# 雙語語音辨識系統應用於 FSR-2020 競賽

# A Bilingual Speech Recognition System for FSR-2020 Challenge

洪孝宗　Hsiao-Tsung Hung
華碩電腦
雲端架構軟體中心
AlexHT_Hung@asus.com

吳孟哲　Meng-Che Wu
mcwu519@gmail.com

## 摘要

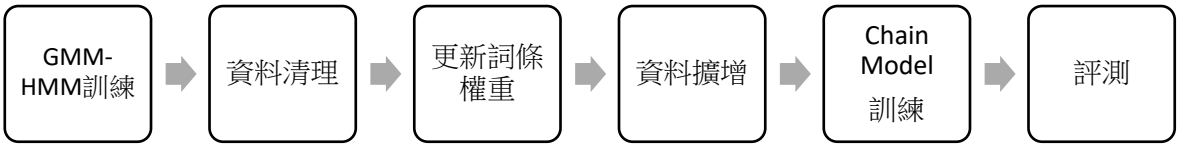
本文說明一個基於雙語語音辨識架構的系統應用於台語語音辨識競賽。為了有效利用台語及華語的文字和語音語料，我們採用福爾摩沙音標來建立聲學模型，並且合併兩個語言的辭典。在語言模型部分，我們分別針對各語言將文字語料進行斷詞處理，再線性組合得到雙語言模型。除了競賽提供的 TAT、PTS 及 FTV 語料，我們額外加入 Aishell-1 和 NER-Vol1 華語語料，包含訓練聲學模型及語言模型兩部分。最後評測結果在 Track-2 中得到約 46.8%字元錯誤率。從辨識結果可以看出辭典及文字語料都有待補充。

關鍵詞：Bilingual speech recognition，Taiwanese speech recognition，Formosa phonetic alphabet

## 一、緒論

本次競賽活動聚焦於台語語音辨識任務，希望能推動台語的自然語言技術發展，而比賽的語料也貼近現實生活中的華語、台語混合使用的情境。競賽包含了三個項目；第一是將台語語音轉寫並翻譯成華語，此問題可以看作是結合語音轉文字和文字翻譯兩個任務。第二及第三項目都是台語語音辨識任務，唯目標書寫系統不同，分別是台文漢字和台語羅馬拼音。本次活動提供了朗讀語料 TAT-Vol1，以及公視與民視節目的對話語料。民視語料屬於翻譯成華語的字幕，對應於第一個比賽項目。公視節目字幕採用教育部推廣的標準用字，可以對應到第二個項目。最後，TAT-Vol1 朗讀語料包含多種格式轉寫標記，可以用在所有情境。

從自然語言處理技術的角度來看，台語文發展的歷史造成許多問題，使得台語語音辨識有相當高的難度。我們認為最大的問題在於書寫系統不一致，導致台語文的文學藝術作品無法有效累積成為豐富的語料庫。其次，台語的詞彙數量並沒有如同主流語言一樣的成長，而此現象也會讓台語能夠使用的場合逐漸萎縮。




圖一、系統建置流程

近年的深度學習技術已經成功應用在多個領域，其中端到端的架構是被廣泛使用的設計之一。我們可以用此技術迴避台語辭典的疏漏以及書寫系統不一致的問題，甚至可以直接完成華語翻譯的任務，輸入台語語音時直接輸出華語翻譯結果。然而，我們認為此方法並不能有效改善前述的台與發展困境。例如，台語語音直接被翻譯成華語變成普遍的現象，也可以視為台語寫作的場域減少。

由於在台灣的語言使用環境還是以華語為大宗，而，所以我們選定第二個台文漢字轉寫項目來探究合適的語料處理方式。而台語羅馬字轉寫問題應較漢字簡單，且比賽提供的基線系統已經有相當低的錯誤率，故本文只聚焦在台文漢字轉寫任務。以下依序介紹辨識系統的各個元件及建置流程，相關程式碼已發布到網路[1]。

## 二、發音辭典

發音辭典是傳統語音處理技術的核心元件，其音標系統直接影響聲學模型的複雜程度，同時單詞本身也是語言模型的最小分析單位。由於我們希望能共享更多語言的語料，且真實對話語料也常混合多個語言使用，所以我們採用福爾摩沙音標（Formosa Phonetic Alphabet, ForPA）[1]作為聲學模型的基本單位，加上台語八聲調的調號。

實作上我們採用教育部《臺灣閩南語常用辭典》和《重編國語辭典修訂本》當作基礎，再透過查表轉換方式將台語羅馬拼音和華語注音轉換為 ForPA 格式。最後為了確保每個漢字都有對應發音，我們也蒐集了網路資源來補足所有單字發音。在資料處理階段，兩個詞典分開來根據語料語言來使用，包含文本段詞和 GMM-HMM 聲學模型訓練。

在建置詞典時我們遭遇些問題。首先，ForPA 在聲調方面並沒有統一的標記建議。我們根據調值的接近程度作判斷依據，將華語第二聲調對應至台語第五聲調，華語第三聲調對應台語第三聲調，華語第四聲調對應至台語第二聲調。其次，外來詞的發音標記和一般詞彙不同。由於至今仍有許多源自日語的詞彙常被使用，我們認為未來應該用不同標記方式或其他的架構來結合日語語料。第三，教育部詞典考量讀音涵蓋完整度而條列許多種發音變異，然而對於系統也會造成辨識時搜尋空間太大的問題。因此我們採用統計方式給予每個條目不同權重，降低冷僻發音方式出現的機率。

## 三、語料前處理及模型訓練

本系統採用 Kaldi[2]工具建立傳統 DNN-HMM 語音模型，模型建置流程如圖一所示。首先，我們發現 TAT-Vol1 語料有相當多的靜音片段，而且公視語料是節目字幕，

[1] https://github.com/alex-ht/TeamAlex-FSR2020。

表一、語料基本資訊

| 訓練集 | 語言 | 類型 | 時數(小時) |
|---|---|---|---|
| TAT-Vol1 | 台語 | 朗讀 | 41.8(24.5) |
| PTS | 台語、華語 | 對話 | 72.6 |
| NER-Vol1 | 華語 | 對話 | 126.7 |
| Aishell-1 | 華語 | 朗讀 | 179.0 |
| 發展集 | | | |
| Pilot-test | 台語 | 朗讀 | 4.8 |
| PTS | 台語、華語 | 對話 | 2.2 |

亦可能存在字幕和口說不匹配的情況。因此，我們採用常見的 SAT-GMM-HMM[2]模型建置流程，再使用 Kaldi 提供的工具進行資料清理及重新切斷長句等處理。

訓練語料的統計資訊如表一所示。除了本次競賽提供的語料，我們額外加入 NER-Vol1[3]及 Aishell-1[4]語料一起訓練聲學模型，而語料的轉寫文字則用來建立語言模型。其中 Pilot-test 語料相當接近 TAT-Vol1 訓練集，同樣屬於朗讀語料。為了測試對話語料的表現，我們將 PTS 語料切分 5%作為發展集。TAT-Vol1 經過清理後剩下 59%時間長度的音檔片段，約剩 24.5 小時，其中多數都是刪除靜音部分。此外，由於我們使用的辭典只收錄漢字詞彙，而 TAT-Vol1 的台文標記含有相當多的台羅拼音字，會造成這些字被標記為未知詞，有可能會被刪除，為此我們用人工查詢教育部字典的方式來校正。

在語言模型方面，我們使用 SRILM[5]對四個訓練語料分別建立三連詞語言模型，再依據發展集的 PTS 語料調整線性組合的權重，合併成一個模型。由於台語、華語皆使用漢字表示，所以兩個語言中有相同漢字的詞彙可以被共享 N 連詞機率。

最後我們採用 Chain Model[6]作為聲學模型，大致上和經典的流程一樣，只有做些許變動。首先，訓練集資料會先經過速度及音量擴增至三倍資料，再用 MUSAN[7]語料對原始資料混響，最後總共擴增至原始資料的六倍。完整的類神經網路架構如圖二，輸入為 64 維 FBank 以及 100 維 I-vector 特徵。我們參考 Kaldi 提供的範例，在 FBank 特徵後接續 SpecAugment[8]層來擴增資料。I-vector 特徵會經過一次仿射變換（Affine transform），再將其視為 Feature map 並與 FBank 合併，最後一起輸入到 CNN 層。由於我們的 FBank 特徵維度較高，所以 CNN 的層數由常用的 2 層增加到 3 層。


圖二、類神經網路架構

## 四、實驗結果

實驗的測試語料包含 Pilot-test 及隨機切分的部分 PTS 語料，結果如表二所示。首

表二、各模型在 Pilot-test 及 PTS 語料評估的字錯誤率(CER%)

| Model | Pilot-test | PTS |
|---|---|---|
| SAT-GMM-HMM-SI | 50.63 | 75.29 |
| SAT-GMM-HMM | 66.79 | 74.94 |
| SAT-BASIS-GMM-HMM-SI | 54.22 | 72.09 |
| SAT-BASIS-GMM-HMM | 51.94 | 71.74 |
| +Lexicon reweighting | 49.97 | 70.61 |
| Chain Model (CNN+TDNN-F) | 19.49 | 43.39 |

先我們發現經過同樣是經過 SAT 訓練的 GMM-HMM 模型，在第一階段語者獨立的方式辨識較第二階段經過 fMLLR 轉換還要好，如同前兩列的 SAT-GMM-HMM-SI 及 SAT-GMM-HMM，而在 PTS 語料中則沒有此現象。其原因可能是其是 Pilot-test 有語者標記，而 PTS 沒有，混和一起訓練所造成此問題。我們嘗試再用 Basis fMLLR[9]訓練更大的模型，來降低語者標記不足的問題。結果顯示經過 Basis fMLLR 轉換在 PTS 語料會較好，而 Pilot-test 部分雖稍稍較差，但 Basis fMLLR 的結果較合理。接著繼續做辭典更新詞條的權重，測試結果也有些許改善。最後在 Chain Model 部分，Pilot-test 和 PTS 可以降低到 19.49%及 43.39%的字錯誤率。最後，我們繳交兩份結果，分別是用一般方式根據 PTS 語料的最佳參數指定聲學模型和語言模型結合的權重，第二份是利用 Minimum Bayes-Risk (MBR)解碼產生答案。比賽結果分別得到 47.4%及 46.8%的字錯誤率。

經過初步分析，我們的系統存在以下幾點問題。首先，最嚴重的問題是我們的辭典沒有處理好。我們的字典缺少台羅拼音字，即使加入到辭典也沒有對語料，導致語言模型中的外來詞出現機率不高。這使得許多常見的用詞都無法正確顯示，例如 ootóobái(腳踏車)，系統仍會輸出”喔都穩”或其他相似發音的字。第二的問題是教育部字典仍有許多字無法涵蓋，例如” 糖醋排骨”被辨識為”當初排骨”。在教育部國語字典中有收錄”糖醋”一詞，但在閩南語常用辭典中沒有收錄，在 N 連詞語言模型的計算過程就會退到兩個單字詞”糖”、”醋”連乘機的情況，而輸給高頻詞”當初”。如何有效共享兩個同是漢字系統的語言模型也是一個待研究的題目，在音節層次建立較複雜的模型或許是可以嘗試的方向，例如遞迴結構的類神經網路語言模型普遍認為可以保留較長距離資訊。

## 五、結論

我們嘗試使用福爾摩沙音標系統讓華語和台語可以共享語音語料，而藉由教育部頒布的建議用字來共享兩種語言的文字語料。從實驗結果可以看出本系統架構仍有許多改善空間，其中台語辭典的完整度和台語文字語料的數量仍然遠遠不足。然而，我們希望未來繼續努力補齊字典，達到和同華語一樣的完備程度，並且盡速統一書寫方式，才能達到累積語料的目的。另外，我們發現外來詞可能在建立辭典過程變成特例處理，部份詞有日文漢字，而教育部辭典並沒有定義標準漢字，只有台語羅馬拼音。最後，我們希望發展技術的同時也能促使語言成長。雖然傳統的辨識架構需要較多的人工介入，但過程產生的詞典和語料都會是可以被進一步研究或發展成新應用。

# 參考文獻


[1] R.-Y. Lyu, M.-S. Liang, and Y.-C. Chiang, “Toward Constructing a Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin,” *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 9, no. 2, pp. 1–12, 2004

[2] D. Povey *et al.*, “The Kaldi Speech Recognition Toolkit,” in *ASRU*, 2011.

[3] Y.-F. Liao, Y.-H. S. Chang, Y.-C. Lin, W.-H. Hsu, M. Pleva, and J. Juhar, “Formosa Speech in the Wild Corpus for Improving Taiwanese Mandarin Speech-Enabled Human-Computer Interaction,” *Journal of Signal Processing Systems*, vol. 92, no. 8, pp. 853–873, 2020

[4] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline,” in *Oriental-COCOSDA*, 2017

[5] A. Stolcke, “SRILM-An Extensible Language Modeling Toolkit,” in *7th International Conference on Spoken Language Porcessing*, 2002, pp. 901–904

[6] D. Povey *et al.*, “Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI,” in *INTERSPEECH*, 2016.

[7] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus.” 2015, arXiv:1510.08484v1

[8] D. S. Park *et al.*, “SpecAugment: A Simple Augmentation Method for Automatic Speech Recognition,” in *INTERSPEECH*, 2019.

[9] D. Povey and K. Yao, “A basis representation of constrained MLLR transforms for robust adaptation,” *Computer Speech & Language*, vol. 26, no. 1, pp. 35–51, 2012