Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 89-138
© The Association for Computational Linguistics and Chinese Language Processing

結合語音辨認及合成模組之台語語音轉換系統

Taiwanese Voice Conversion based on Cascade ASR and TTS Framework

許文漢 *、廖元甫#、王文俊 +、潘振銘+

Wen-Han Hsu, Yuan-Fu Liao, Wern-Jun Wang, and Chen-Ming Pan

摘要

台語已被聯合國列為瀕危消失語言，急需傳承。因此，本論文研究如何做出一個可以用任何人的聲音，合成出任何台語語句的台語語音合成系統。為達到此目的，我們首先(1)建置一 Taiwanese Across Taiwan (TAT) 大規模台文語音語料庫，其共有 204 位語者，約 140 小時的語料，其中有兩男兩女，每人約 10 小時的台語語音合成專用語料。然後(2)基於 Tacotron2 之語音合成架構，並加上前端中文字轉台羅拼音模組與後端 WaveGlow 即時語音生成器，建立中文文字轉台語語音合成系統。最後(3)基於串接台語語音辨認與語音合成架構，建置一台語語音轉換系統，並完成同語言：台語對台語語音轉換；以及跨語言：華語對台語語音轉換，兩種台語語音轉換功能。為評估此台語語音轉換系統的成效，我們透過網路公開招募到 29 位實驗者，進行同語言及跨語言轉換台語語音兩項評分任務，並分別進行針對「自然度」與「相似度」的 MOS 分數之主觀評測。實驗結果顯示，在同語言部分，若使用目標語者 10 分鐘，3 分鐘與 30 秒語料進行測試，自然度平均 MOS 分數依序為 3.45 分，3.02 分與 2.23 分，相似度平均 MOS 分數依序為 3.38 分，2.99 分與 2.10 分；而在跨語言部分，若使用目標語者 6 分鐘與 3 分鐘語料進行測試，自然度平均 MOS 分數依序為

* 國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
E-mail: jeff3136169@gmail.com
* 國立陽明交通大學智能系統系
Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University
E-mail: yfliao@nycu.edu.tw
+ 中華電信研究院
Chunghwa Telecom Laboratories
E-mail: {wernjun, chenming }@cht.com.tw

2.90 分與 2.70 分，相似度平均 MOS 分數依序為 2.84 分與 2.54 分。由實驗結果，可以顯示我們確實初步達成一個可以用任何人的聲音，合成出任何台語語句的台語語音合成系統。

Abstract

Taiwanese has been listed as an endangered language by the United Nations and is urgent for passing on. Therefore, this study wants to find out how to make a Taiwanese speech synthesis system that can synthesize any Taiwanese sentences via anyone's voice. To achieve this goal, we first (1) built a large-scale Taiwanese Across Taiwan (TAT) corpus, with in total of 204 speakers and about 140 hours of speech. Among those speakers, two men and women, each one has especially about 10 hours of speech recorded for the purpose of speech synthesis, then (2) establish a Chinese Text-to-Taiwanese speech synthesis system based on the Tacotron2 speech synthesis architecture, plus with a frontend sequence-to-sequence-based Chinese characters to Taiwan Minnanyu Luomazi Pinyin (shortened as Tâi-lô) machine translation module and the backend WaveGlow real-time speech generator, and finally, (3) constructed a Taiwanese voice conversion system based on the concatenated speech recognition and speech synthesis framework where two voice conversion functions had been implemented including (1) same-language: Taiwanese to Taiwanese voice conversion, and (2) multi-language: Chinese to Taiwanese voice conversion. In order to evaluate the Taiwanese voice conversion system, we publically recruited 29 subjects from the Internet to conduct two kinds of scoring task: same-language and cross-language voice conversion and carried out the subjective "naturalness" and "similarity" mean opinion score (MOS) evaluations respectively. The test result shows that in the Intra-lingual session, the average naturalness MOS is 3.45, 3.02 and 2.23 points, and average similarity MOS score's 3.38, 2.99 and 2.10 points while using 10 minutes, 3 minutes, and 30 seconds target speech, respectively; in cross-lingual part, the average naturalness MOS score is 2.90 and 2.70 points; average similarity MOS score is 2.84 and 2.54 points while using 6 minutes and 3 minutes target speech, respectively. From those results, it shows that our proposed system indeed could synthesize any Taiwanese sentences via anyone's voice.

關鍵詞：台文語音語料庫、台語語音合成、台語語音轉換

Keywords: Taiwanese Across Taiwan, Taiwanese Speech Synthesis, Taiwanese Voice Conversion

1. 緒論 (Introduction)

1.1 動機與目的 (Motivation)

台灣的台語人口逐漸式微，普遍有越年輕越不使用台語的趨勢，最多情況就是面對長輩時，不懂得如何用台語溝通。因此本論文針對現況，希望能建置（1）中文文字轉台語語音合成系統，與（2）具語音轉換功能的多語者台語語音合成系統，讓使用者在面對講不出台語的窘境時，能知道要如何用台語表達，並能聽到以自己講話的聲音合成的台語語音。

阻礙達成此目標的主要問題是缺乏台語語音語料，因此，我們首先針對此台語語料基礎建設之需求，建置台文語音語料庫（Taiwanese Across Taiwan，TAT）作為研發台語語音合成系統之基礎，其包括（1）TAT-Vol1~2，其為橫跨台灣錄製的 100 小時/200 人之台語語音辨認（automatic speech recognition, ASR）與多語者語音合成（multi-speaker text-to-speech, TTS）用語料庫，與（2）TAT-TTS-M1~2 與 TAT-TTS-F1~2，此為依據台語強勢（高雄）腔與次強勢（台北）腔，各錄製一男一女，每人十小時之台語語音合成用語料庫。此外，為使台語語音合成的自然度能逼近真人，本論文也針對 TAT-TTS 語料庫進行人工台語變調與韻律標註，包括校正語料中每個字的音調，與加上語音韻律階層邊界標記，以改善合成語音的韻律流暢度。我們即利用此 TAT 語料庫，訓練前述之台語語音合成與語音轉換系統。

此外，我們考慮在中文文字轉台語語音合成系統系統方面，需要能即時將中文文字轉譯成台語語音，因此除採用高品質（state-of-the-art）的 end-to-end（E2E）Tacotron 2 語音合成主架構，並再加上基於 convolution neural network（CNN）之 sequence-to-sequence 中文文字轉台語台羅拼音機器翻譯前級，與可即時合成語音的 WaveGlow 語音合成後級。

而為盡量能多樣化合成台語語音的（人物）音色，在語音轉換架構方面，我們改用基於串接語音辨認（automatic speech recognition, ASR）與語音合成（text-to-speech, TTS）模組之 cascaded ASR+TTS 架構。其中 ASR 與 TTS 模組，都將使用基於 end-to-end（E2E）架構的 Transformer 類神經網路。

最後，我們使用 mean opinion score（MOS）主觀評分方式，測試台語語音合成與語音轉換系統的『自然度』，與比較語音轉換系統的合成音檔與目標語者音色的「相似度」。此外，如何在有限的目標語者訓練語料限制情況下，盡量維持住語音轉換後音檔的自然度和相似度，也是語音轉換系統評估成效的重點。

1.2 背景 (Background)

目前主流的語音合成系統，幾乎都是基於類神經網路技術，比如 Google 提出的 Tacotron2+WaveNet Vocoder 架構(Shen *et al*., 2018)。Tacotron2 可直接以類神經網路，進行文脈訊息處理，建立一「文字」轉「Mel-Spectrogram」的 end-to-end 架構。WaveNet Vocoder 接著將「Mel-Spectrogram」轉成「Speech Waveform」。此架構出現以後，語音

合成的音質就幾乎接近人聲。但此處的 WaveNet Vocoder，是一個以 sample 為單位做計算的序列式遞迴網路架構，sample 需要一個接著一個照前後順序產生。除計算量相當大外，也不易平行化，導致語音生成速度非常慢。而 NVIDIA 提出的 WaveGlow (Prenger *et al*., 2018)則正好可以解決此問題，WaveGlow 是一個基於流的生成模型，其利用批次取樣與分佈轉換處理，避開遞迴網路架構計算量大，且不易平行化的問題，合成所需時間比大幅減少，若是合成約 10 秒以下的語音，大幅減少的合成時間已經幾乎接近體感的即時合成，且其公開的平均意見得分（MOS）測試也表明，WaveGlow 的音質也不遜於 WaveNet。

而在語音轉換方面，國際研究社群在 2016，2018 和 2020 年，分別舉辦多次的 voice conversion challenge（VCC）競賽。2020 年的比賽(Zhao *et al*., 2020)分成了 Task1：同語言的半平行語音轉換（Intra-lingual semi-parallel VC)，和 Task2：跨語言語音轉換（Cross-lingual VC）兩種任務。在 VCC 2020 中，各參賽隊伍在語音特徵轉換和後端聲碼器的選擇如圖 1 所示。由 Task1 的語音轉換音檔自然度及相似度比賽結果（如圖 2 所示），顯示在同語言的情況下，得益於 ASR 和 TTS 技術近幾年的快速發展，一些基於聲學後驗圖譜（phonetic posteriorgram，PPG），或是 ASR 加 TTS 架構的語音轉換技術，在語音轉換音檔的自然度與相似度上，被普遍認為優於以往的自動編碼器（Auto-Encoder）或是生成對抗網路（GAN）等系統。而由 Task2 比賽結果（如圖 3 所示）可以看出來，PPG 和 cascaded ASR+TTS 的方法，都很適合用在跨語言語音轉換情況。

Team ID	Task 1		Task 2	
	VC model	Vocoder	VC model	Vocoder
T01	PPG-VC (Tacotron)	Parallel WaveGAN	N/A	N/A
T02	PPG-VC (Tacotron)	WaveGlow	PPG-VC (Tacotron)	WaveGlow
T03	AutoVC	WaveRNN	AutoVC	WaveRNN
T04	VQVAE	WaveNet	N/A	N/A
T05	N/A	N/A	PPG-VC (IAF)	WORLD & WaveGlow
T06	StarGAN	WORLD	StarGAN	WORLD
T07	NAUTILUS (Jointly trained TTS VC)	WaveNet	NAUTILUS (Jointly trained TTS VC)	WaveNet
T08	VTLN + Spectral differential	WORLD	VTLN + Spectral differential	WORLD
T09	AutoVC	Parallel WaveGAN	AutoVC	Parallel WaveGAN
T10	ASR-TTS (Transformer) / PPG-VC (LSTM)	WaveNet	PPG-VC (LSTM)	WaveNet
T11	PPG-VC (LSTM)	WaveNet	PPG-VC (LSTM)	WaveNet
T12	ADAGAN	AHOcoder	ADAGAN	AHOcoder
T13	PPG-VC (Tacotron)	WaveNet	PPG-VC (Tacotron)	WaveNet
T14	One shot VC	NSF	N/A	N/A
T15	N/A	N/A	AutoVC	MelGAN
T16	CycleVAE	Parallel WaveGAN	CycleVAE	Parallel WaveGAN
T17	Cotatron	MelGAN	N/A	N/A
T19	VQVAE	Parallel WaveGAN	VQVAE	Parallel WaveGAN
T20	VQVAE	Parallel WaveGAN	VQVAE	Parallel WaveGAN
T21	CycleGAN	MelGAN	N/A	N/A
T22	ASR-TTS (Transformer)	Parallel WaveGAN	ASR-TTS (Transformer)	Parallel WaveGAN
T23	Transformer VC (Jointly trained TTS VC)	Parallel WaveGAN	CycleVAE	WaveNet
T24	PPG-VC (Tacotron)	LPCNet	PPG-VC (Tacotron)	LPCNet
T25	PPG-VC (CBHG)	WaveRNN	PPG-VC (CBHG)	WaveRNN
T26	One shot VC	Griffin-Lim	One shot VC	Griffin-Lim
T27	ASR-TTS (Transformer)	Parallel WaveGAN	PPG-VC / ASR-TTS (Transformer)	Parallel WaveGAN
T28	Tacotron	WaveRNN	Tacotron	WaveRNN
T29	PPG-VC (CBHG)	LPCNet	PPG-VC (CBHG)	LPCNet
T31	Multi-speaker Parrotron	WaveGlow	Multi-speaker Parrotron	WaveGlow
T32	ASR-TTS (Tacotron)	WaveRNN	ASR-TTS (Tacotron)	WaveRNN
T33	ASR-TTS (Tacotron)	Parallel WaveGAN	PPG-VC (Transformer)	Parallel WaveGAN

圖 1. VCC 2020 參賽者使用模型架構詳細資訊(Zhao et al., 2020)
[Figure 1. Summary of adopted approaches in Voice Conversion Challenge (VCC) 2020]

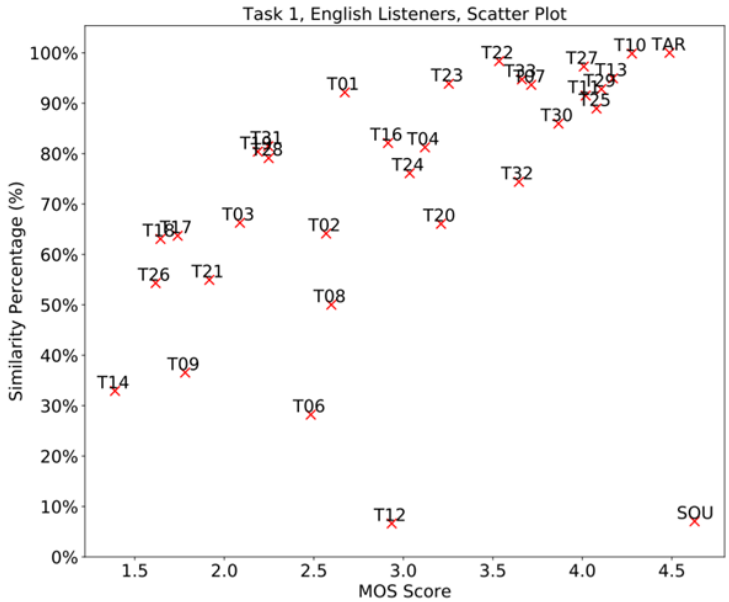


圖 2. VCC 2020 Task1 比賽結果(Zhao et al., 2020)
[Figure 2. Benchmark Results on Task1 of Voice Conversion Challenge (VCC) 2020]

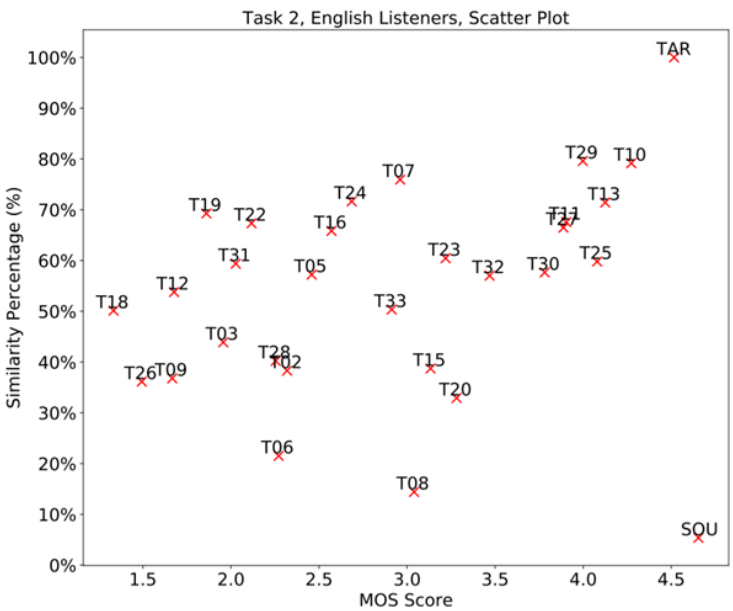


圖 3. VCC 2020 Task2 比賽結果(Zhao et al., 2020)
[Figure 3. Benchmark Results on Task2 of Voice Conversion Challenge (VCC) 2020]

此外，從整體比賽結果來看，同語言任務的自然度與相似度分數，都已經可以達到相當不錯的分數。而跨語言語音轉換，因難度較高，在自然度和相似度都還有進步的空間，不同語言說話的方式還是一定程度上影響了語音轉換的成效。

1.3 研究方法 (Approaches)

目前還較少有人嘗試台語的語音合成與語音轉換技術。以下說明我們建置（1）中文文字轉台語語音合成系統，與（2）具語音轉換功能的多語者台語語音合成系統的主要目標與研究方法。

1.3.1 單人台語語音合成系統 (Single-Speaker Chinese Text-to-Taiwanese Speech Synthesis)

此系統使用 sequence-to-sequence-based 中文文字轉台語台羅拼音機器翻譯前級，串接 Tractron 語音合成主架構，與 WaveGlow 語音合成後級，以實現高品質且即時之台語語音合成系統。其系統流程如圖 4 所示：使用者輸入中文文字後，先透過機器翻譯成相對應的台語語句的台羅拼音文本，再使用 Tacotron 2 將台羅拼音文本轉換成台語合成語音的頻譜，最後用 WaveGlow 將台語語音的頻譜解碼成實際的台語語音。此系統最後並被做成線上展示網頁(http://tts001.iptcloud.net:8801)，公開供大眾測試。

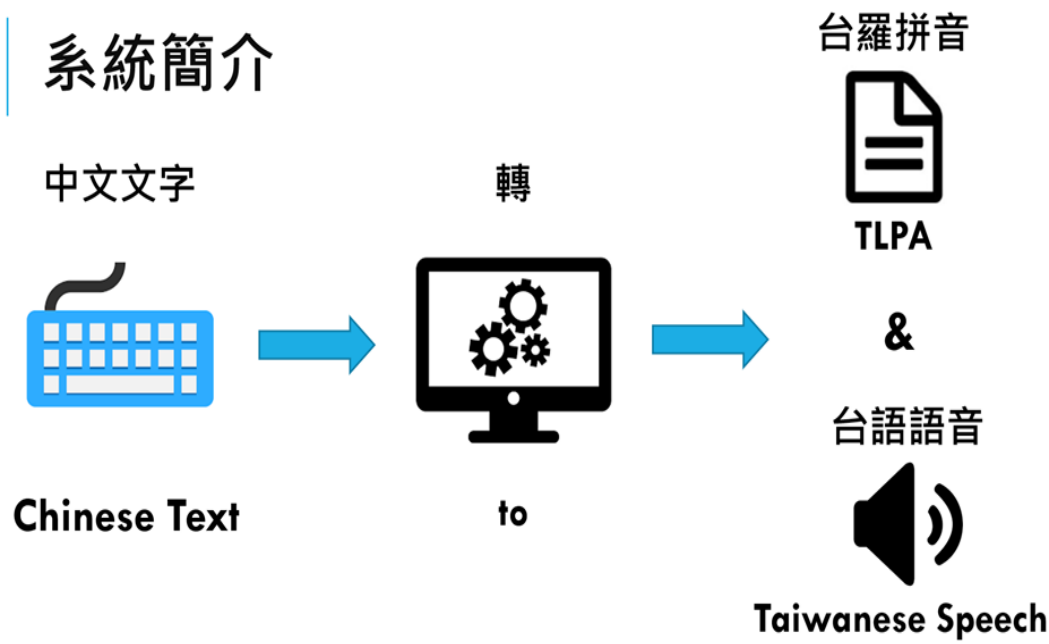


圖 4. 單人台語語音合成系統流程簡介
[Figure 4. Block-diagram of the Single-Speaker Taiwanese Text-to-Speech System]

1.3.2 結合語音辨認及合成模組之台語語音轉換系統 (Cascade ASR and TTS-based Taiwanese Voice Conversion System)

此部分採用 Cascade ASR+TTS 架構(Huang *et al*., 2020)進行台語語音轉換系統的初步建置，其架構如圖 5 所示。其包含三個模型，分別是負責台語語音辨認的語者獨立 Transformer-based ASR model，負責擷取目標語者語音特徵的 X-Vector，以及負責依據目標語者 X-Vector，進行台語語音合成的 Multi-speaker Transformer-TTS model。

其在訓練階段，利用少量目標語者的聲音，求取其 X-Vector，並微調（fine-tuning）事先預備好的多語者 TTS 模型，產生目標語者的 TTS 模型。因此，在測試階段，就能利用 ASR 前級，把來源語者的音檔辨識出文字，並用微調好的 TTS 合成出具目標語者語音特徵的語音檔。

此外，我們也針對跨語言語音轉換任務，將原先 VCC 2020 Task2 中的華語轉英語的專案(Kamo, 2021)，替換成華語轉台語的語音轉換系統。

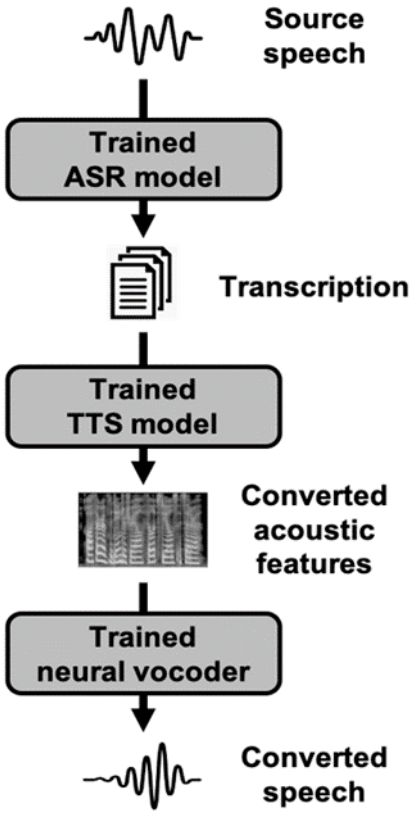


圖 5. 結合語音辨認及合成模組之台語語音轉換系統流程簡介 (Huang et al., 2020)

[Figure 5. Flowchart of the Cascade ASR and TTS-based Taiwanese Voice Conversion System]

2. 相關研究 (Related Works)

2.1 語音合成 (Text-to-Speech)

語音合成，是一種將文本轉換成語音（Text-To-Speech，TTS）的技術，從最早期的音檔串接合成，發展到使用隱藏式馬可夫模型，一直到現今的類神經網路，電腦合成語音的聲音已經到了幾乎跟真人相似的程度。

2.1.1 Tacotron2

Tacotron2 的架構圖(Shen *et al*., 2018) 如圖 6 所示，其使用 encoder-decoder + Location Sensitive Attention 的架構。將文本資訊輸入 encoder 後，encoder 類神經網路進行文本分析，萃取出輸入文句的文脈特徵參數，decoder 接著依據文脈特徵參數，以 attention 權重機制，對齊（alignment）輸入文字與產出的合成語音的梅爾頻譜圖，最後再由 WaveNet Vocoder 將梅爾頻譜進行解碼，合成高品質的語音。

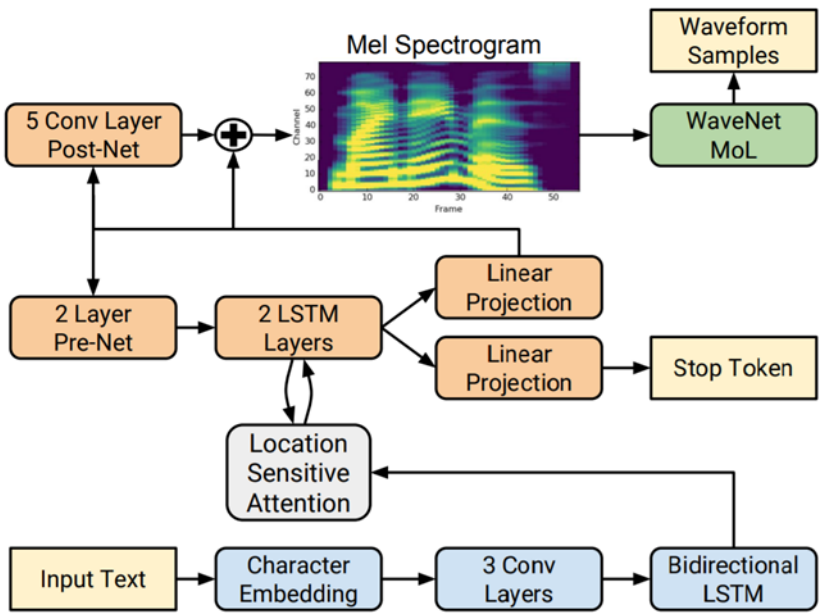


圖 6. Tacotron2 架構流程圖(Shen et al., 2018)
[Figure 6. Architecture of the Tacotron2 Speech Synthesis System]

與同是 Google 提出的對前一代的 Tacotron (Wang *et al*., 2017)相比，其主要改良是取消 CBHG，改用普通的 LSTM 和 Convolution layer，接著每一步 decoder 改成只生成一個 frame，並在後面增加了一個 5 層的後置 CNN 網路，來使產出的梅爾頻譜更正確，最後重點是將 Tacotron 後端的 Griffin-Lim，改為能合成出更像真人講話聲音的 WaveNet Vocoder。

2.1.2 WaveGlow

上述提到的 Tacotron2 已經可以做到合成出接近人聲的音檔，但後端的 WaveNet Vocoder 卻有著語音生成速度緩慢的缺點。而 NVIDIA 提出的 WaveGlow 架構(Prenger *et al*., 2018)即是此問題的解決方法，WaveGlow 是一種 flow-based generative networks，其架構圖如圖 7 所示，主要是透過一可以完美執行逆轉換的音檔波形正規化轉換類神經網路架構，在給定輸入音檔的相對應頻譜的條件下，將所有可能的語音音檔波形折疊投影到一高斯分布 z 空間。因此，在合成時只需在 z 空間中作取樣，再依據給定的合成音檔的對應頻譜條件，即可以類神經網路執行逆轉換，將取樣出的 z 向量，反折疊回實際的音檔波形。此外，因為這些逆轉換的類神經網路運算過程，都可以進行批次平行化處理，所以最終可以即時合成高品質語音。

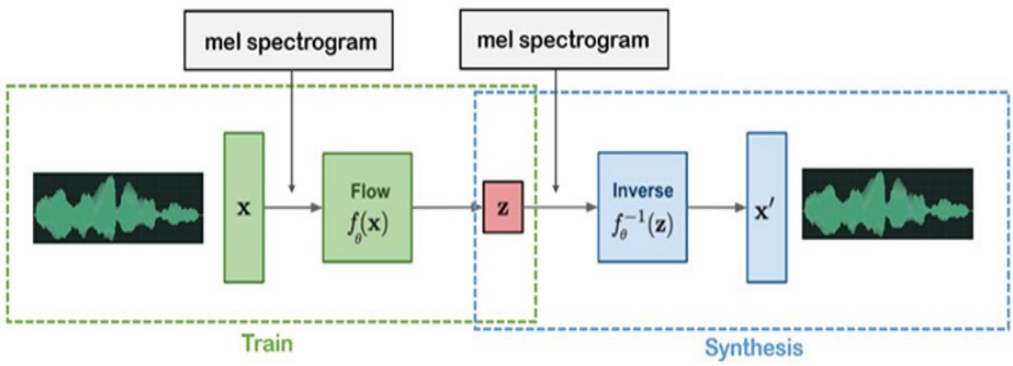


圖 7. WaveGlow 訓練及合成過程(Prenger et al., 2018)
[Figure 7. Training and Synthesis processing of the WaveGlow Approach]

其中，可完美執行逆轉換的類神經網路的實際架構圖如圖 8 所示。訓練時依據基於輸入與輸出音檔相似度的 cost function 導引，依據給定的合成音檔頻譜條件，多次利用可逆卷積（invertible convolution）與耦合層（coupling layer）網路，進行函式轉換，逐步學習如何將真實語音波形訊號 x，投射到一具高斯分佈之隱藏變數 z 的向量空間。並在訓練時限制 mapping 函式為可逆函式。如此，WaveGlow 在生成語音波型時，即可在隱藏變數 z 空間進行取樣，再依據給定的合成音檔頻譜條件，經多次逆函式轉換，逐步將 z 向量，反解成真實語音波形訊號 x。

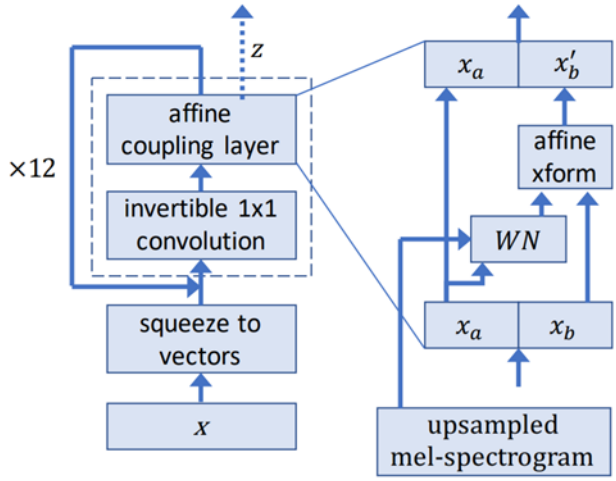


圖 8. WaveGlow 網路結構圖(Prenger et al., 2018)
[Figure 8. The architecture of the WaveGlow System]

2.2 語音轉換 (Voice Conversion)

較早的語音轉換方法，在有平行錄音語料的情形下，有統計式及樣本式兩種方法，例如高斯混合模型（Gaussian mixture model, GMM）與基於局部線性嵌入(locally linear embedding, LLE)的語音轉換方法等。其中，統計式語音轉換方法的主要缺點，為轉換後所得到的語音頻譜有過度平滑的現象，因而降低了語音品質與語者相似度。而樣本式語音轉換方法的優點，則是不需要模型訓練過程，但是樣本數量的多寡會影響轉換後語音的音質，而且轉換過程的運算量會隨著樣本數量增加而提高。

隨著類神經網路的高度發展，基於神經網路的語音轉換技術也漸漸成為主流。像是變分式自動編碼器（variational autoencoder, VAE）架構(Hsu *et al*., 2016)，可以用來串在一通用語音合成系統的後級，用以在只有少量語者語音資料的情形下，轉換通用語音合成系統的輸出音色，如圖 9 所示。

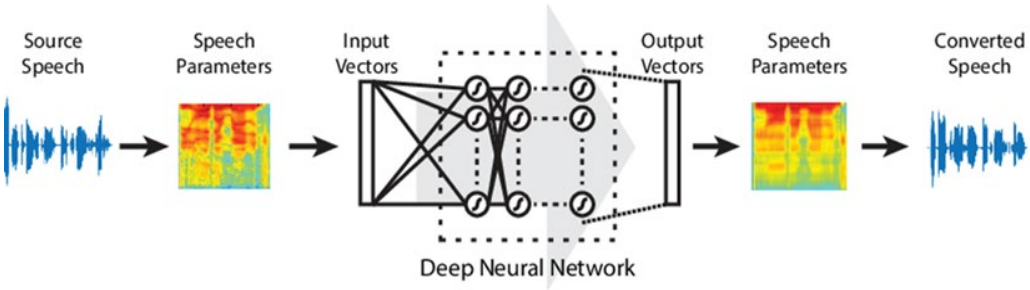


圖 9. 基於變分式自動編碼器（variational autoencoder, VAE）之語音轉換方法 (Hsu et al., 2016)
[Figure 9. Variational autoencoder-based voice conversion]

此外，亦有利用語音辨認器作為輔助的方法，例如，以語音辨認器，計算語音中的聲學後驗圖譜（phonetic posteriorgram，PPG）(Sun *et al*., 2016)，並依此圖譜當作語音中說話內容的資訊，用以輔助語音轉換，如圖 10 所示。此做法可以容忍因前級語音辨認器的錯誤辨認輸出，導致後級語音轉換合成的錯誤。

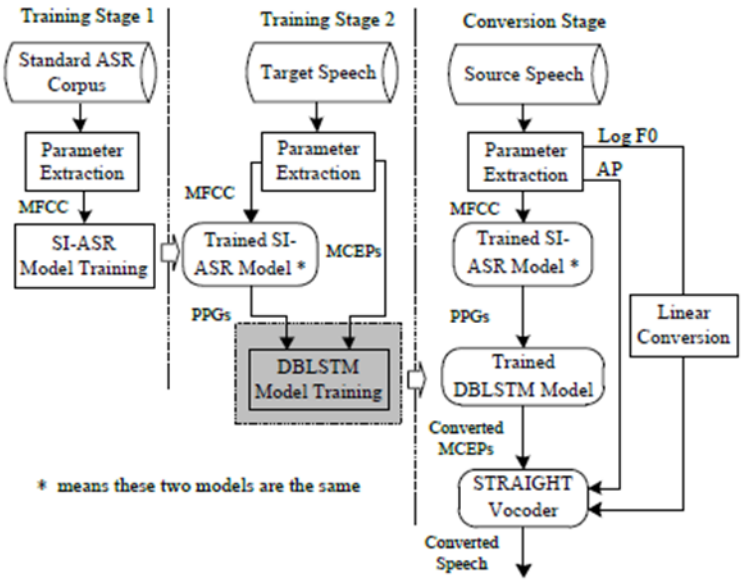


圖 10. 基於聲學後驗圖譜（phonetic posteriorgram，PPG）之語音轉換架構(Sun et al., 2016)
[Figure 10. Phonetic Posteriorgram-based voice conversion]

3. Taiwanese Across Taiwan語料庫建置 (Taiwanese Across Taiwan Corpus)

為了處理在此論文中，所述之台語語音合成與語音轉換系統開發，所需的大量台語語料問題，我們首先規劃並建置一大規模台文語音語料庫（Taiwanese across Taiwan，TAT）。其包括 TAT-Vol1~2，共有 200 位語者，約 100 小時的台語語音辨認/多語者語音合成用語料，與 TAT-TTS-M1~2 與 TAT-TTS-F1~2，考慮台語強勢（高雄）與次強勢（台北）腔，共兩男兩女，每人約 10 小時的台語語音合成專用語料。以下說明 TAT 語料庫的設計，錄音標準作業程序，人工校正作業程序與語料庫發行等程序。

3.1 語料庫目的與設計 (Corpus Design)

我們的目的是想要製作一個能反應台文與台語使用現況，並涵蓋大部分台語腔調的台文語音語料庫。因此我們需要收集豐富的原生台語文本，並在台灣各地招募台語發音人，進行語料收集。並整理成具錄音資訊後製資料（metadata）的電子化語料庫。所以，此台語語音語料庫設計如下：

- 目標台語文本內容：使用多種題材的台語原生短文章（不用翻譯文章），以充分反應台文與台語使用現況。並包含日常生活對話、數字，地址等常用語句，以利開發語音辨認與語音合成應用。
- 目標台語發音語者：在台灣各區域招募不同性別與年紀的當地台語語者，並要求其照自己平常的習慣自然發音，以蒐集台灣不同區域的人真實使用的不同台語腔調。
- 目標錄音環境：針對語音辨認應用，需使用多種麥克風，在一般安靜辦公室環境錄音，以收集多種麥克風通道與環境變化。針對語音合成應用，則需進具高階隔音與殘響抑制的專業錄音室，以高階電容式麥克風，錄制無任何背景噪音與殘響的高品質音檔。
- 電子化檔案格式：語料庫必須包含錄音資訊後製資料的電子化檔案格式，最終形式需為一個 Microsoft Waveform 格式的語音音檔，配合一個對應的 json 格式文字檔，其中記錄了人工校正後產生的台羅數字調，音檔長度或是錄音者使用的腔調等相關資訊，檔案格式範例如圖 11 所示。


```
{
    "音檔長度": "9.61",
    "漢羅台文": "我厝內的電話是空二三三六六九空五四",
    "台羅": "guá tshù-lāi ê tiān-uē sī khòng jī sam sam lio͘k lio͘k kiú khòng ngóo sù",
    "台羅數字調": "gua2 tshu3-lai7 e5 tian7-ue7 si7 khong3 ji7 sam1 sam1 liok8 liok8 kiu2 khong3 ngoo2 su3",
    "白話字": "góa chhù-lāi ê tiān-ōe sī khòng jī sam sam lio͘k lio͘k kiú khòng ngó͘ sù",
    "字數": "17",
    "提示卡編號": "0011",
    "句編號": "1.1",
    "發音人": "IUF001",
    "性別": "女",
    "年齡": "51",
    "教育程度": "博士",
    "出生地": "高雄市",
    "現居地": "台北市文山區",
    "腔調": "漳泉濫",
    "錄音環境": "安靜隔音室內",
    "提示卡切換速度": "快",
    "總錄音時間(分)": "90"
}
```


圖 11. TAT 語料庫 json 文檔範例
[Figure 11. Recording Metadata Example]

3.2 錄音作業協定 (Recording Protocol)

依據上述語料庫目的，我們訂定了以下語料庫建置的錄音標準作業流程，如圖 12 所示，包括文本蒐集、提示卡製作、發音人招募、發音人錄音預備、實際錄音、人工校正與語料庫釋出等流程。以下將分別介紹各程序的進行方式。

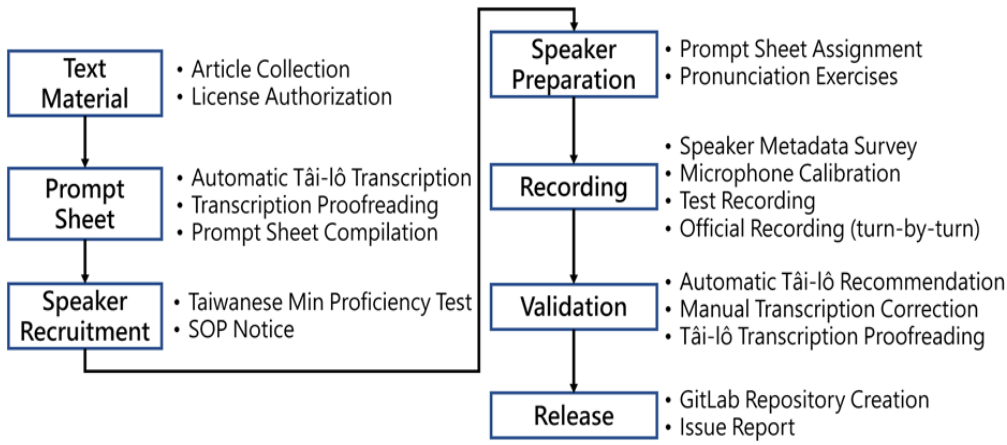


圖 12. TAT 語料庫建置流程
[Figure 12. The recording protocol of the TAT corpus]

3.2.1 台語原生文本蒐集與提示卡製作 (Native Taiwanese Prompt Sheets)

錄音的文本的蒐集來源，主要來自於李江卻台語文教基金會[1]，由其聯繫曾在基金會出版品發表文章的 50 位作者，每人蒐集約 6000 字的文本。再加上基金會本身出版的台語日常對話課程教材，與程式依樣板隨機產生的數字串、電話號碼與地址等常用語句。

依據上述文本，製作出的約 50 份提示卡，其內容包含三大類，範例如圖 13，圖 14 以及圖 15 所示，分別為(1)數字資料，像是地址、日期、電話等等，還有(2)日常對話以及(3)短文等，分別以句或是短段落（1~3 短句）為單位編號條列之。每一個句子，除提供台文文句外，並標上參考（不限制其發音）用的對應台羅拼音文句，以便不熟習台文

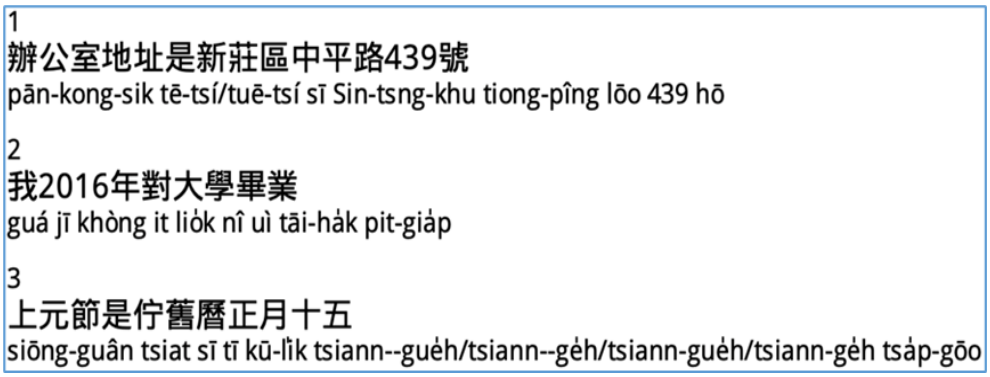


圖 13. 錄音提示卡-數字資料部分範例
[Figure 13. A typical exampleof prompt sheet: address, date and digits]

[1] i kang khioh taiwanese cultural and educational foundation https://www.tgb.org.tw/

的發音人事先溫稿。最後，每份提示卡內容長度，則是設定成一人份，總共約錄製出 30 分鐘語音檔。

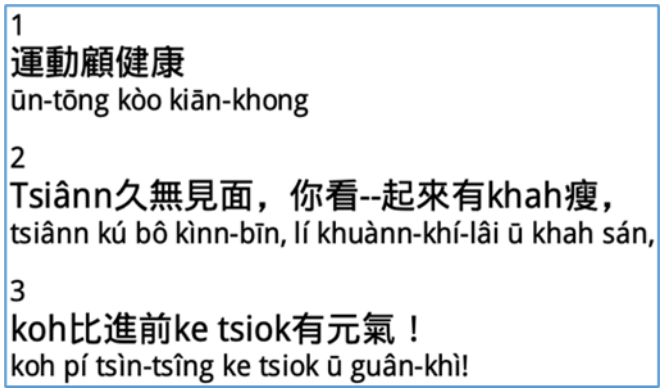


圖 14. 錄音提示卡-日常對話部分範例

[Figure 14. A typical exampleof prompt sheet: daily conversation session]

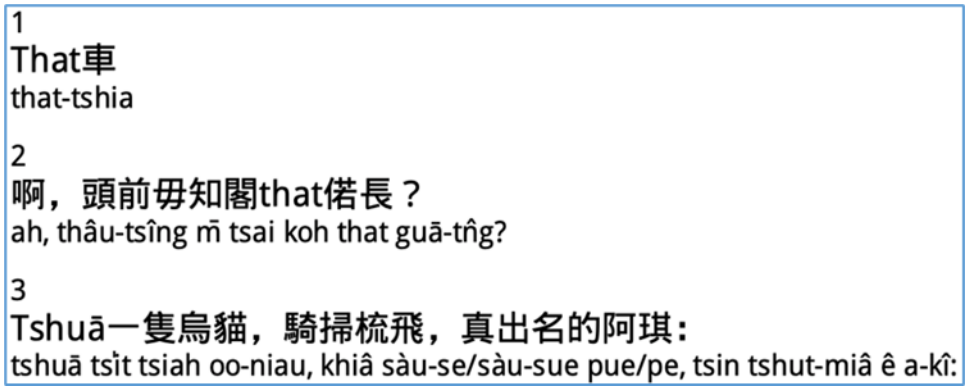


圖 15. 錄音提示卡-短文部分範例

[Figure 15. A typical exampleof prompt sheet: short article session]

3.2.2 ASR錄音程序 (Recording Procedure for ASR Corpus)

3.2.2.1 錄音員招募與錄音員預備 (Speaker Recruitment)

錄音提示卡準備好後，我們與台灣各地的教授合作，就近在其學校附近，招募適合的台語發音人，並指定其所使用的提示卡的編號，要求其先溫稿後，約定時間進行錄音。以涵蓋全台灣各地使用各種不同台語腔調的台語使用者。在 TAT-Vol1~2 語料庫錄製過程中，我們共與五位教授合作，包括師範大學許慧如教授，台中教育大學楊允言教授及程俊源教授，中正大學蔡素娟教授與成功大學陳麗君教授，其學校所在區域如圖 16 所示，每位教授負責招募與錄製約 50 位語者，因此一份提示卡，最多只會能給四個人使用。

圖 16. TAT 語料庫錄製合作教授分布區域圖
[Figure 16. Distribution of locations of recording campuses]

3.2.2.2 錄音設備配置 (Equipment Configuration)

ASR 語料蒐集，通常選在安靜的一般辦公室，會議室或教室作為錄音環境。錄音時，為了模擬不同的使用情境下錄到的聲音，整個錄音設備配置如圖 17 所示，使用筆電，透過 USB 介面，連接 Zoom H6[2]多軌數位錄音介面，同時抓取 6 隻麥克風的訊號，一次錄製 6 軌音檔，以蒐集不同麥克風與錄音通道的影響。

[2] ZOOM CORPORATION, H6 Handy Recorder Operation Manual, Available: https://www.zoom.co.jp/sites/default/files/products/downloads/pdfs/E_H6v2.pdf

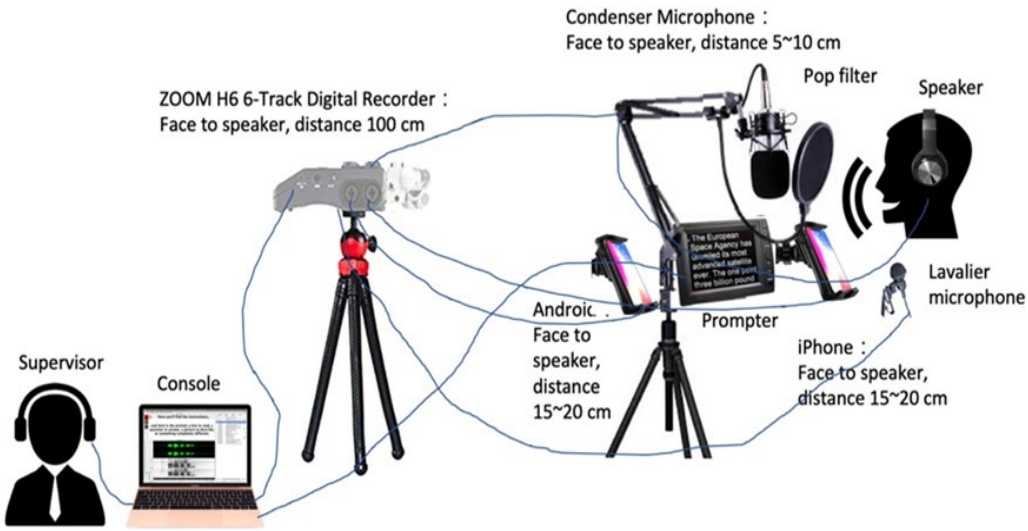


圖 17. TAT 語料庫錄製情形模擬圖
[Figure 17. Configuration of the recording equipments]

其中，Zoom H6 的最高取樣頻率，可以設置到 192 kHz。而此六隻麥克風的模擬使用情境，可分為三類，第一種是距離最遠的兩個麥克風，約距離錄音者 1 公尺，分別為 ZOOM XYH-6 左聲道和 ZOOM XYH-6 右聲道，用以模擬遠距麥克風收音效應，其會含有較多空間殘響與背景噪音：第二種是距離最近的兩個麥克風，約距離錄音者 5 到 10 公分，分別為放在發音人嘴巴正前方的電容式麥克風，和別在胸口的領夾式麥克風，用以收集近場語音，其聲音應該最為乾淨清楚：第三種是約距離錄音者 15 到 20 公分，位居下方偏右的 ios 手機，以及下方偏左的 android 手機的內建麥克風，分別用來呈現一般手機使用情形下的收音情況。

3.2.2.3 錄音程序 (Recording Procedures)

錄音現場示意圖如圖 18 所示，由一名監錄員使用筆電，執行德國慕尼黑大學開發的 SpeechRecoder 語料蒐集錄音軟體[3]，一次一句將提示卡內容，投放至發音人正前方的提示卡螢幕（外接螢幕），給發音人觀看。請發音人按照提示卡螢幕上的文句，進行錄音。同時，監錄員利用耳機，與筆電主螢幕上 SpeechRecoder 錄音軟體所畫出的六軌音檔波形，進行音檔監聽與波形監看。確認每次錄音，音檔聲音的大小聲，錄音說話內容及說話順暢度皆正確後，再接續投放提示卡中的下一個文句。如此重複進行直至全部錄音文本錄音完畢。

[3] SpeechRecorder. Available: https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

圖 18. ASR 錄音現場示意圖
[Figure 18. Photo of the recording site]

3.2.3 TTS錄音程序 (Recording Procedures for TTS)

3.2.3.1 錄音員招募與錄音員預備 (Speaker Recruitment)

TTS 的部分則是請李江卻台語文教基金會招募了共兩男兩女 4 名語者，代號分別為 M1，M2，F1 與 F2，4 位語者的相關資訊如表 1 到表 4 所示。其中 M1 以及 F1 語者的腔調為台灣台語強勢腔（偏漳州腔/高雄腔），M2 跟 F2 則為次強勢腔（偏泉州腔/台北腔）。此四位發音人需把所有提示卡全部念完，完成共約十小時的台語語音，以蒐集足夠台語語音合成用語料。

表 1. M1 語者資訊
[Table 1. Personal information about Speaker M1]

性別	年齡	教育程度	出生地	現居地	腔調
男	34 歲	大學	台北市士林區	台北市士林區	偏漳州腔

表 2. M2 語者資訊
[Table 2. Personal information about Speaker M2]

性別	年齡	教育程度	出生地	現居地	腔調
男	55 歲	大學	新北市汐止區	新北市汐止區	泉州安溪腔

表 3. F1 語者資訊
[Table 3. Personal information about Speaker F1]

性別	年齡	教育程度	出生地	現居地	腔調
女	52 歲	碩士	高雄市新興區	新北市新店區	漳州腔

表 4. F2 語者資訊
[Table 4. Personal information about Speaker F2]

性別	年齡	教育程度	出生地	現居地	腔調
女	41 歲	碩士	台中市梧棲區	台北市中正區	泉州腔

3.2.3.2 錄音設備配置 (Equipment Configuration)

與 ASR 不同， TTS 用合成語料必須排除背景雜音，讓聲音越乾淨越好。因此 TTS 專用語料是在能隔絕背景底噪（如圖 19 左方所示）及抑制空間殘響（如圖 19 右方所示）的專業錄音室中錄音。並且錄製 TTS 時必須使用音質最好的高階電容式麥克風，近距離進行錄製。

圖 19. 專業錄音室示意圖
[Figure 19. Photos of the professional recording studio]

3.2.3.3 錄音程序 (Recording Procedures)

TTS 語料蒐集的錄音現場示意圖如圖 20 所示。需由一名音響工程師，操作專業錄音工作站軟體，確認音檔聲學特性平穩一致（大小聲、速度與韻律等等）。此外，並需再加上一名具台語老師等級的發音監錄員，同時一句一句監督每句台語發音的正確性。此外，第一次錄音時，需先錄製數句校正句，作為範本，在每次開始新的錄音工作（session）前，播放給發音人聆聽，讓發音人校正對齊其發音特性。而且，一次錄音工作的時間不能太長，必須讓發音人有足夠休息時間，以維持其發音特型一致，不要偏掉。

圖 20. TTS 錄音現場示意圖
[Figure 20. Photo of the recording workstation]

3.2.4 人工校正 (Transcription)

錄製好的語音檔，最後分別由五位合作教授的團隊（負責 TAT-Vol1~2），與李江卻台語文教基金會 的工作人員（負責 TAT-TTS-M1~2 與 TAT-TTS-F1~2），利用意傳科技的線上語料庫校正輔助工具（如圖 21 所示），以人工逐句聆聽校正文字檔，與依據發音人的實際發音，標上台羅正確拚音，產生最終的語料庫。

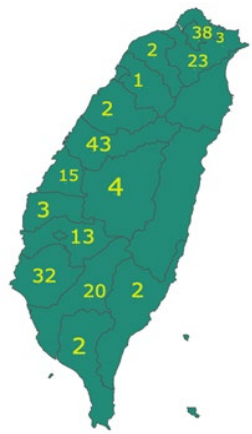
圖 21. 意傳科技語料庫校正輔助工具示意圖
[Figure 21. User interface of the online corpus annotation tool]

3.3 語料庫內容統計 (Statistics of TAT Corpus)

以下介紹完成的 TAT-Vol11~2 與 TAT-TTS-M1~2 與 TAT-TTS-F1~2 語料庫的內容統計資料。

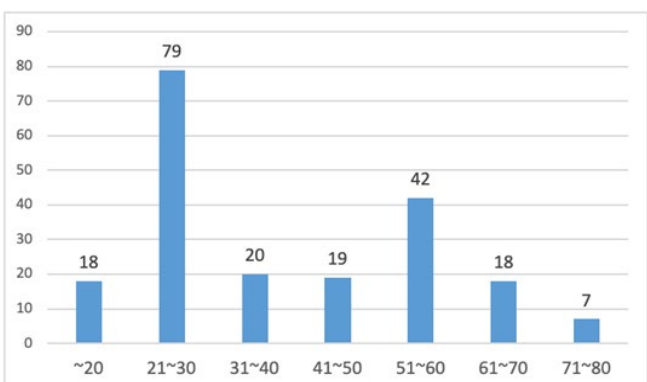
3.3.1 TAT-Vol1~2

在台語語音辨認用語料蒐集部分，經過一年的辛苦錄音後，最終招募了台灣各地使用各種不同台語腔調的台語語者共約 200 人，包括男生 91 人，女生 109 人，其在台灣的地域/人數分布如圖 22 所示，年齡分佈從 18 到 80 歲都有，分佈如圖 23 所示。

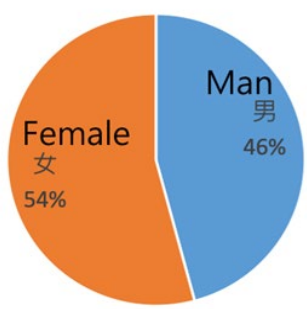


Region Distribution

圖 22. TAT-Vol1~2 語料庫，錄音語者地域分布
[Figure 22. Distribution of speakers in TAT-Vol1~2 corpus]



Age Distribution



Gender Distribution

圖 23. TAT-ASR 錄音語者年齡分布及性別比
[Figure 23. Distribution of ages and sexs of recorded speakers in TAT-Vol1~2 corpsu]

目前完成的語料庫總時數共約 104.36 小時，經均分成 TAT-Vol1~2 兩集語料庫，委託『社團法人中華民國計算語言學學會』[4]公開發行。各子集的詳細的句數，字數與時數統計數據如圖 24 所示。

TAT-ASR-Vol1			
Speakers	Sentences	Characters	Hours
100	28833	339592	51.94
TAT-ASR-Vol2			
Speakers	Sentences	Characters	Hours
100	28978	340607	52.42
Total			
Speakers	Sentences	Characters	Hours
200	57811	680199	104.36

圖 24. TAT-ASR 語料庫詳細統計數據
[Figure 24. Statistics of the TAT-Vol1~2 corpus]

3.3.2 TAT-TTS-M1~2與TAT-TTS-F1~2 (TAT-TTS-M1~2 and TAT-TTS-F1~2)

在台語語音合成專用的語料庫方面，最終錄製了 2 名男生和 2 名女生的語料。語料庫完成品形式同樣為一個音檔，配對一個相同檔名的 json 檔。此 json 檔的格式，如圖 25 所示，含有漢羅台文、台羅數字調、音檔長度或是發音人的屬性等相關資訊。

M1_1-1.json

```
{
    "音檔長度": "4.33",
    "漢羅台文": "台灣需要主動的孤單事務大臣",
    "台羅": "Tâi-uân su-iàu tsú-tōng ê koo-tuann sū-bū tāi-sîn",
    "台羅數字調": "tai5-uan5 su1-iau3 tsu2-tong7 e5 koo1-tuann1 su7-bu7 tai7-sin5",
    "白話字": "Tâi-oân su-iàu chú-tōng ê ko͘-toaⁿ sū-bū tāi-sîn",
    "字數": "13",
    "提示卡編號": "M1_1",
    "句編號": "M1_1-1",
    "發音人": "M1",
    "性別": "男",
    "年齡": "34",
    "教育程度": "大學",
    "出生地": "台北市士林區",
    "現居地": "台北市士林區",
    "腔調": "偏漳州腔",
    "錄音環境": "專業錄音室",
    "提示卡切換速度": "",
    "總錄音時間(分)": ""
}
```

圖 25. TAT-TTS json 文檔範例
[Figure 25. A typical example of the recording metadata]

[4] https://www.aclclp.org.tw/corp_c.php

此語料也已經委託『社團法人中華民國計算語言學學會』公開發行。目前共完成 4 位語者，每位語者約 10 小時的語料，總時數共約 40.6 小時。音檔詳細資訊如圖 26 所示，以人為單位分成四集，包括 TAT-TTS-M1~2 與 TAT-TTS-F1~2，其中 M 與 F 分別為男生與女生語者的代號，1 與 2 則分別為強勢腔與次強勢腔的編碼。

TAT-TTS-M1						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
9625	10.4	wav	2	192000	24-bit	24-bit Floating Point PCM
TAT-TTS-M2						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
11532	10.1	wav	2	192000	25-bit	32-bit Floating Point PCM
TAT-TTS-F1						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
12917	10.0	wav	2	48000	24-bit	24-bit Signed Integer PCM
TAT-TTS-F2						
Sentences	Hours	Extension	Channels	Sample Rate	Precision	Sample Encoding
12422	10.1	wav	2	48000	24-bit	24-bit Signed Integer PCM

圖 26. TAT-TTS-M1~2 與 TAT-TTS-F1~2 語料庫的詳細音檔資料與句數、時數等統計資訊
[Figure 26. Statistics of the TAT-TTS-M1~2 and TAT-TTS-F1~2 corpus]

3.4 台語變調與韻律邊界標註 (Annotation of Tone and Prosodic Boundary)

我們以 TAT-TTS 的 M1 語者作為訓練語料，完成單人台語語音合成系統後，發現合成音有時有台語變調與停頓不通順的問題，這主要是當初因人力問題，在建置 TAT-TTS-M1~2 與 TAT-TTS-F1~2 時，只以台語本調做標記，並沒有針對台語變調，進行人工校正，或是加上韻律邊界標注。因此，我們針對 TAT-TTS-M1 語料庫，進一步加上中文翻譯、標註台語變調與台語韻律詞或韻律片語邊界。

3.4.1 語料庫設計 (Corpus Design)

我們訂定了以下的變調與韻律標註標準作業程序。標註方法為在原先語料的 json 檔裡面新增兩行字串，分別是(1)中文翻譯，(2)台語變調校正以及（3）加入兩種新韻律邊界符號的台羅數字調。新增標註資訊後的 json 檔的前後比較如圖 27 所示。

圖 27. TAT-TTS 台語語料校正前後比較
[Figure 27. Comparison of metadata before and after Chinese Text, tone and prosodic boundary annotation]

其中加入中文是為了以後能進行中文轉台文與中文轉台羅拼音兩種機器翻譯。而校正變調與加入自訂的兩種新韻律符號，是為了讓語音合成系統，可以學習變調規則與韻律停頓方式。

3.4.2 人工標註程序協定 (Annotation Protocol)

校正者應先聆聽每一句台語音檔，以人工整句翻譯成對應的中文文字，並在 json 中加入一行中文文字標註。範例如圖 28 所示。

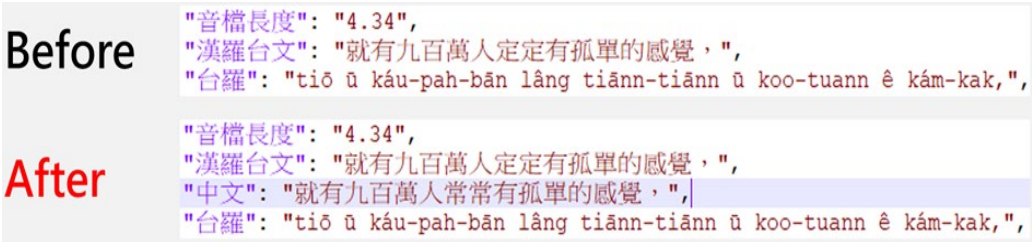


圖 28. TAT-TTS 台語語料之加入中文翻譯範例
[Figure 28. A typical example of Chinese translation]

此外，校正者需一句一句聆聽台語音檔中的變調與韻律邊界結構現象，首先將需要變調的數字調進行更正。接下來必須留意音檔停頓的位置，語音停頓的地方如為空格或標點符號為正常，不須理會。講者連續念過去的地方如為連字號或是輕聲符號(雙重連字號)，也不須理會。但如連續念過去的地方為空格，則需將空格取代為適當的韻律符號，經討論後我們定義出兩種新的韻律符號，分別為(1)韻律詞，代表符號為加號"+"，(2)韻律片語，代表符號為底線"_"。json 檔校正完成後的範例如圖 29 所示，此範例音檔為 TAT-TTS-M1 裡面的 M1_1-6.wav 音檔。

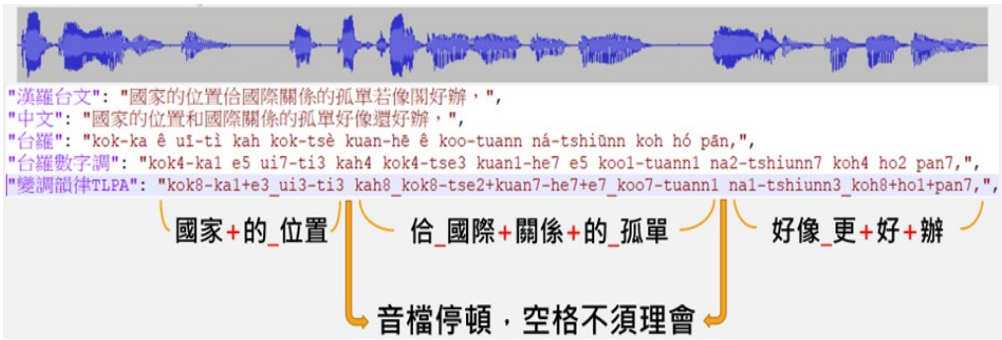


圖 29. TAT-TTS 台語語料之變調與韻律符號校正範例
[Figure 29. A typical example of tone sandhi and prosodic boundary annotation]

3.4.3 語料庫完成品統計 (Statistics of Corpus)

經過校正人員的努力，目前已經將 M1 語者的語料全數校正完畢，校正完成後的部分 json 文檔範例如圖 30 所示。後續也將針對剩下的 M2，F1 和 F2 語者的語料進行校正。

圖 30. TAT-TTS-M1 校正後 json 文檔範例
[Figure 30. A typical example of recording metadata]

4. 中文文字轉台語語音合成系統 (Chinese Text to Taiwanese Speech Synthesis System)

在踏入較複雜的台語語音轉換系統前，我們先以做出單人台語語音合成系統為目標，從中汲取台語語音合成相關的經驗，後面再繼續做多人台語語音合成系統，跟台語語音轉換系統。此外，因大多數人無法讀寫台文或是台羅拼音，因此我們額外在台語語音合成系統的前端，再加上一個中文文字轉台羅拼音的機器翻譯模組，製作一『中文文字轉台語語音合成系統』。

4.1 系統架構 (System Architecture)

此次單人台語語音合成系統的建置，以前端的中文轉台羅拼音（Chinese to Taiwanese Tâi-Lô Pinyin (TLPI), C2T）機器翻譯模組，加上後端的 Tacotron2+WaveGlow 語音合成架構為 baseline。訓練 Tacotron2 的台語語料則選用 TAT-TTS-M1 的男生強勢腔語者，並選擇 json 文檔中「台羅數字調」當作訓練文本，機器翻譯使用語料則為開源之 iCorpus 臺華平行新聞語料庫漢字臺羅版(Sih4, 2015)。具體系統架構如圖 31 所示，使用者輸入中文後，C2T 將中文轉換成台羅拼音（依據『臺灣閩南語羅馬字拼音方案』(教育部，2008)，台羅拼音作為文本輸入 Tacotron2 轉為頻譜，最後透過 WaveGlow 將頻譜即時的轉為波形並合成語音。

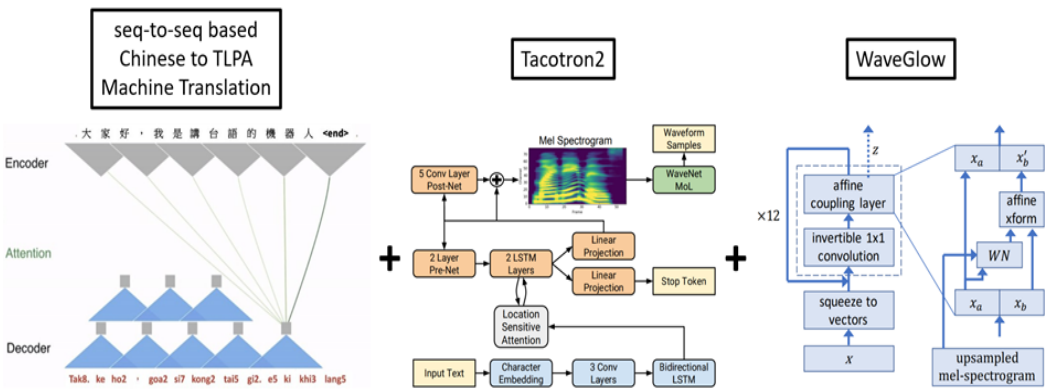


圖 31. 中文文字轉台語語音合成系統

[Figure 31. The architecture of the Chinese text to Taiwanese speech synthesis system]

4.1.1 中文文字轉台羅拼音機器翻譯 (Chinese to Taiwanese translation)

在訓練此機器翻譯所使用的中文對應台語拼音平行語料方面，使用了開源之 iCorpus 臺華平行新聞語料庫漢字臺羅版(Sih4, 2015)。並以上述語料為基礎，自行人工去除專有名詞，地名及人名等英文，以及少部分中文錯誤的地方，使訓練文本更為正確，經整理後得到 60323 句原始平行語料。而原始的 iCorpus 平行語料並不包含標點符號，為使成品之機器翻譯能夠看懂基本的標點符號，進而對語料進行以下處理。

做法為將文本複製六份後，每一份負責加入一種標點符號，分別為逗號、句號、驚嘆號、問號、冒號與分號六種標點符號。加入的方式為，每一句平行語料的結尾加入標點符號後，隨後接上一句隨機分配的平行語料，這種將標點符號加在前後為不一樣句子的方式，能讓標點符號的存在較為自然，使標點符號能正確地被訓練進去。最後將平行語料中文的部分以空格隔開每一個中文字，台羅拼音的部分連字號也改成以空格表示，讓平行語料以 phone 對 phone 的方式對應。以上全部完成處理後的部分範例如表 5 與表 6 所示。

表 5. iCorpus 平行語料中文部分範例
[Table 5. Examples of the Chinese transcriptions in the iCorpus corpus]

駐美特派員曹郁芬華府報導，出海捕獲一條將近三百公斤
蔡英文也說到時薪應該調整？每天得花四個小時在訓練上面
現在正好芒果盛產的季節：北部早晚低溫可能只有二十度上下

表 6. iCorpus 平行語料台羅拼音部分範例
[Table 6. Examples of the Taiwanese transcriptions in the iCorpus corpus]

tsu3 bi2 tik8 phai3 uan5 tso5 hiok4 hun1 hua5 hu2 po3 to7 , tshut4 hai2 liah8 tioh8 tsit8 tiau5 ua2 beh4 sann1 pah4 kong1 kin1
tshua3 ing1 bun5 ma7 tam5 kau3 si5 sin1 ing1 kai1 tiau5 tsing2 ? tak8 kang1 khai1 iong7 si3 tiam2 tsing1 ti7 hun3 lian7 bin7 ting2
tsit4 ma2 tu2 ho2 suainn7 a2 tua7 tshut4 e5 kui3 tsiat4 : pak4 poo7 tsa2 am3 ke7 un1 kho2 ling5 tsi2 u7 ji7 tsap8 too7 ting2 e7

準備好平行語料後，我們參考網路上開源的 fairseq 機器翻譯演算法(Hsu, 2021)，其基於 sequence-to-sequence 架構，網路如圖 32 所示，包括一 encoder 前端與一 decoder 後端。前端 encoder 負責接收輸入中文文字序列，分析其語意並擷取出文脈資訊向量。後端 decoder 在文脈資訊向量之間加入 attention 之機制與 convolutional neural network 之訓練模型下每個 encoder 權重，利用中文對應台語拼音平行語料庫進行訓練，以此得到最佳的轉譯台羅拼音序列。

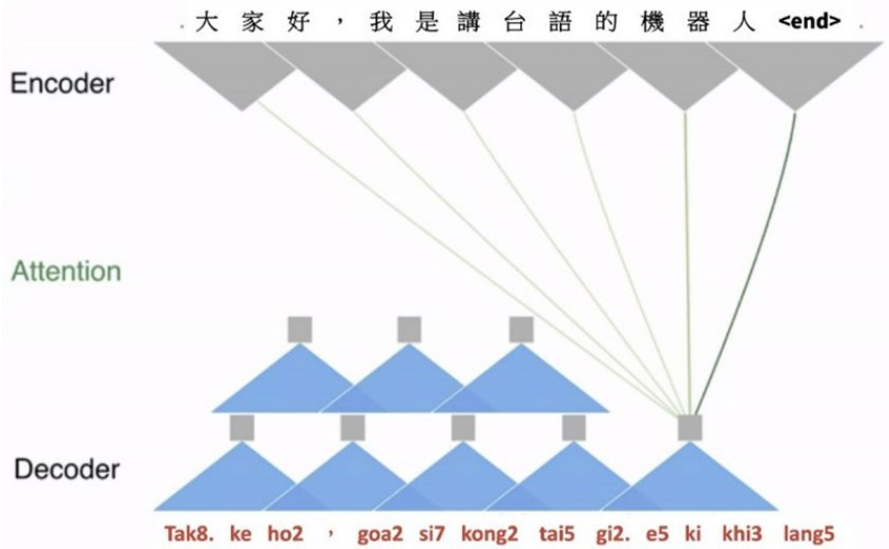


圖 32. 中文轉台羅拼音機器翻譯訓練架構
[Figure 32. The architecture of the Chinese to Taiwanese machine translation system]

4.1.2 Tacotron2+WaveGlow

我們使用 TAT-TTS-M1 約 10.4 小時的台語語料，以 22050 赫茲的音檔取樣頻率，以及原始的「台羅數字調」當作文本，進行 Tacotron2 的訓練(Valle, 2020)。WaveGlow 聲碼器僅負責將頻譜合成語音，訓練時僅使用大量的音檔，不須配合文本，意即使用的語料語言與前端 Tacotron2 並無衝突。因此 WaveGlow 的部份，使用實驗室已經事先用數量與豐富度較多的英文語料 LJ Speech 訓練出的 WaveGlow 模型(Valle, 2020)，不須使用台語語料重新訓練。

另外我們也使用了校正好變調以及韻律符號的 TAT-TTS-M1 語料，裡面新增的"變調韻律 TLPI"作為訓練文本，將新增的兩種符號，加號"+"以及底線"_"新增進去訓練模型時考慮的特殊符號，訓練了一版考慮變調以及新增兩種韻律符號的 Tacotron2，作為後續實驗的比較。

4.1.3 雛型系統展示網頁 (Prototype System Demonstration)

我們將結合了中文轉台羅拼音機器翻譯的單人台語語音合成系統，做成了一展示網頁[5]，如圖 33 所示。使用者輸入中文文字後，按下合成按鈕就能撥放對應的台語語音，並能一併顯示出翻譯過後的台羅拼音供使用者查詢。且另外設計了可輸入台羅拼音的欄位，讓擁有相關台羅知識的使用者可以鍵入不同的發音並合成想要的台語語音。

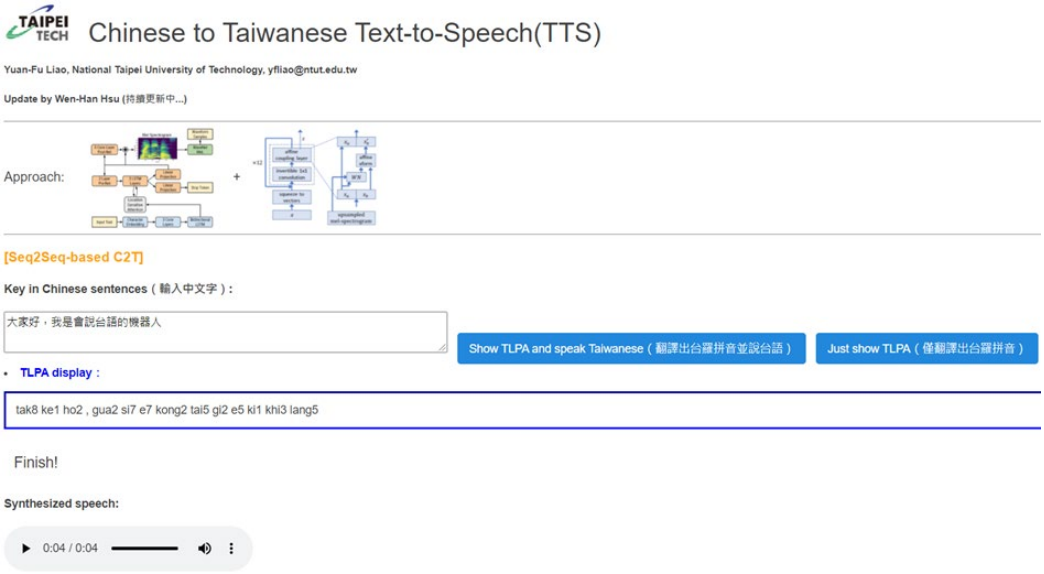


圖 33. 中文文字轉台語語音合成系統展示網頁
[Figure 33. Demo website of the Chinese to Taiwanese machine translation]

[5] NTUT's Chinese to Taiwanese Text-to-Speech(TTS), http://tts001.iptcloud.net:8801/

5. 結合語音辨認及合成模組之多語者台語語音轉換系統 (Multi-Speaker Voice Conversion System based on Cascade ASR and TTS framework)

有了台語語音合成相關經驗的累積後，我們開始探究適合的台語語音轉換進行方法，目標做出一個初版可行的台語語音轉換系統。

5.1 同語言之台語對台語語音轉換系統建置 (Intra-Lingual Voice Conversion)

我們以 VCC 2020 中出現的 Cascade ASR and TTS 方法(Huang *et al*., 2020)為 baseline，目標建置台語對台語語音轉換系統。具體系統架構如圖 34 所示，將來源語者的音檔，以台語語音辨認器轉成台羅拼音後，輸入已經預先以目標語者的語料微調過的台語多語者語音合成器，合成出符合來源語者文本以及目標語者音色的語音轉換音檔。

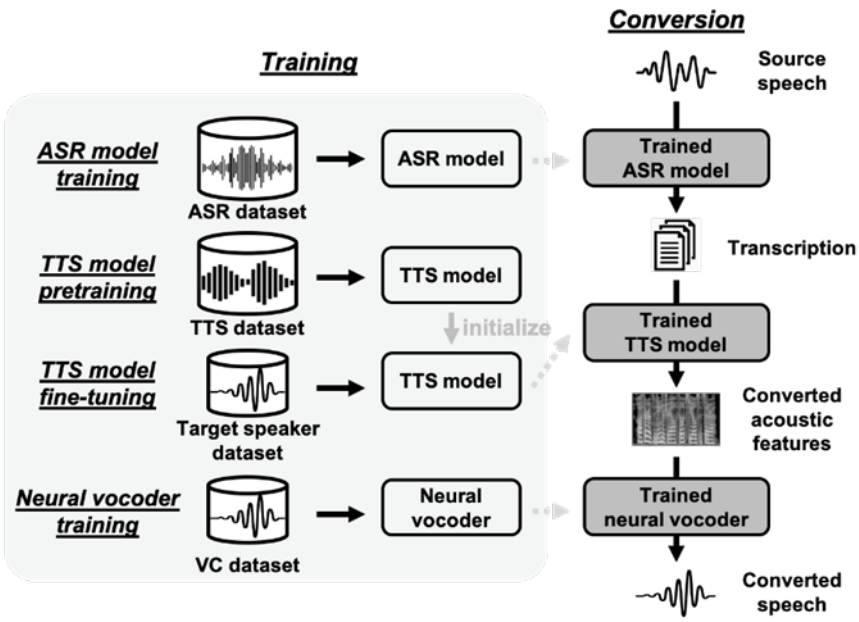


圖 34. 結合語音辨認及合成模組之台語轉台語語音轉換系統架構 (Huang et al., 2020)
[Figure 34. The architecture of the Taiwanese voice conversion system]

此方法需以三種預訓練模型為基礎，分別是(1)X-Vectors，(2)Transformer-based ASR model 以及(3)Multi-speaker Transformer-TTS model。在語音轉換的領域，因為涉及到語者辨識的技術，因此訓練語料的語者數量是越豐富越好。我們訓練三個預訓練模型使用的台語語料分別使用了(1)TAT-TTS- M1~2 與 TAT-TTS-F1~2 四位語者，共 2 男 2 女，每人約有 10 小時語料，總長度約為 40.6 小時，以及(2) TAT-Vol1~2 語料庫裡面的 200 位全部語者，共 91 男 109 女，每人約有半小時語料，總長度約為 104.4 小時。

在訓練文本的選擇上，我們一樣使用"台羅數字調"，並將文本的連字號取消，以較單純的字對字去訓練，部分範例如圖 35 所示，這樣一方面可以降低訓練的難度，也可以跟後面跨語言任務使用的華語語料做相同的對應。

台文意思1	阿明的護照號碼是八五四一二三六五五
訓練文本1	A1 bing5 e5 hoo7 tsiau3 ho7 be2 si7 pat4 ngoo2 su3 it4 ji7 sam1 liok8 ngoo2 ngoo2
台文意思2	啊我明明就共講台語
訓練文本2	ah4 gua2 bing5 bing5 toh8 ka7 kong2 tai5 gi2
台文意思3	就隨有買一枝鍊仔鋸來做工課的心念
訓練文本3	tioh8 sui5 u7 be2 tsit8 ki1 lian7 a2 ku3 lai5 tso3 khang1 khue3 e5 sim1 liam7

圖 35. 台語語料訓練文本部分範例
[Figure 35. A typical Example of the Taiwanese speech transcription data]

而後端聲碼器則沿用原本的 Parallel WaveGAN（PWG），原因同單人台語語音合成的 WaveGlow，聲碼器的部分不需要重新訓練。

5.1.1 語者向量編碼器 (Speaker Embedding)

使用了較為基礎的 X-Vectors 方法(Snyder *et al*., 2018)，如圖 36 所示。把輸入的語音截成多段，將每一小段語音信號輸出的特徵算一個 mean 以及 variance 並且 concat 起來，輸入 DNN 後來判斷這一小段語音是哪位語者的語者資訊，最後各小段語音的平均結果即為 speaker embedding。將 204 人的台語音檔以 train set 194 人，test set 10 人的設定進行語者向量編碼器的訓練(esdeboer, 2020)。

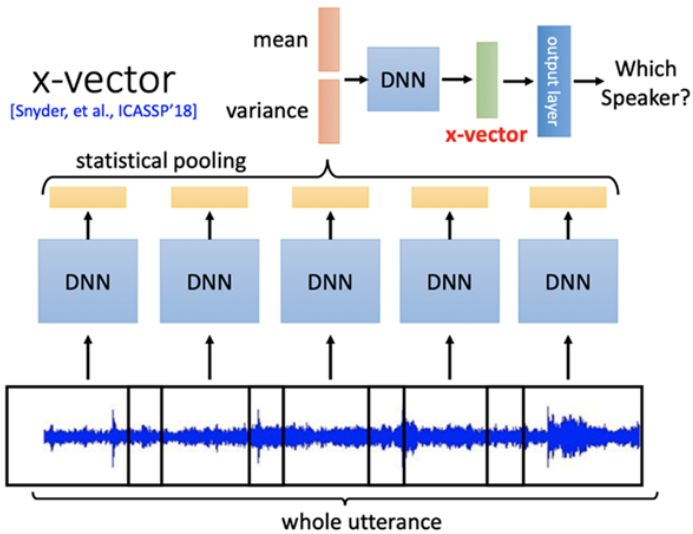


圖 36. X-Vectors 架構(Snyder et al., 2018)
[Figure 36. The architecture of X-Vector speaker embedding encoder]

5.1.2 台語語音辨認器 (Taiwanese Speech Recogntizer)

以端對端 ASR 架構(Dong *et al.*, 2018)，如圖 37 所示，以台語音檔和對應的文本，和上述已經訓練好的 X-Vectors，進行 Transformer-based ASR model 的訓練(shirayu, 2021)。在資料集分配上使用跟上述語者向量編碼器一樣 194 人的 train set，並將相同 10 人的 test set 分出 5 人給 dev set。

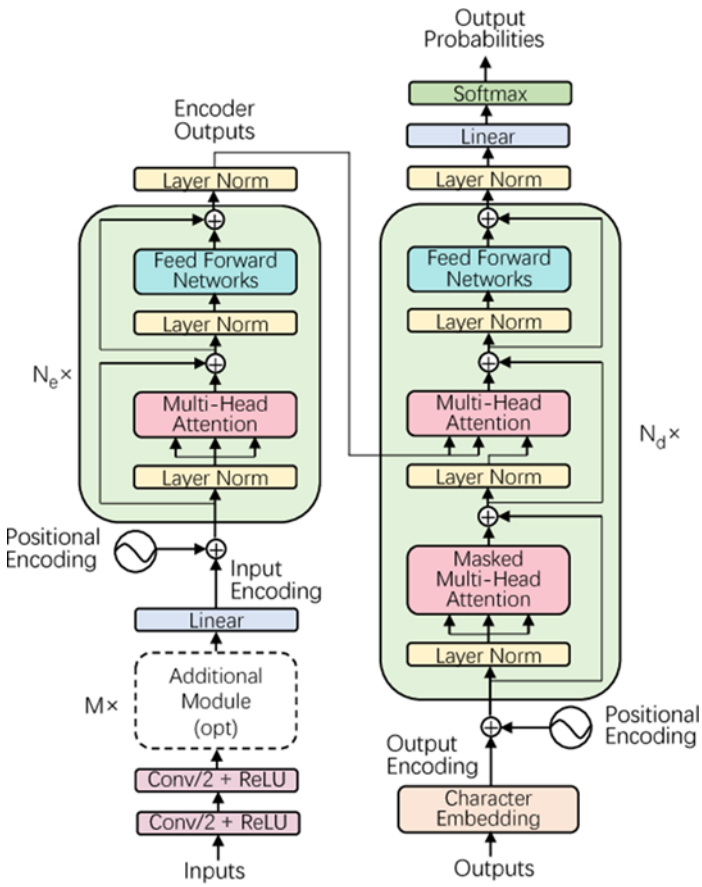


圖 37. E2E-ASR 架構(Dong et al., 2018)
[Figure 37. The architecture of Taiwanese speech recognizer]

5.1.3 多語者語音合成器 (Multi-speaker TTS)

多語者語音合成器的部分採用類似如圖 38 的架構(Chen *et al.*, 2020)，以台語音檔和對應的文本，和上述已經訓練好的 X-Vectors，進行 Multi-speaker Transformer-TTS model 的訓練(shirayu, 2021)。資料集分配上則跟上述 ASR 完全一致，train：dev：test = 194：5：5。

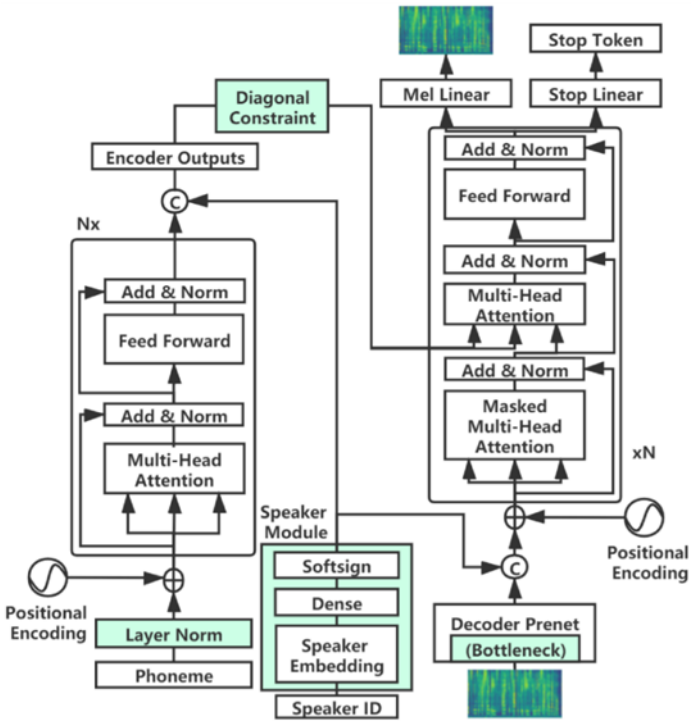


圖 38. E2E-TTS 架構(Chen et al., 2020)
[Figure 38. The architecture of multi-speaker Taiwanese speech synthesis]

5.2 跨語言之華語對台語語音轉換系統建置 (Cross-Lingual Voice Conversion)

在 VCC 2020 中，比賽又分為 Task1：同語言任務以及 Task2：跨語言任務，如圖 39 所示。而在跨語言任務中，比賽中設置目標語者的語言分別為華語，德語和法語，並需要利用語音轉換使用目標語者的語音特徵說出英文句子。

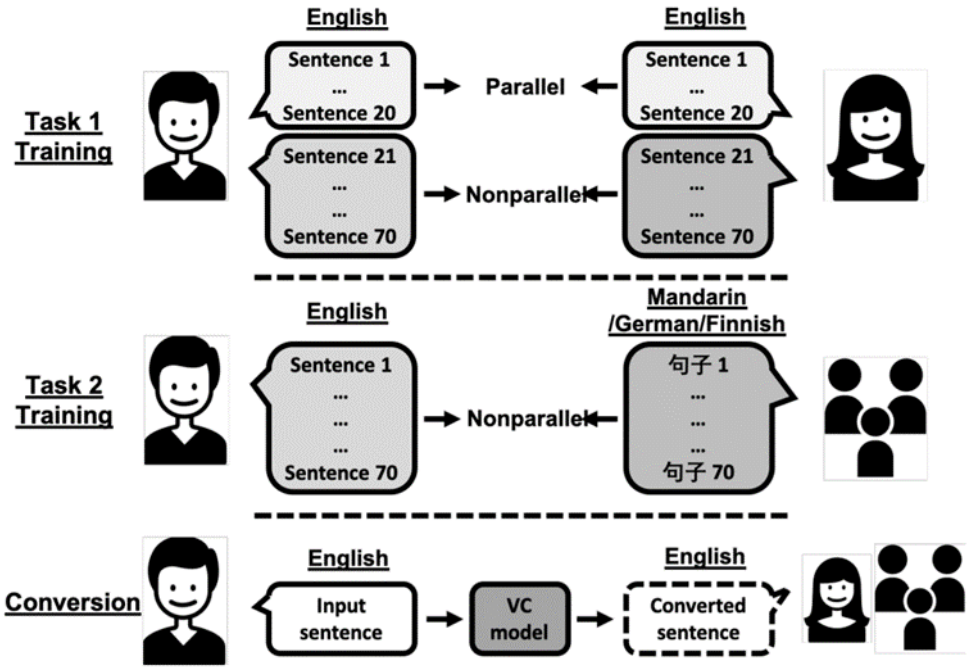


圖 39. VCC 2020 兩種任務介紹(Zhao et al., 2020)

[Figure 39. The Intra- and cross-lingual voice conversion tasks in VCC 2020]

Cascade ASR and TTS 方法因為架構的特性，也可以做到跨語言的語音轉換，我們將目標語者的語言設定為華語，目標建置一個華語對台語語音轉換系統。跟同語言的台語對台語語音轉換系統相比，跨語言最大的差異在於 TTS 端的部分，如圖 40 所示。目標語者的語言變成華語後，要做到跨語言的語音轉換就必須重新以華語和台語 2 種語言去訓練雙語言的多語者語音合成器。

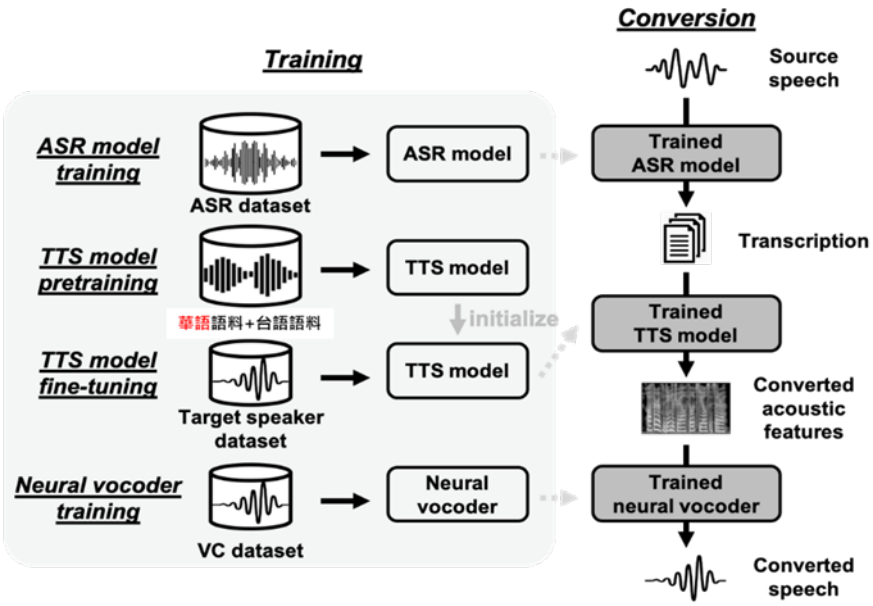


圖 40. 結合語音辨認及合成模組之華語轉台語語音轉換系統架構 (Snyder et al., 2018)

[Figure 40. The architecture of the cross-lingual voice conversion framework]

我們在同語言的部分，也使用了原先由 librispeech 英文語料所訓練的 X-Vectors，並以英文的 X-Vectors 重新對台語的多語者語音合成器進行訓練後，發現語音轉換的音檔聽感其實差異不大，因此在跨語言的部分，我們在 X-Vectors 的部分沿用了原先英文語料訓練的模型。而 ASR 的部分沿用同語言任務已經訓練好的模型。聲碼器也依舊使用原本訓練好的 PWG。

此處應用在跨語言語音轉換的華語語料，為 VCC 2020 中，Cascade ASR and TTS 方法在華語對英語的跨語言語音轉換中，訓練雙語言多語者語音合成器(Kamo, 2021)時使用的語料。名稱為 csmsc，共有 10000 筆音檔，語料總長度約 11.86 小時，為一大陸腔女生所錄製的華語語料。

為了使華語文本和台語文本一致，我們使用華語語料中跟台語文本一樣使用子音加母音的文本形式作為訓練文本，部分範例如圖 41 所示。

中文意思1	赵荻约曹云腾去鬼屋
訓練文本1	zhao4 di2 yue1 cao2 yun2 teng2 qu4 gui3 wu1
中文意思2	高压铁塔下的低矮棚屋
訓練文本2	gao1 ya1 tie2 ta3 xia4 de5 di1 ai3 peng2 wu1
中文意思3	用不用我替你捂着嘴
訓練文本3	yong4 bu2 yong4 wo3 ti4 ni2 wu3 zhe5 zui3

圖 41. 華語語料訓練文本部分範例
[Figure 41. A typical example of the Mandarin speech transcription data]

因為使用了兩種語言作為訓練文本，為了區分文本的語言屬性，需要在訓練文本前加上語言碼，如英文預設為<en_US>，華文預設為<zh_ZH>，我們在台語方面則訂為<tw_TW>，加上語言碼後的訓練文本部分範例如圖 42 所示。

```
TS_TSM0020_99 <tw_TW> hiong3 kok4 tse3 tshut4 siann1 ting7 gi7
csmsc_000001 <zh_ZH> ka2 er2 pu3 pei2 wai4 sun1 wan2 hua2 ti1
```

圖 42. 跨語言任務語言碼部分範例
[Figure 42. A typical example of the language code embedding for cross-lingual voice conversion task]

在使用 csmsc 華語語料以及同語言任務使用的台語語料，以及原先英文語料訓練的 X-Vectors，混合訓練好雙語言多語者語音合成器(Kamo, 2021)後，我們也成功建置了一個跨語言之華語對台語語音轉換系統。

6. 實驗 (Experimental Results)

6.1 單人台語語音合成系統實驗 (Single-Speaker Taiwanese Speech Synthesis)

我們使用變調更正以及增加韻律符號的新文本，重新訓練台語語音合成的新模型，並以原本使用沒有變調以及沒有考慮兩種新的韻律符號的舊文本訓練的原模型做比較，簡單設計了以下實驗。分別準備 10 句中文句子原始還沒校正過的台羅拼音，作為原模型的輸入文本合成語音，然後請校正人員以相同規則為這 10 句台羅拼音進行變調更正以及韻律符號的添加，作為新模型輸入的文本並合成語音。10 句實驗句子（S1~S10）校正前後的台羅拼音比較如圖 43 所示，標記紅色的地方為變調更正和韻律符號不同的地方。

(S1)大家好，我是會說台語的機器人。	
原模型	tak8-ke1 ho2,gua2 si7 e7 kong2 tai5-gi2 e5 ki1-khi3-lang5.
新模型	tak4-ke7+ho2,gua1_si3_e3+kong1+tai7-gi2+e7_ki7-khi2-lang5.
(S2)今天一早起來，天氣就非常炎熱。	
原模型	kin1-a2-jit8 thau3-tsa2 khi2-lai5,thinn1-khi3 to1 hui1 siong5 pik4-juah8.
新模型	kin7-a1-jit8_thau2-tsa2_khi2--lai3,thinn7-khi3_to3_hui7+siong5_pik8-juah8.
(S3)一千兩百三十四萬五千六百七十八點零九美元。	
原模型	tsit8-tshing1 nng7-pah4 sann1-tsap8-si3-ban7 goo7-tshing1 lak8-pah4 tshit4-tsap8-peh4-tiam2-khong3-kau2 bi2-kim1.
新模型	tsit4-tshing1+nng3-pah8+sann7-tsap4-si2-ban7_goo3-tshing1+lak4-pah8+tshit8-tsap4-peh8-tiam1-khong2-kau1+bi1-kim1.
(S4)現在為您報導晚間新聞。	
原模型	tsit4-ma2 ui7 lin2 po3-to7 am3-si5 sin1-bun5.
新模型	tsit8-ma2_ui3_lin1_po2-to3_am2-si5+sin7-bun5.
(S5)武漢肺炎的出現，讓全世界的人都開始戴口罩。	
原模型	bu2-han3 hi3-iam7 e5 tshut4-hian7,hoo7 tsuan5-se3-kai3 e5 lang5 long2 khai1-si2 ti3 tshui3-am1.
新模型	bu1-han3+hi2-iam7_e7_tshut8-hian7,hoo3_tsuan7-se2-kai3+e7_lang5_long1_khai7-si1_ti2_tshui2-am1.
(S6)昨天地震時，我們家的花瓶掉下來摔破了。	
原模型	tsoh8-jit8 te7-tang7 si5,gun2-tau1 e5 hue1-kan1 lak4-loh8-lai5 siak4-phua3-ah4.
新模型	tsoh8--jit8_te3-tang7+si5,gun1-tau1+e3_hue7-kan1_lak4--loh4-lai3_siak8-phua3--ah4.
(S7)失敗為成功之母。	
原模型	sit4-pai7 ui5 sing5-kong1 tsi1 bo2.
新模型	sit8-pai7_ui7_sing7-kong1+tsi7+bo2.
(S8)歡迎光臨，請問有幾位?	
原模型	huan1-ging5 kng1-lim5, tshiann2-bun7 u7 kui2-ui7?
新模型	huan7-ging5+kong7-lim5, tshiann1-mng7_u3_kui1-ui7?
(S9)有颱風從太平洋來的時候，中央山脈常常幫台灣的西部擋去很多災情。	
原模型	u7 hong1-thai1 tui3 thai3-ping5-iunn5 lai5 e5 si5-tsun7, tiong1-iang1-suann1-meh8 tiann7-tiann7 pang1 tai5-uan5 e5 se1-poo7 tong3-khi3 tsin1-tsue7 tsai1-tsing5.
新模型	u3_hong7-thai1_tui2+thai2-ping7-iunn5+lai5_e7_si7-tsun7, tiong7-iong7-suann7-meh8_tiann3-tiann3_pang7_tai7-uan5+e7_se7-poo7_tong2-khi2_tsin7-tse3_tsai7-tsing5.
(S10)龜笑鱉無尾，鱉笑龜粗皮。	
原模型	ku1 tshio3 pih4 bo5 bue2, pih4 tshio3 ku1 tshoo1-phue5.
新模型	ku1_tshio2_pih4_bo7+bue2, pih4_tshio2_ku1_tshoo7-phue5.

圖 43. 校正前後的台羅拼音比較

[Figure 43. Comparison of Taiwanese transcriptions before and after tone sandi annotation]

我們請聽者對原模型以及新模型各 10 個句子合成出的音檔進行自然度的評分，最後收集到 27 位聽者的評分，原模型以及新模型的實驗結果如圖 44 及圖 45 所示。

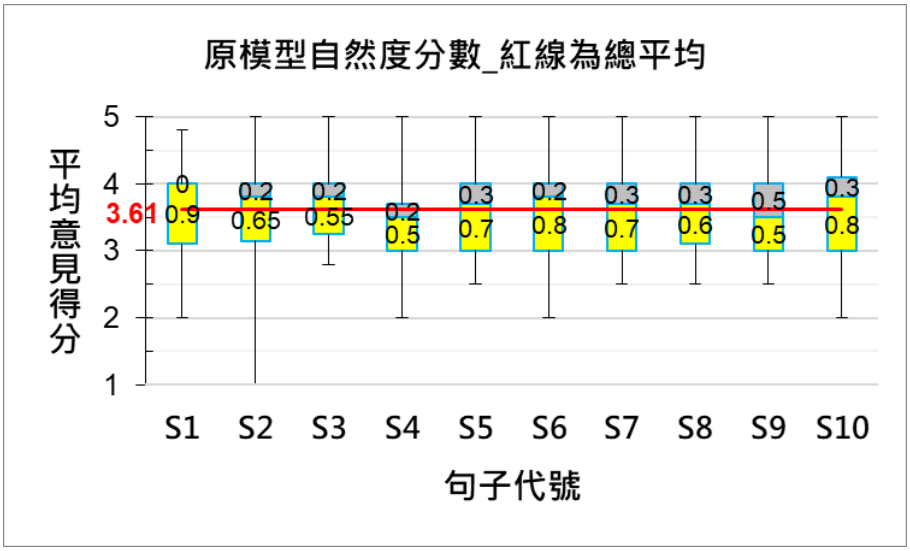


圖 44. 原模型自然度分數實驗結果盒鬚圖

[Figure 44. The box-and-whisker plot of naturalness scores of the baseline model]

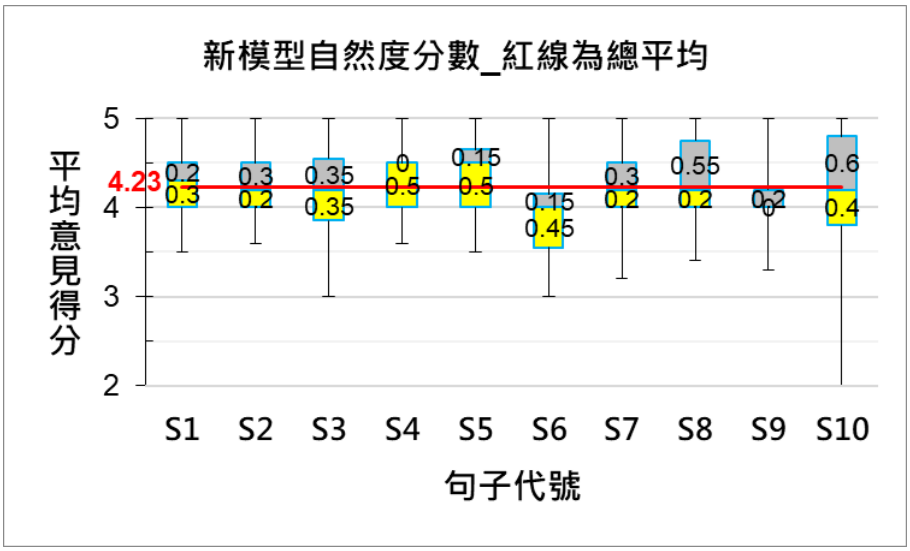


圖 45. 新模型自然度分數實驗結果盒鬚圖

[Figure 45. The box-and-whisker plot of naturalness scores of the improved model]

由實驗結果得知，輸入同樣韻律規則的文本合成的語音，比起只有以連字號和空格作為韻律符號的文本合成的語音，的確在聽感上有更加接近真人在講同一句話時該有的順暢度，較少會在奇怪的地方停頓。因此下一個目標，即為使用此新語料訓練能產出同樣韻律規則的中文轉台語機器翻譯，以完成整個新台語語音合成系統的建置。

6.2 結合語音辨認及合成模組之台語多語者語音轉換系統實驗 (Multi-Speaker Voice Conversion)

以下將介紹使用 TAT-ASR 以及 TAT-TTS 台文語料庫總共 204 名語者，共約 145 小時的台語語料訓練的模型成果，以及結合語音辨認及合成模組之台語語音轉換系統的相關實驗。

6.2.1 語者向量編碼器EER結果 (Performance on Speaker Recognition)

將 204 人的台語音檔以 train set 194 人，test set 10 人的設定進行語者向量編碼器的訓練 (esdeboer, 2020)。並以 test set 的台語語料製作 EER 的測試檔案，將 10 位測試語者取出 1 位，以相同語者，非相同語者的設計平均的跟另外 9 位測試語者做語者辨識的測試，並且以此方法將 10 位測試語者全部測試完畢，測試檔案部分範例如圖 46 所示。最後用 194 人台語語料訓練出的語者向量編碼器，得出了 EER 為 5.506%的測試結果，minDCF 在 p-target = 0.01 的情況下為 0.7101，在 p-target = 0.001 的情況下為 0.8318。

```
IU_IUF0023_1 IU_IUF0023_22 target
IU_IUF0023_1 IU_IUM0017_43 nontarget
IU_IUF0023_1 IU_IUF0023_184 target
IU_IUF0023_1 KH_KHF0030_190 nontarget
IU_IUF0023_1 IU_IUF0023_214 target
IU_IUF0023_1 KH_KHM0024_38 nontarget
IU_IUF0023_1 IU_IUF0023_2 target
IU_IUF0023_1 KK_KKF0015_105 nontarget
IU_IUF0023_1 IU_IUF0023_169 target
IU_IUF0023_1 KK_KKM0015_81 nontarget
IU_IUF0023_1 IU_IUF0023_19 target
IU_IUF0023_1 TA_TAF0020_248 nontarget
IU_IUF0023_1 IU_IUF0023_9 target
IU_IUF0023_1 TA_TAM0020_55 nontarget
IU_IUF0023_1 IU_IUF0023_112 target
IU_IUF0023_1 TH_THF0022_317 nontarget
IU_IUF0023_1 IU_IUF0023_83 target
IU_IUF0023_1 TH_THM0018_44 nontarget
IU_IUF0023_10 IU_IUF0023_14 target
IU_IUF0023_10 IU_IUM0017_115 nontarget
IU_IUF0023_10 IU_IUF0023_16 target
IU_IUF0023_10 KH_KHF0030_234 nontarget
```

圖 46. 語者向量編碼器測試檔案部分範例

[Figure 46. A typical example of the speaker tranacriptions for speaker recognition]

6.2.2 台語語音辨認器錯誤率 (Performance on Taiwanese Speech Recognition)

在資料集分配上使用跟上述語者向量編碼器一樣 194 人的 train set，並將相同 10 人的 test set 分出 5 人給 dev set，最後訓練出來的語音辨認器，錯誤率約為 2.9%，詳情如圖 47 所示。訓練過程相關 loss 資訊如圖 48 所示。

exp/train_pytorch_train_specaug/decode_test_model.val5.avg.best_decode_lm/hyp.wrd.trn

SPKR	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
iu_iuf0023	233	2419	97.9	1.9	0.2	0.0	2.2	17.2
iu_ium0017	233	2415	97.0	2.5	0.5	0.1	3.1	26.2
kh_khf0030	282	2652	98.2	1.7	0.1	0.2	1.9	16.0
kh_khm0024	262	2506	97.0	3.0	0.0	0.2	3.2	23.7
kk_kkf0015	264	2688	96.0	3.7	0.3	0.3	4.3	28.8
Sum/Avg	1274	12680	97.2	2.6	0.2	0.1	2.9	22.3
Mean	254.8	2536.0	97.2	2.6	0.2	0.1	2.9	22.4
S.D.	21.4	128.2	0.9	0.8	0.2	0.1	0.9	5.6
Median	262.0	2506.0	97.0	2.5	0.2	0.2	3.1	23.7

exp/train_pytorch_train_specaug/decode_test_model.val5.avg.best_decode_lm/hyp.wrd.trn

SPKR	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
iu_iuf0023	233	2419	2367	47	5	1	53	40
iu_ium0017	233	2415	2343	61	11	2	74	61
kh_khf0030	282	2652	2605	45	2	4	51	45
kh_khm0024	262	2506	2431	74	1	4	79	62
kk_kkf0015	264	2688	2580	100	8	7	115	76
Sum	1274	12680	12326	327	27	18	372	284
Mean	254.8	2536.0	2465.2	65.4	5.4	3.6	74.4	56.8
S.D.	21.4	128.2	120.9	22.6	4.2	2.3	25.9	14.4
Median	262.0	2506.0	2431.0	61.0	5.0	4.0	74.0	61.0

圖 47.台語語音辨認器訓練結果

[Figure 47. Experimental results of the Taiwanese speech recognitizer]

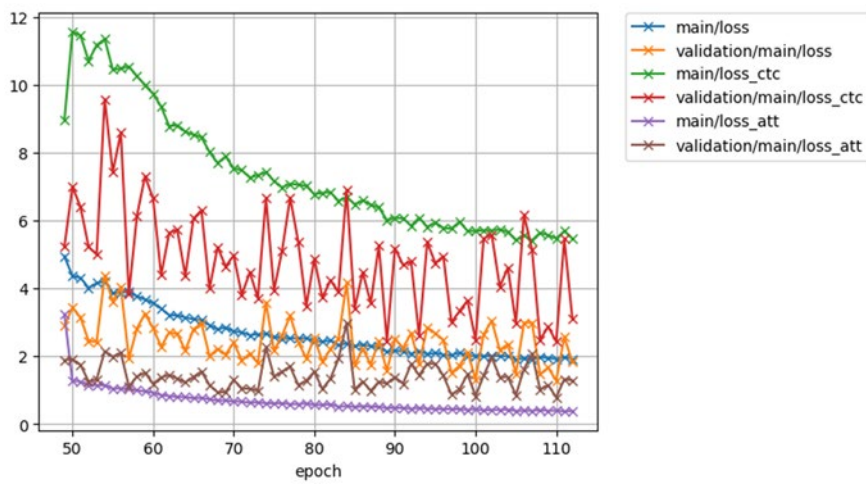


圖 48.台語語音辨認器訓練過程 loss

[Figure 48. The learning curves of the Taiwanese speech synthesis system]

6.2.3 台語語音轉換系統實驗 (Taiwanese Voice Conversion)

最後，我們針對使用台語語料做出的結合語音辨認及合成模組之台語語音轉換系統，設計了轉換音檔「自然度」和「相似度」的主觀評測實驗。此實驗最終收集了 29 位評分者的評分結果，評分人分別有來自中華電信研究院的語音相關專業人士，長問科技的語音相關專業人士，台語老師以及同實驗室的研究生。

6.2.3.1 實驗方法 (Experimental Settings)

實驗問卷分成(1)台語對台語語音轉換和(2)華語對台語語音轉換兩部分，第一部分有 4 位目標語者的合成音檔，第二部分有 3 位目標語者的合成音檔。每 1 位目標語者有數個需要進行評分的合成音檔，和 1 個原始音檔以供對照。

(1) 台語對台語語音轉換的部分，12 個合成音檔分別為

(1-1)4 個使用約 10 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

(1-2)4 個使用約 3 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

(1-3)4 個使用約 30 秒 fine-tuning 語料量做出的語音轉換合成音檔

(2) 華語對台語語音轉換的部分，8 個合成音檔分別為

(2-1)4 個使用約 6 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

(2-2)4 個使用約 3 分鐘 fine-tuning 語料量做出的語音轉換合成音檔

評分者聽完合成音檔後，依據主觀感受對每個合成音檔評兩種分數。

(一) 自然度分數

根據聽到的「自然度」進行 1.0 到 5.0 的評分，最多評分到小數第一位

最低分 1.0 分 為完全不像真人講話的聲音

最高分 5.0 分 為完全像是真人講話的聲音

(二)相似度分數

根據聽到的「相似度」進行 1.0 到 5.0 的評分，最多評分到小數第一位

最低分 1.0 分 <原始音檔>和<合成音檔>完全不像同一個人講話的聲音

最高分 5.0 分 <原始音檔>和<合成音檔>完全像同一個人講話的聲音

6.2.3.2 同語言任務實驗結果 (Intra-Lingual Voice Conversion)

台語對台語語音轉換音檔的 MOS 分數盒鬚圖（Box-Plot）

(1-1)10 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 49 和圖 50 所示

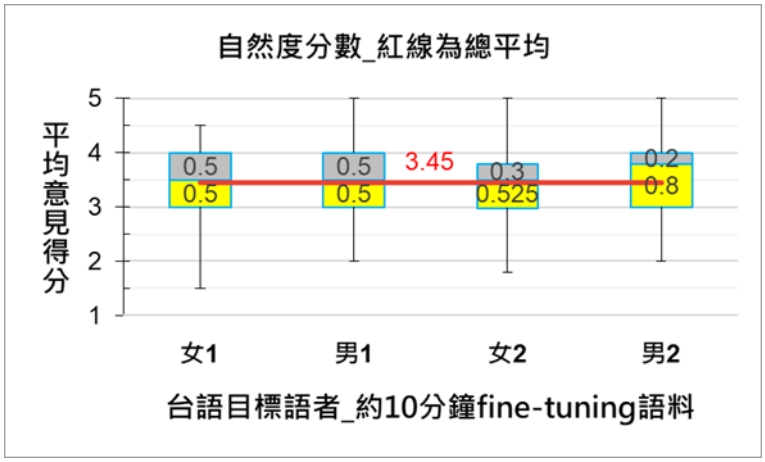


圖 49. 同語言任務使用 10 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 49. The box-and-whisker plot of naturalness scores using 10-minute fine-tuning data for intra-lingual voice conversion]

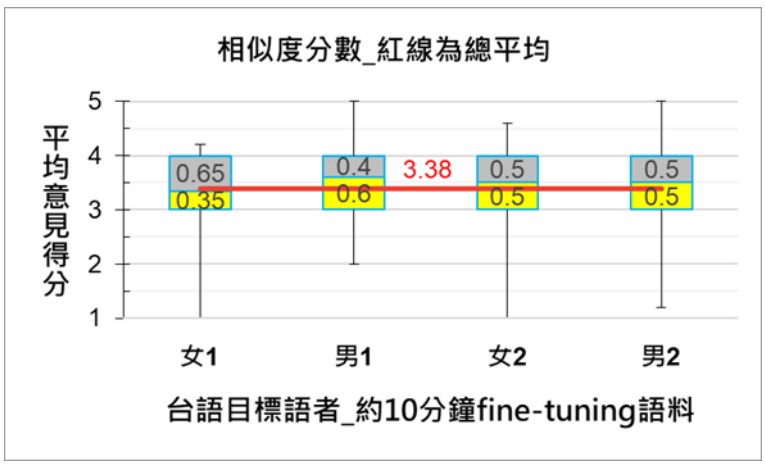


圖 50. 同語言任務使用 10 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 50. The box-and-whisker plot of similarity scores using 10-minute fine-tuning data for intra-lingual voice conversion]

(1-2)3 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 51 和圖 52 所示

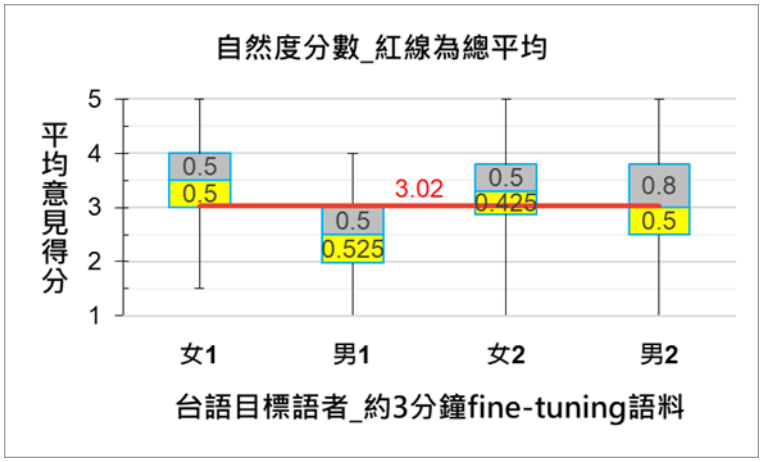


圖 51. 同語言任務使用 3 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 51. The box-and-whisker plot of naturalness scores using 3-minute fine-tuning data for intra-lingual voice conversion]

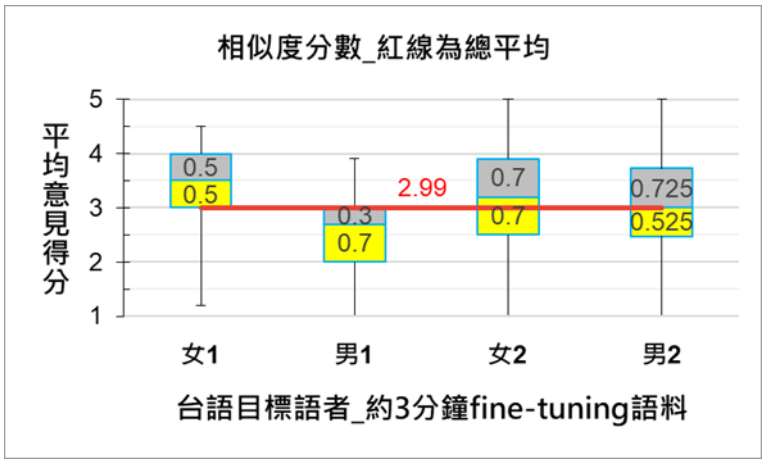


圖 52. 同語言任務使用 3 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 52. The box-and-whisker plot of similarity scores using 3-minute fine-tuning data for intra-lingual voice conversion]

(1-3)30 秒 fine-tuning 語料量，自然度分數和相似度分數如圖 53 和圖 54 所示

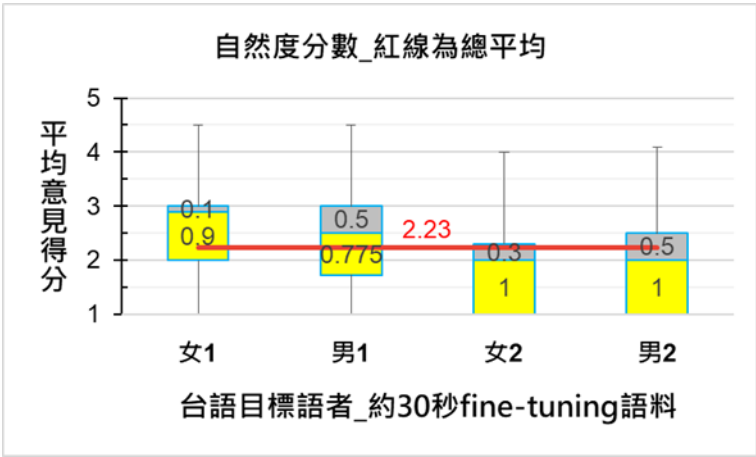


圖 53. 同語言任務使用 30 秒 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 53. The box-and-whisker plot of naturalness scores using 30-second fine-tuning data for intra-lingual voice conversion]

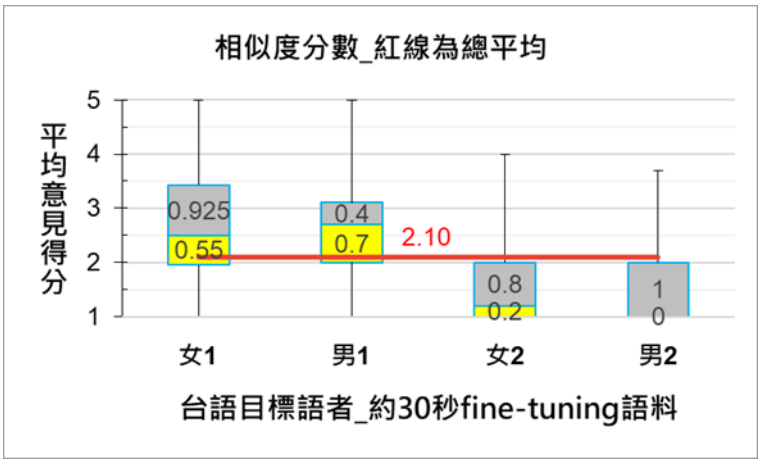


圖 54. 同語言任務使用 30 秒 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 54. The box-and-whisker plot of similarity scores using a 30-second fine-tuning data for intra-lingual voice conversion]

6.2.3.3 跨語言任務實驗結果 (Cross-Lingual Voice Conversion)

華語對台語語音轉換音檔的 MOS 分數盒鬚圖（Box-Plot）

(2-1)6 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 55 和圖 56 所示

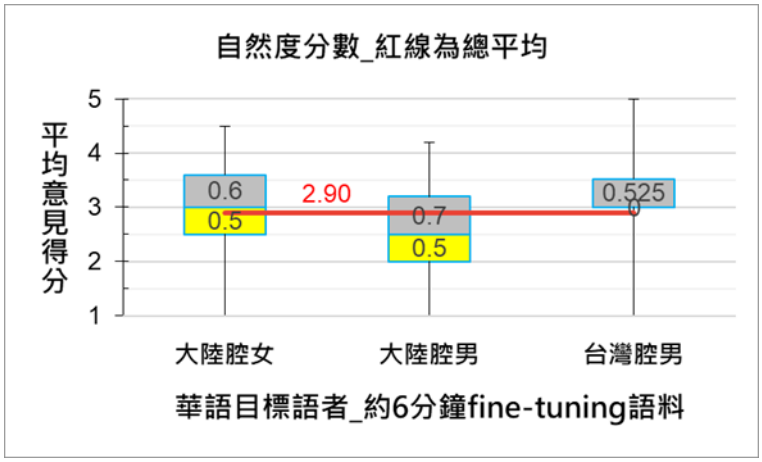


圖 55. 跨語言任務使用 6 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 55. The box-and-whisker plot of naturalness scores using 6-minute fine-tuning data for cross-lingual voice conversion]

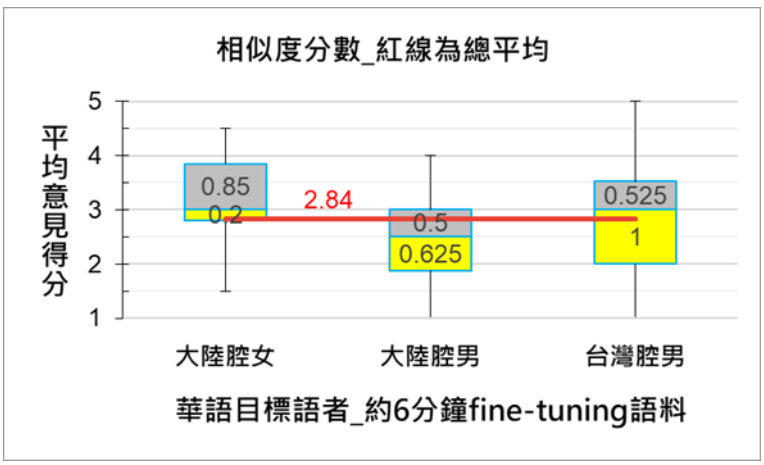


圖 56. 跨語言任務使用 6 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 56. The box-and-whisker plot of similarity scores using 6-minute fine-tuning data for cross-lingual voice conversion]

(2-2)3 分鐘 fine-tuning 語料量，自然度分數和相似度分數如圖 57 和圖 58 所示

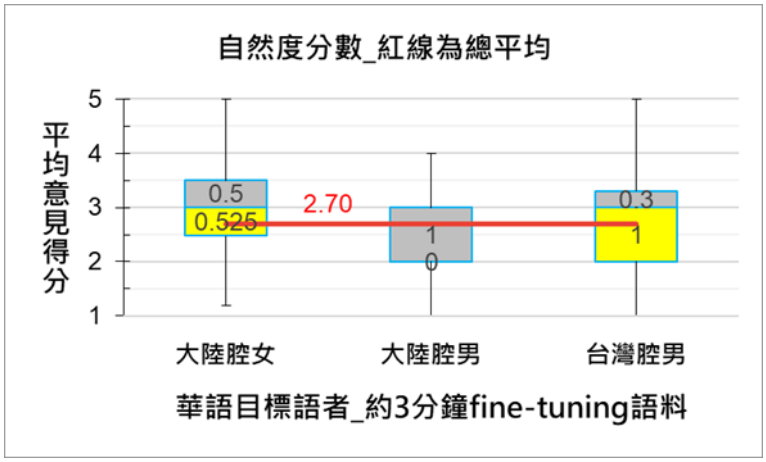


圖 57. 跨語言任務使用 3 分鐘 fine-tuning 語料量之自然度分數盒鬚圖
[Figure 57. The box-and-whisker plot of naturalness scores using 3-minute fine-tuning data for cross-lingual voice conversion]

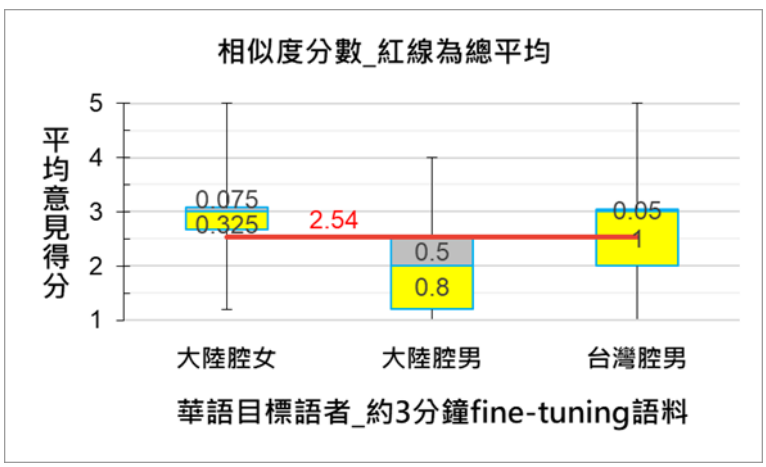


圖 58. 跨語言任務使用 3 分鐘 fine-tuning 語料量之相似度分數盒鬚圖
[Figure 58. The box-and-whisker plot of similarity scores using 3-minute fine-tuning data for cross-lingual voice conversion]

從實驗結果可以得知，無論是同語言還是跨語言任務，使用的 fine-tuning 語料量越少，音檔越難達到高品質的自然度與相似度，且跨語言的情況又比同語言更艱難。

7. 結論 (Conclusions)

在此論文中，我們利用所蒐集的 Taiwanese Across Taiwan (TAT) 大規模台文語音語料庫，包括，TAT-Vol1~2、TAT-TTS-M1~2 與 TAT-TTS-F1~2，完成了中文文字轉台語語音合成系統，與台語語音轉換系統（包括同語言（台語對台語）與跨語言（華語對台語）兩項語音轉換任務）。

其中的中文文字轉台語語音合成系統，在經利用校正台語變調以及新增兩種韻律符號，訓練出的新模型後，由實驗的結果也得知，合成音檔的自然度，可提升到 4.23 分，如圖 59 所示。

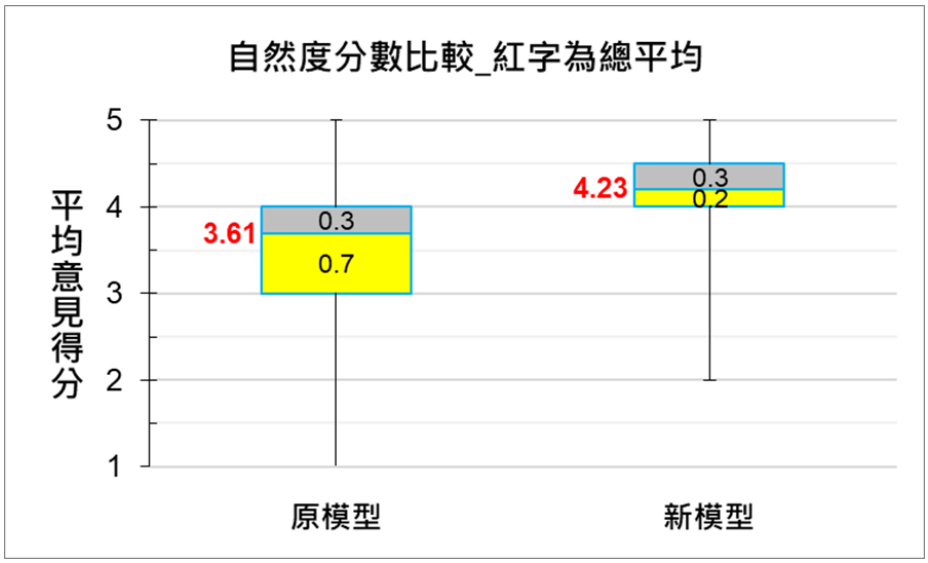


圖 59. 單人台語語音合成系統校正前後自然度實驗結果
[Figure 59. Experimental results on the naturalness of the single-speaker Taiwanese synthesis with and without tone snadi and prosodic boundary annotations]

在結合語音辨認及合成模組之台語多語者語音轉換實驗部分，同語言（台語對台語）語音轉換任務在語料量較充足，如 10 分鐘的情況下，已經可以達到音檔自然度與相似度都 3.45 與 3.38 分，如圖 60 和圖 61 所示。

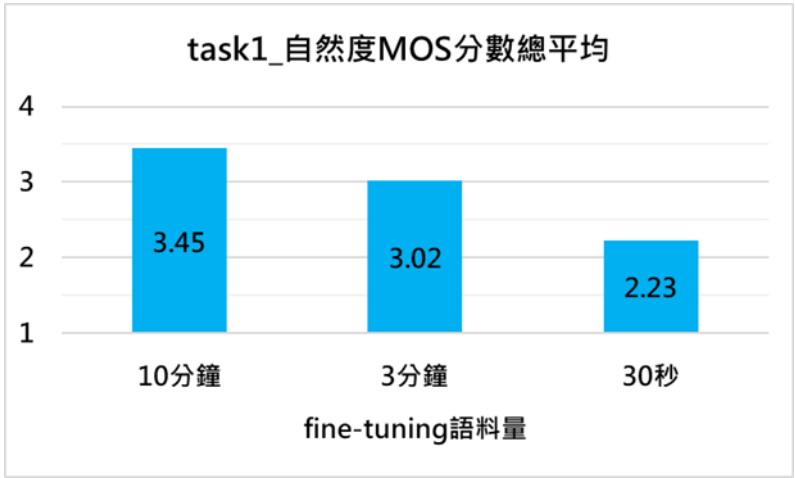


圖 60. 台語轉台語語音轉換系統自然度實驗結果
[Figure 60. Experimental results on the naturalness of the intra-lingual voice conversion]

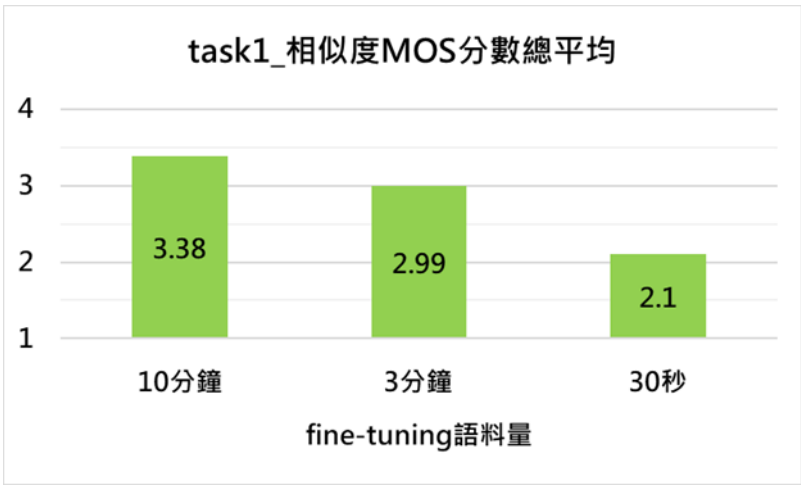


圖 61. 台語轉台語語音轉換系統相似度實驗結果
[Figure 61. Experimental results on the similarity of the intra-lingal voice conversion]

而在跨語言（華語對台語）語音轉換任務難度較高，但在只使用 6 分鐘語料量的情況下，自然度以及相似度的，也還可以達到 2.9 分跟 2.84 分，如圖 62 和圖 63 所示。

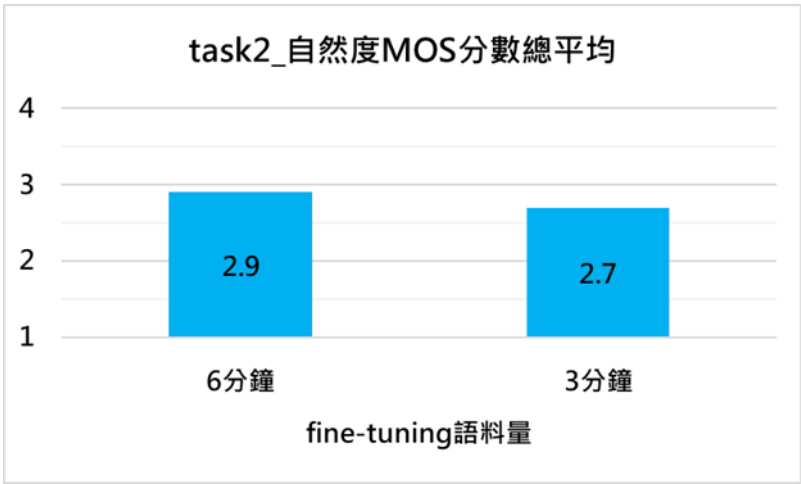


圖 62. 華語轉台語語音轉換系統自然度實驗結果
[Figure 62. Experimental results on the naturalness of the cross-lingual voice conversion]

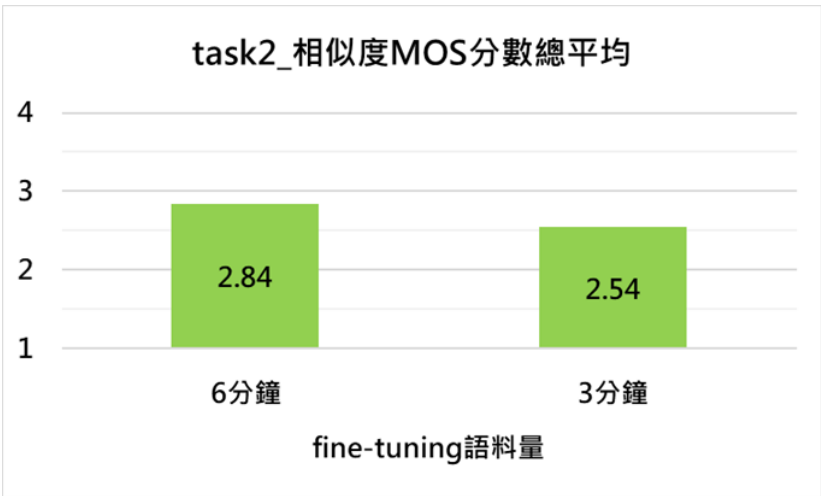


圖 63. 華語轉台語語音轉換系統相似度實驗結果
[Figure 63. Experimental results on the similarity of the cross-lingual voice conversion]

因此，由以上結果可知，我們所蒐集的 TAT 大規模台文語音語料庫，的確可以有效做為開發台語語音合成技術的語料庫。所做出來的中文文字轉台語語音合成系統，與台語語音轉換系統有都有還不錯的效能。

致謝 (Acknowledgements)

This work is supported partially by Taiwan's Ministry of Education under project "教育部閩

南語語音語料庫建置計劃", partially by Ministry of Science and Technology under contract No. 109-2221-E-027-108, 110-2221-E-027- 082, 110-2622-8-002-018, partially by National Taiwan University (NTU 109-3111-8-002-002), Cathay Life Insurance, Cathay Century Insurance, Cathay United Bank, Cathay Securities Corporation, and Cathay Securities Investment Trust, partially by Chunghwa Telecom. Lab. under contract No. TL-109-D304 and partially by the Department of Industrial Technology of Ministry of Economic Affairs under contract No. 110-EC-17-A-02-S5-008.

參考文獻 (References)

Chen, M., Tan, X., Ren, Y., Xu, J., Sun, H., Zhao, S., Qin, T., & Liu, T.-Y. (2020). MultiSpeech: Multi-Speaker Text to Speech with Transformer. arXiv preprint arXiv: 2006.04664v2

Dong, L., Xu, S., & Xu, B. (2018). Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). https://doi.org/10.1109/ICASSP.2018.8462506

esdeboer. (2020). GitHub-X-Vector,. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

howardhsu. (2021). GitHug-Facebook AI Research Sequence-to-Sequence Toolkit written in Python. Available https://github.com/pytorch/fairseq

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., & Wang, H.-M. (2016). Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder. arXiv preprint arXiv:1610.04019

Huang, W.-C. Huang, Hayashi, T., Watanabe, S., & Toda, T. (2020). The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS. arXiv preprint arXiv:2010.02434

kamo-naoyuki. (2021). GitHub-Mandarin to English Multi-speaker Transformer-TTS model. Available https://github.com/espnet/espnet/tree/master/egs/vcc20/tts1_en_zh

Prenger, R., Valle, R., & Catanzaro, B. (2018). WaveGlow: A Flow-based Generative Network for Speech Synthesis. arXiv preprint arXiv:1811.00002

rafaelvalle. (2020).GitHub-Tacotron 2 - PyTorch implementation with faster-than-realtime inference. Available: https://github.com/NVIDIA/tacotron2

rafaelvalle. (2020). GitHub-A Flow-based Generative Network for Speech Synthesis. Available https://github.com/NVIDIA/waveglow

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv preprint arXiv:1712.05884v2

shirayu. (2021). GitHub-Transformer-based ASR model. Available: https://github.com/espnet/espnet/tree/master/egs/librispeech/asr1

shirayu. (2021). GitHub-Multi-speaker Transformer-TTS model. Available: https://github.com/espnet/espnet/tree/master/egs/libritts/tts1

sih4sing5hong5. (2015). GitHub-iCorpus 臺華平行新聞語料庫語料加漢字. Available https://github.com/Taiwanese-Corpus/icorpus_ka1_han3-ji7

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. In proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). https://doi.org/10.1109/ICASSP.2018.8461375

Sun, L., Li, K., Wang, H., Kang, S., & Meng, M. (2016). Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In proceedings of 2016 IEEE International Conference on Multimedia and Expo (ICME). https://doi.org/10.1109/ICME.2016.7552917

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. arXiv preprint arXiv:1703.10135v2

Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., &Toda, T. (2020). Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. arXiv preprint arXiv:2008.12527

教育部。(2008)。臺灣閩南語羅馬字拼音方案使用手冊。Available: https://ws.moe.edu.tw/001/Upload/FileUpload/3677-15601/Documents/tshiutsheh.pdf [Ministry of Education. (2008). Tâi-oân Bân-lâm-gú Lô-má-jī Pheng-im Hong-àn.]