學長，讓我們能夠順利完成這項研究。未來，我希望能夠進一步改善系統，善用學習者資訊增加個性化學習功能，並透過更多的實地測試來提高模型的準確性與實用性。這次的研究經驗讓我深刻感受到技術在文化傳承中的力量，也激勵我繼續在這個領域中探索和貢獻。

林宜蓁

The research experience has been challenging yet rewarding over the past year. By attending the lab meetings, I have been exposed to a variety of projects currently going on in the lab; on top of that, I was fortunate to get personal teaching from a senior to build up the foundational knowledge for our work. In addition, regular discussion with the professor provided critical insights and suggestions that helped us overcome various obstacles and further refined the direction of the project.

During the second half of the year, my main task was to create a website to visualize and present the system we developed. This was a big challenge because it depended on my efforts whether the platform would be functional and user-friendly. Although at the very beginning I had little experience in web development, step by step I acquired the necessary knowledge to develop an intuitive and transparent interface.

The project taught me to do the work patiently and to be flexible with something that was out of the comfort zone. Other than web design, the concept of audio processing and machine learning, integrating all these into one system, was really different. When I reflect on this experience, it has been very self-satisfactory, and I wish our work will help the beginners learn the language and speak it with confidence in the near future.

許修睿

從今年二月開始參與實驗室參與會議，除了旁聽實驗室學長姐正在做的題目之外，協助專題的學長會另找時間教專題相關的背景知識，教授也針對我們遇到的問題提供建議以與延伸方向，讓我對這個領域有更多的了解。進入到下半學期完成系統後，我學到了完整的模型微調流程，並首次接觸網頁設計，讓系統能透過視覺化的方式呈現。這次的專題讓我在音訊處理這個領域、機器學習以及網頁設計有更進一步的收穫，也希望未來此系統能成功幫助初學者講出標準的泰雅語。最後，由衷感謝指導教授、學長以及我的組員，他們的幫助讓這個專題得以順利完成，也讓我收穫滿滿。