Computational Linguistics and Chinese Language Processing
Vol. 25, No. 2, December 2020, pp. 69-84

基於深度學習之中文文字轉台語語音合成系統初步探討

A Preliminary Study on Deep Learning-based Chinese Text to Taiwanese Speech Synthesis System

許文漢*、曾證融*、廖元甫*、王文俊+、潘振銘+

Wen-Han Hsu, Cheng-Jung Tseng, Yuan-Fu Liao,

Wern-Jun Wang and Chen-Ming Pan

摘要

台語在台灣歷史悠久，使用的族群眾多，有著很重要的存在價值。語音合成在追求跟人類一樣的聲音以及語調的同時，語言的多樣性也是一個需要深入探討的領域。本論文針對目前較少有的台語語音合成系統來作探討，利用翻譯模型 Chinese to Taiwanese(C2T)將輸入的中文文字轉成台羅拼音數字調(TLPA)，再將拼音輸入 Tacotron2 模型(Text to Spectrogram)後輸出頻譜，最後由 WaveGlow 模型(Spectrogram to Waveform)來實現語音合成。同時有架設網頁可供使用者一同來測試成效。

本文 C2T 機器翻譯的實驗方面採取三種模式，包括(1)輸入中文字詞，先進行斷詞，再輸出每個中文詞的台語台羅（Tâi-lô）拼音。(2)輸入中文字元串，直接輸出台羅拼音串。(3)輸入中文字元串，輸出台語的台羅拼音串與台語詞的斷詞關係。若不考慮聲調，方法(1)的 syllable error rate(SER)為 15.66%。而方法(2)的 SER 更可達 6.53%。這表示我們所用的 sequence-to-sequence 模型確實可以正確地將輸入的中文字元串，直接輸出台羅拼音串。

*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; {t107368030, yfliao}@ntut.edu.tw
+中華電信實驗室
Chunghwa Telecom Laboratories
E-mail: {wernjun, chenming}@cht.com.tw

在台語語音合成品質實驗方面，我們找了 20 位聽者，各聽取 15 句不同內容的合成音檔後，以平均主觀意見進行評分(mean opinion score，MOS，完全不像人講話的聲音為 1 分，完全像真人講話聲音為 5 分)。總計收集到 300 個評分，最後得到我們系統的 MOS 得分為 4.30 分。這表示我們所用的 Tacrtron2 與 WaveGlow 模型確實可以正確將台羅拼音串轉成台語語音。此外此系統的語音合成速度為一秒可合成約 3.5 秒之音檔，的確可以達到即時語音合成的要求。

Abstract

This paper focuses on the development and implementation of a Chinese Text-to-Taiwanese speech synthesis system. The proposed system combines three deep neural network-based modules including (1) a sequence-to-sequence-based Chinese characters to Taiwan Minnanyu Luomazi Pinyin (shortened to as Tâi-lô) machine translation (called C2T from now on), (2) a Tacotron2-based Tâi-lô pinyin to spectrogram and (3) a WaveGlow-based spectrogram to speech waveform synthesis subsystems.

Among them, the C2T module was trained using a Chinese-Taiwanese parallel corpus (iCorpus) and 9 dictionaries released by Academia Sinica and collected from internet, respectively. The Tacotron2 and Waveglow was tuned using a Taiwanese speech synthesis corpus (a female speaker, about 10 hours speech) recorded by Chunghwa Telecom Laboratories. At the same time, a demonstration Chinese Text-to-Taiwanese speech synthesis web page has also been implemented.

From the experimental results, it was found that (1) the best syllable error rate (SER) of 6.53% was achieved by the C2T module, (2) and the average MOS score of the whole speech synthesis system evaluated by 20 listeners gains 4.30. These results confirm that the effectiveness of integration of C2T, Tacrtron2 and WaveGlow models. In addition, the real-time factor of the whole system achieved 1/3.5.

關鍵詞：機器翻譯、臺灣閩南語羅馬字拼音、台語語音合成

Keywords: Machine Translation, Taiwanese Speech Synthesis, Tacotron2, Waveglow

1. 緒論 (Introduction)

造成台語使用人口式微的原因很多，最早可追溯至民國 34 年，國民政府接管臺灣後極力推行的「國語運動」，使得當時的學校禁止各省方言及原住民語，並嚴格推行「國語教育」(李玟逸、李祐萱、周楚，2017)。時至今日，人民生活中大多說國語為主，導致現代人熟悉台語的人數越來越少，尤其是年輕人，大部分懂的台語詞彙不多，講得也不甚

流利。此外，台語語言的演進也慢慢地與生活脫節，常有一些新的時、事、物，例如“磐石艦、討拍、滑鼠”等等，都不知道如何用台語來說，造成大家在用台語講話時，只好常常夾雜國語。

針對台語現階段的困境，若能做出一套中文文字轉台語語音合成的機器翻譯系統，讓使用者輸入中文文字後，能自動合成台語語音，就可以教大家如何講台語。讓使用者對台語提起興趣，並加強台語在日常生活中的應用，進而活化台語。尤其若能同時顯示教育部官方推薦的台羅拼音書寫系統，就能讓學生對台羅拼音有初步的認識及瞭解，進而能直接書寫台語。建立起一套中文文字轉台語語音合成的機器翻譯系統，通常需要三個模組，包括(1)將中文文字轉成以台羅（Tâi-lô）拼音表示的台語講法，(2)將台羅拼音轉為台語合成語音參數，最後(3)將合成語音參數轉成實際台語合成音檔。其中，其中以將中文文字轉成台羅拼音的機器翻譯模組最為重要，因為若翻譯的正確率不高，合成端有再好的音色品質和合成速度都是徒勞。

較早期的機器翻譯方法，有基於規則的字對字機器翻譯(RBMT)，基於範例的句對句機器翻譯(EBMT)，以及統計機器翻譯(SMT) (“機器翻譯，” 2020)。詞對詞的規則法即為將一個中文詞，依據規則與台華平行辭典，對照到一個台語拼音的翻譯法，適用於只注重單詞的非完整句子之翻譯，但翻譯出來的台語文法可能不正確。句對句的範例法為一整串中文句子對照到一整串台語的台羅拼音，可適用於句子的翻譯，也較能考慮文法差異。但此法常需依賴語料庫中台華平行句子的多樣化和數量，如要翻譯從未出現於語料庫中的句子，通常會較為困難。統計法目前為非限定領域機器翻譯中性能較佳的一種方法，通過對大量的台華平行語料進行統計分析，構建統計翻譯模型並進行翻譯，已經可以融合文句中語法等信息進一步提高翻譯的精確性。

例如，交大陳信宏(Kuo, Wang & Chen, 2004) (趙良基，2012)，中興余明興(潘能煌、余明興、許書豪，2011)與台大陳信希(Lin & Chen, 1999)等老師與都曾進行過中文文字轉成台羅拼音機器翻譯的相關研究。其中，陳信希老師曾在 1999 年，就發展出一個基於辭典翻譯，具有語音合成功能的 Mandarin to Taiwanese Min Nan Machine Translation System(目前已終止維護)。而且意傳科技也採用統計方法訓練出一套網頁版本的中文轉台語機器翻譯[1]。不過，此種機器翻譯模組，還需要先有一個華文斷詞與 POS 剖析器(自然語言剖析器，NLP parser)，才能順利進行後續的機器翻譯程序。但 NLP parser 本身就已經是一個難解的問題，而且通常會有大約 5%的分析錯誤(包括斷詞與 POS 標記)。若還是使用此傳統兩階段架構，就會讓前級產生的錯誤，連帶導致後面的翻譯與語音合成錯誤，而且後級只能接受，無法再加以挽救。

而近年來主流的機器翻譯方法為類神經網路機器翻譯(NMT)，顧名思義使用類神經網路(Neural Network)來做機器翻譯，其通常是基於 sequence-to-sequence 模型，使用 encoder-decoder 架構來學習輸入來源語言與輸出目標語言間的對應關係。NMT 尤其常使用 CNN 或是 RNN，來學習自然語言這種具有時間順序的序列數據(Sequence Data)的關

[1] 鬥拍字，https://suisiann.ithuan.tw/

係。例如給 encoder 端的 RNN 輸入一個來源語言的句子後，先利用 RNN 分析來源文字的語意，編碼成一個能代表原語句的語意向量序列。再讓 decoder 端的 RNN，以目標語言的語言模型知識，重新解譯該語意，輸出合乎目標語言架構的語句(Lee, 2019)。這樣就可以讓翻譯結果同時符合詞彙、文法與語意。

另一方面，目前的主流語音合成，也幾乎都是基於類神經網路技術，尤其以 Google 提出的 Tacotron2+WaveNet Vocoder 較為出名。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 Vocoder 出現以後，語音合成的音質就幾乎接近人聲。Tacotron2+WaveNet Vocoder 兩者的組合基本上就是目前的 State-of-the-Art 語音合成技術。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，幾乎無法用一般的 GPU 設備達到 real-time 的效能要求。

因此，目前語音合成研究主要是要解決合成速度問題。例如 Wave-RNN 與 WaveGlow。其中，NVIDIA 提出的 WaveGlow 跟 WaveNet 相比，可以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，約為 1:400，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。

因此，基於以上討論，我們將使用 sequence-to-sequence + Tacotron2 + WaveGlow 等模型來實現高品質且即時之台語語音合成。其架構為使用者輸入的中文文本透過 C2T (Ott, Edunov, Grangier & Auli, 2018)轉為台羅拼音，再透過 Tacotron2 (Shen *et al.*, 2018) 將台羅拼音轉為頻譜，最後透過 WaveGlow (Prenger, Valle & Catanzaro, 2018)將頻譜合成出台語語音，如圖 1 所示。

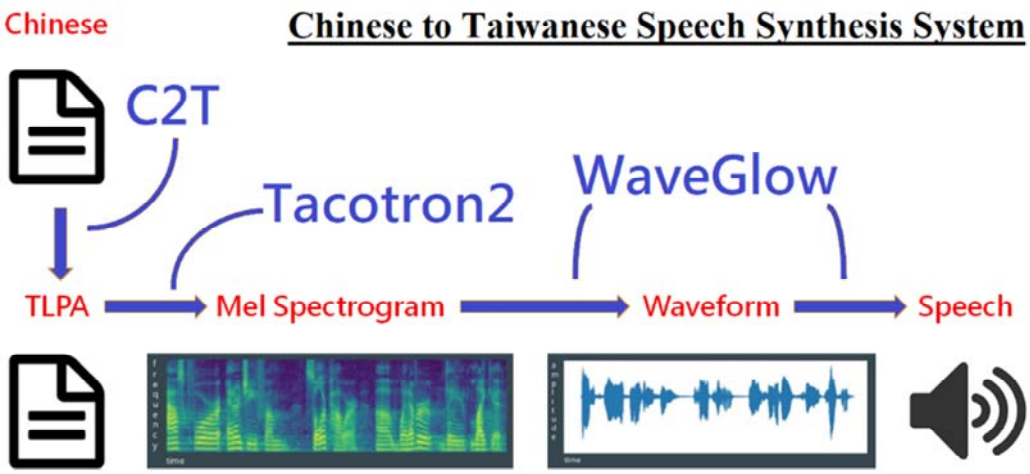

圖 1. 中文轉台語語音合成系統流程圖

[Figure 1. Chinese text to Taiwanese speech synthesis system flow chart]

為訓練此系統中的 C2T 模組，我們將利用中研院的 iCorpus 台華平行語料庫，與從網路上收集的多本台華平行辭典（包含教育部閩南語常用詞辭典），並採用 sequence-to-sequence 深度類神經網路架構，讓模型去學習如何將中文文字，轉換成台羅拼音。並利用中華電信錄製的單一語者台語語音合成語料庫，訓練 Tacotron2 與 Waveglow。希望能盡可能地達到中文文字翻譯成台羅拼音的正確性，與合成台語語音的高度自然度。

2. 中文文字轉台語語音合成系統 (Chinese text to Taiwanese speech synthesis system)

此系統基於深度學習之語音合成技術，以實現高品質且即時之台語語音合成。以下將進一步敘述 C2T，Tacotron2 和 WaveGlow 三個模型的實際作法。

2.1 Chinese to Taiwanese (C2T)

本文中的 C2T 採用 Facebook AI 研究院發表的 fairseq (Ott *et al*., 2018)為架構進行訓練，選用此法的原因為 fairseq 使用了比 RNN 效率和成果表現都更為優秀的 CNN 架構作為基礎。RNN 相比 CNN 有以下幾項缺點，(1)RNN 的模型是時序的，在處理序列的信息時只能逐項處理，不能並行操作，導致運行速度慢。(2)RNN 在處理較長的語句時，間格較遠的詞很難去學到詞與詞之間的依賴關係，並不能很好地處理句子中的結構化等更複雜的信息。(3) RNN 輸入多個單詞時，第一個單詞會經過 n 次單元的計算和非線性，但是最後一個單詞只會經過 1 次。

相比之下，CNN 改善了以上問題，除了能夠並行處理數據，Position Embedding 時輸入除了詞向量還加入位置向量，且 CNN 為層級結構，可顧慮到整段文句的每一個單詞，較底層 CNN 捕捉間隔較近的詞之間的依賴關係，較高層 CNN 則捕捉間隔較遠的詞之間的依賴關係。CNN 在 encoder 端，以 GLU 作為非線性單元，輸入與輸出相加後，才輸入到下一層網絡中，在 decoder 端，有 multi-hop attention 機制，encoder 端的輸出進行加權時，會考慮原始的輸入向量，在每一個卷積層都會進行 attention 的操作，使得模型在得到下一個 attention 時，能夠考慮到之前的已經 attention 過的詞。從 IBM Research 發表的研究論文”Comparative Study of CNN and RNN for Natural Language Processing“ (Yin, Kann, Yu & Schütze, 2017)，也可以看出在處理句子配對的任務上，比起 RNN 的 GRU，LSTM 等模型，CNN 擁有一定的優勢。

2.1.1 語料及辭典 (Corpus and Lexicon)

要訓練一個 C2T 模型，必須先準備好台華平行語料以及台華平行辭典。本模型使用的語料，為中研院資訊所陳孟彰老師計畫內的 iCorpus，此語料庫收集 3266 篇新聞，共 83544 句。算標點符號，台語 504037 詞、1030671 字，華語 501202 詞、1028218 字。以下為 iCorpus 的部份文章內容，如圖 2 所示。

	在地 力量 企業 贊助 萬國 戲院 重生		chai7-te7 lek8-liang7 khi3-giap8 chan3-chou7 Ban7-kok hi3-hng5 tiong5-seng
012	民視新聞報導	012	Bin5-si7-sin-bun5-po3-to7
	位於 大林鎮 的 萬國 戲院 曾經 風光一時 ，		ui7-ti7 Toa7-na5-tin3 e5 Ban7-kok hi3-hng5 chan-keng chhiann-iann7-chit8-si5 ，
	隨著 產業 沒落 經濟 蕭條 人潮 散去 ，		toe3-tioh8 san2-giap8 pang-pai7 keng-che3 chhin3-chhi7 jin5-tiau5 soann3-khi3 ，
	最後 發生 火災 停業 二十多年 。		siang7-boe2 hoat-seng hoe2-chai theng5-giap8 ji7-chap8-goa7-ni5 。
	現在 在 當地人 的 努力 規劃 ，		chit-ma2 ti7 chai7-te7-lang5 e5 lou2-lek8 kui-oe7 ，
	企業 的 贊助 加上 電視台 戲劇 演出 。		khi3-giap8 e5 chan3-chou7 ka-siang7 tian7-si7-tai5 hi3-kiok8 ian2-chhut 。
	配合 當時 的 時空 背景 場景 重建 ，		phoe3-hap8 tong-si5 e5 si5-khong poe3-keng2 tiunn5-keng2 tiong5-kian3 ，
	讓 萬國 戲院 彷彿 回到 風華 時代 。		hou7 Ban7-kok hi3-hng5 na2-chhiunn7 tng2-kau3 hong-hoa5 si5-tai7 。

圖 2. iCorpus 華台平行語料庫
[Figure 2. iCorpus Chinese-TLPA parallel corpus]

辭典方面，則是透過”ChhoeTaigi 找台語 ”網站之台語字詞資料庫，蒐集到的 9 本不同台語辭典合併成的台華平行辭典，各辭典統計的華語詞數，如圖 3 所示。

台語辭典	詞數
1. 台文華文線頂辭典	87670
2. 台日大辭典（台文譯本）	69552
3. Maryknoll台英辭典	55903
4. Embree台語辭典	36820
5. 教育部台語辭典	27487
6. 甘字典	24367
7. iTaigi華台辭典	8713
8. 台灣白話基礎語句	5301
9. 台灣植物名彙	1722
總共	317526

圖 3. 台語辭典華語詞數統計
[Figure 3. statistics of number of lexicon words]

因各個辭典由不同作者所撰寫，格式並非一致，為能在合成系統中使用，需要經過多次校正，校正目的主要是將台語詞翻譯成華語詞，華語詞即為平常生活中口語的慣用文字，檢查華語詞的意思與格式是否正確，並且與之對應的台羅拼音是否為一對一。台羅拼音也需檢查，剔除多餘之意思或符號。最終可使用的台華詞條數 225965，華語詞條數 88881，台語詞條數 153132。

2.1.2 模型訓練 (Model training)

中文轉台羅拼音為一種機器翻譯，做法為將中文文字序列轉台羅拼音序列，利用基於 sequence-to-sequence 深度類神經網路架構，學習如何進行轉換。此 C2T 採用網路上開源的 fairseq 架構(Ott *et al.*, 2018)進行訓練，其包括一 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，利用一中文對應台語拼音平行語料庫(iCorpus)，再加上台華平行辭典進行訓練，以此得到最佳的轉譯台羅拼音序列，如圖 4 所示。

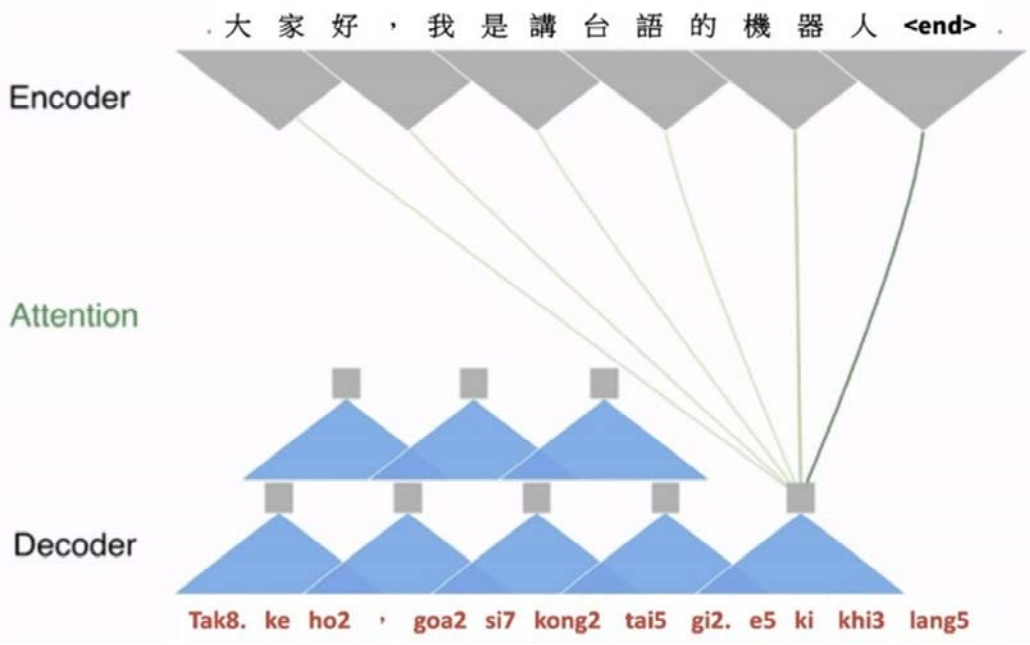

圖 4. 中文轉台語拼音模組
[Figure 4. C2T seq-to-seq model]

2.2 Tacotron2

此處模型訓練用之語料庫來源，為臺北科技大學和李江確台語文教基金會以及意傳科技合作產製，為一有大學教育程度之 34 歲男性錄製，台語腔調偏漳州腔，音檔筆數 9625 筆，長度約 10.4 小時。Tacotron2 為一 end-to-end 方式做訓練與推論之模型，使用架構為 encoder-decoder + Location Sensitive Attention。做法為將翻譯完成的台羅拼音輸入後，類神經網路進行文本分析，把語法與語意轉成語言特徵參數，讓系統知道文本中哪些是詞，哪些是句子，發什麼音，怎麼發音，發音時到哪應該停頓，停頓多長等等。語言特徵參數接著送入韻律產生器來產生文本裡每個音節的對應韻律訊息，包含基頻軌跡，音量，音長等，然後把說話的聲調，語氣，停頓方式，發音長短轉換成韻律參數(朱孝國，2005)。

最後輸出梅爾頻譜圖，再經由對齊達成台羅拼音與頻譜一對一的對應，如圖 5 所示。

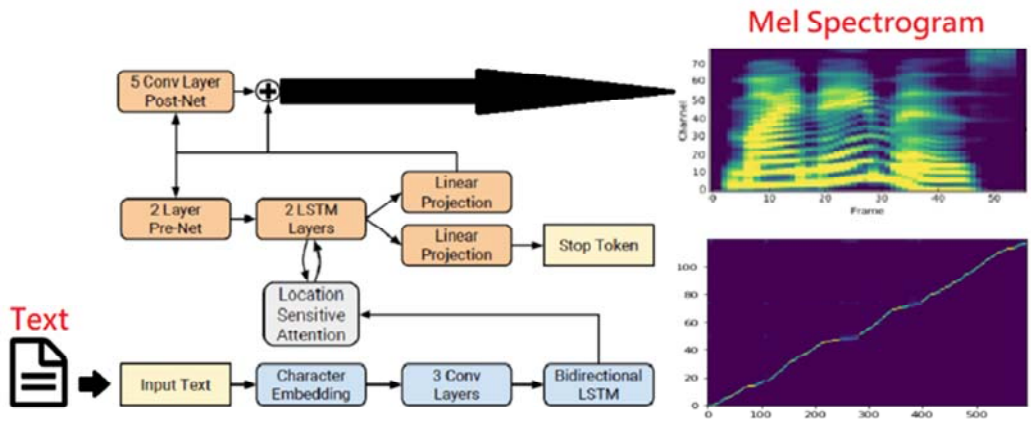

圖 5. Tacotron2 流程圖
[Figure 5. Tacotron2 flow chart]

2.3 WaveGlow

此處模型訓練用之語料庫來源同 Tacotron2，WaveGlow 是一種 flow-based generative networks，結合 Glow 和 WaveNet 的原理，透過輸入音檔與其生成之頻譜，僅使用單個網路與單個損失函式進行訓練，生成一高斯分布 z，合成時只需透過 z 與頻譜就可即時合成高品質語音，如圖 6 所示。

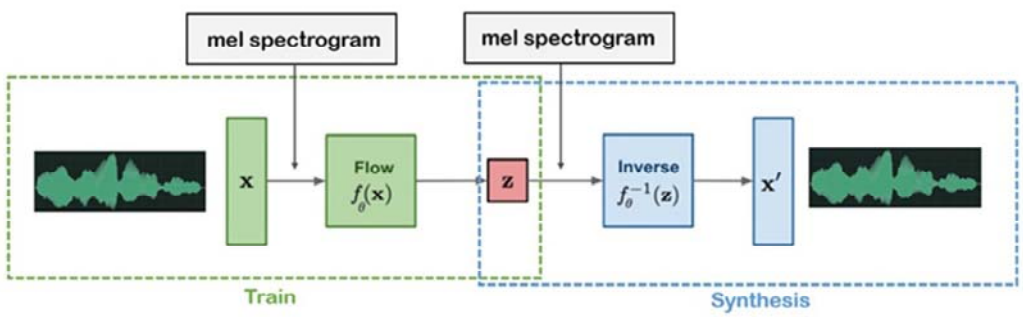

圖 6. WaveGlow 訓練及合成過程
[Figure 6. WaveGlow process of synthesis and training]

訓練時依據基於機率之 cost function 導引，多次利用函式轉換，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 的空間。並在訓練時限制 mapping 函式為可逆函式，WaveGlow 在生成波型圖時即可依據隱藏變數 z 空間取樣的結果，經多次函式轉換，逐步轉換成真實語音波形訊號 x。最後根據需要發出的聲音從資料庫中選擇出合適的聲學參數，然後根據在韻律模型中得到的韻律參數，透過語音合成演算法

產生語音(朱孝國，2005)。如圖 7 所示。

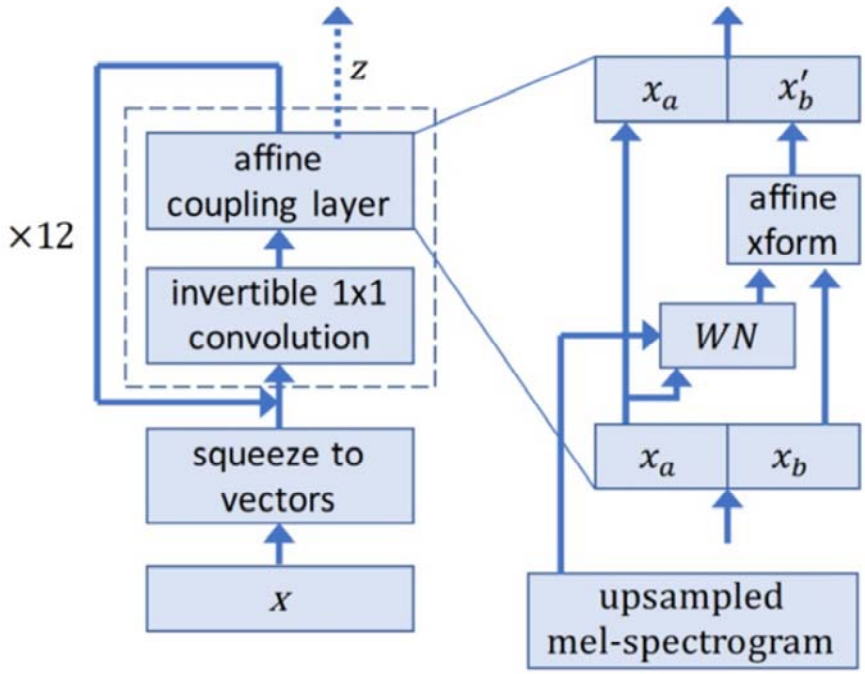

圖 7. WaveGlow 網路結構圖
[Figure 7. WaveGlow network structure diagram]

3. 中文文字轉台語語音合成雛形展示系統 (Website of Chinese text to Taiwanese speech synthesis system)

此基於深度學習之中文轉台語語音合成系統，已架設網頁版本以供使用，連結網址為 http://140.115.54.90:31810/。使用者輸入中文文字，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方式並合成語音。圖 8 展示為本文之使用者介面。網頁之紅色分隔線上方使用本文之 C2T 機器翻譯，下方為作為比較用之意傳科技統計法機器翻譯，合成端一律使用本文之 Tacotron2+WaveGlow。

其中，初步測試統計式機器翻譯後可以發現，統計式翻譯的結果較為接近中文文字本身念法的音譯，亦即如果要翻譯出中文轉台語後正確的台羅拼音，應輸入台文為佳，所謂台文即為中文轉為閩南語的另一種書寫表示形式。如中文的「現在是晚上八點」，統計式機器翻譯結果為「hian7 tsai7 si7 mng2 siong7 peh4 tiam2」，不是正確的台語發音，而台文的「這馬是暗時八點」翻譯後的「tsit4 ma1 si7 am3 si5 peh4 tiam2」，才是正確的台語發音。因此意傳科技此機器翻譯的「中文」轉台語翻譯，還是有所牽強。

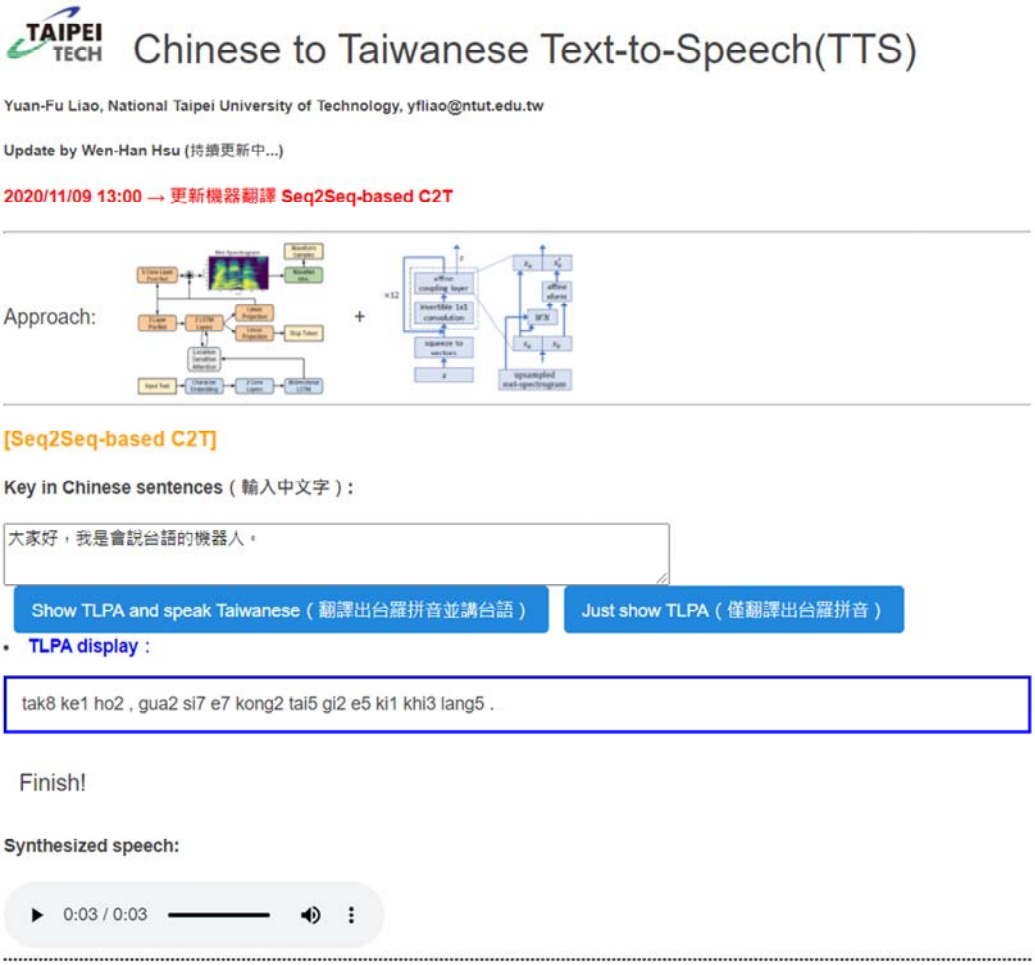

圖 8. 中文轉台語語音合成系統網頁
[Figure 8. Website of Chinese text to Taiwanese speech synthesis system]

4. 實驗 (Experiment)

4.1 C2T效能實驗 (C2T efficacy experiment)

本文之 C2T 模型分別以三種運行模式進行效果測試。模式一為中文句子透過斷詞後，判斷各個詞之詞性，再將各詞轉為台羅拼音，因此可以得到斷詞後含詞性之台羅拼音，如表 1 所示；模式二為將整句中文句子直接轉換為台羅拼音，如表 2 所示；模式三為將整句中文句子直接轉換為台羅拼音，同時學習台語之斷句規則，如表 3 所示。

表 1. 中文轉台羅拼音^斷詞/詞性
[Table 1. Chinese to TLPA^Hyphenation/ Part of speech]

中文句子	傅達仁今將執行安樂死，卻突然爆出自己 20 年前遭緯來體育台封殺，他不懂自己哪裡得罪到電視台。
斷詞	傅達仁 今 將 執行 安樂死 ， 卻 突然 爆出 自己 20 年 前 遭 緯來 體育台 封殺 ， 他 不 懂 自己 哪裡 得罪到 電視台。
詞性	Nb Nd D VC Na COMMACATEGORY D D VJ Nh Neu Nf Ng P Nb Na VC COMMACATEGORY Nh D VK Nh Ncd VJ Nc PERIODCATEGORY
台羅拼音^斷詞/詞性	poo3^B/Nb tat8^I/Nb jin5^E/Nb kim1^S/Nd tsiong3^S/D tsip4^B/VC hing5^E/VC an^B/Na lok8^I/Na si2^E/Na , khiok^S/D tut8^B/D jian5^E/D pok8^B/VJ chhut^E/VJ ka^B/Nh ki7^E/Nh ji7^B/Neu tsap8^E/Neu ni5^S/Nf tsing5^S/Ng cho^S/P hu7i^B/Nb la5i^E/Nb the2^B/Na iok8^I/Na tai5^E/Na hong^B/VC sat^E/VC , i^S/Nh bo5^S/D tong2^S/VK ka^B/Nh ki7^E/Nh to2^B/Ncd ui7^E/Ncd tioh8^B/VJ choe7^I/VJ kau3^E/VJ tian7^B/Nc si7^I/Nc tai5^E/Nc .

表 2. 中文轉台羅拼音
[Table 2. Chinese to TLPA]

中文句子	中央流行疫情指揮中心，今日表示，國內無新增確診個案。
台羅拼音	Tiong iang liu5 heng5 ek8 cheng5 chi2 hui tiong sim , kin a2 jit8 piau2 si7 , kok lai7 bo5 sin cheng7 chin2 ko3 an3 .

表 3. 中文轉台語詞
[Table 3. Chinese to words of TLPA]

中文句子	里長的言論在 PTT 引發熱議許多網友紛紛留言。
台語詞	li2-tiunn2-e5 gian5-lun7 ti7 PTT in2-huat4 jiat8-gi7 tsiann5-tse7 bang7-iu2 hun1-ue7 .

為測試以上三種 C2T 模式的效能，我們以 iCorpus 台華平行語料與辭典進行測試。實驗資料庫包括 iCorpus(78821 句)與台華辭典合集(225965 詞條)，並切分成三個子集，包括 Train 90%，Valid 5%，Test 5%。系統效能則以 Perplexity，及 Word error rate(WER)來衡量結果。考慮聲調的情況下，模式一到模式三的 WER 分別為 25.265%，7.102%以及 9.211%。不考慮聲調的情況下，模式一到模式三的 WER 分別為 18.660%，6.530%以及 8.699%。綜上數據得知，模式二由中文句子直接轉換為台羅拼音的效果最佳，如圖 9 所示。因此本文之 C2T 最終採用將中文直接轉換為台羅拼音的轉換法，以利接下來的台語語音合成工作。

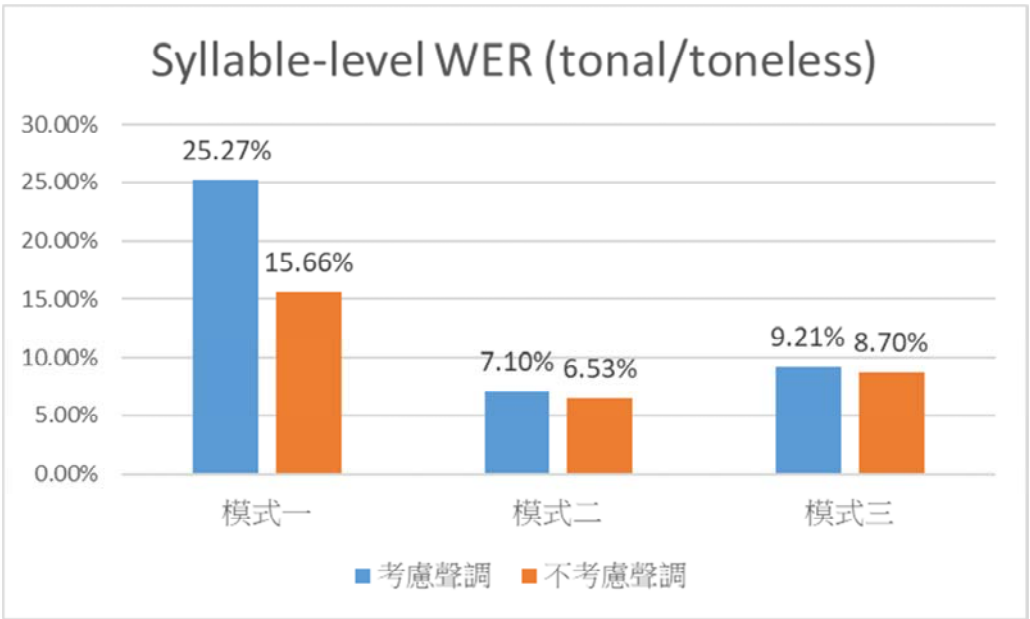

圖 9. C2T 錯誤率比較
[Figure 9. C2T Syllable-level WER]

4.2 Tacotron2+WaveGlow 合成語音品質實驗 (Tacotron2+WaveGlow synthesized speech quality experiment)

將事先準備好的 15 句合成音檔放上 Google 表單，並開放一般人對每個句子單獨進行評分。忽略語音內容翻譯錯誤或是語意不順等因素，僅根據聽到的「品質」評分 1.0 到 5.0 分。最低分為 1.0 分 為最接近機器人講話的聲音；最高分為 5.0 分 為最接近真人講話的聲音。評分到小數點後一位。開放評分時間約兩天，截止時有 20 位聽者評分。表 4 之 S1 到 S15 代表 15 句中文語句內容，為了不讓翻譯錯誤或是語意不通順等因素影響聽者對音檔品質的評分，所有句子皆有透過人工校正台羅拼音，且適當的加入斷詞標記使句子整體語氣通順，讓聽者可專心針對音檔「品質」評分。實驗最終結果之盒鬚圖如圖 10 所示。

總共 300 筆評分資料，最終平均意見得分(mean opinion score，MOS)約為 4.30 分。此實驗所使用之 15 個音檔皆由系統雛型展示網頁合成，由此高得分可知本文所用的 Tacrtron2 與 WaveGlow 模型確實可以正確合成出接近真人聲音之台語語音。

表 4. 測試中文語句內容
[Table 4. Chinese sentence content for experiment]

S1	大家好，我是會說台語的機器人
S2	請根據聽到聲音的音質打分數
S3	今天一早起來，天氣就非常炎熱
S4	一千兩百三十四萬五千六百七十九點零一美元
S5	韓國瑜是台灣歷史上，第一位被罷免的縣市首長
S6	武漢肺炎的出現，讓全世界的人都開始戴口罩
S7	現在是晚上八點，有一些老人應該想睡了
S8	歐美國家如:美國、加拿大、德國、法國、英國、西班牙、瑞典、瑞士、挪威、芬蘭等等
S9	這個猴死囝仔，竟然偷恁爸的錢去買那個垃圾
S10	吃予肥肥，裝予錘錘，裝予水水，等領薪水
S11	昨天地震時，我們家的花瓶掉下來摔破了
S12	龜笑鱉無尾，鱉笑龜粗皮
S13	歡迎光臨，請問有幾位
S14	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
S15	謝謝你付出寶貴的時間，參加這次的調查

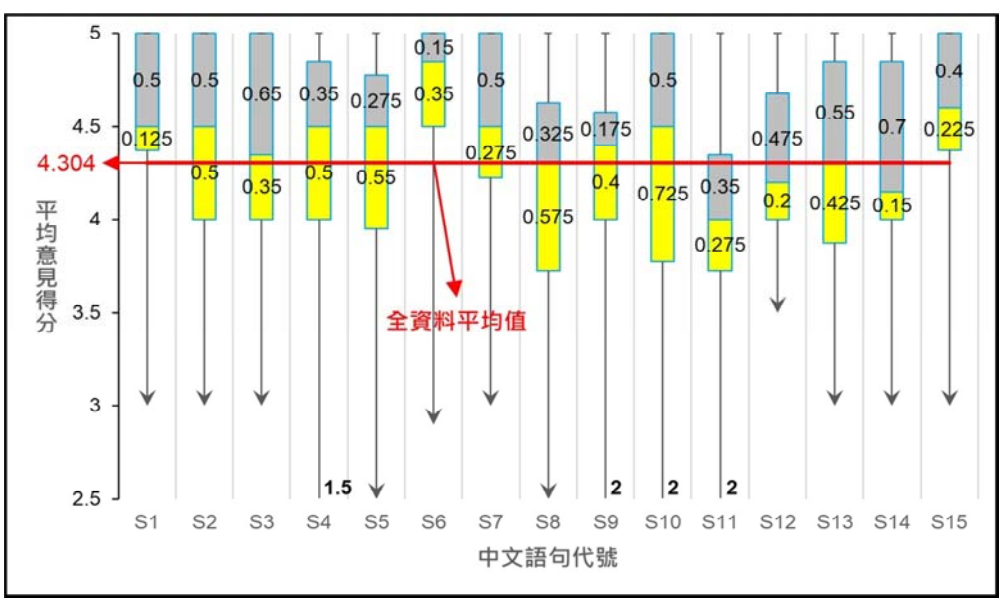

圖 10. 合成語音品質實驗結果盒鬚圖
[Figure 10. Box-plot of experimental results]

4.3 WaveGlow語音合成速度實驗 (WaveGlow speech synthesis speed experiment)

選用 WaveGlow 當作合成端的好處在於即時的語音合成速度，表 5 為一個簡單的 WaveGlow 合成音檔的速度實驗，時間的單位為秒。

表 5. WaveGlow 合成速度實驗
[Table 5. WaveGlow synthesis speed experiment]

音檔長度	5.83	4.25	7.48	3.96	9.16
合成花費時間	1.90	1.22	1.83	1.08	2.74

由表 5 可以得知，本文中的 WaveGlow 一秒約可合成 3.5 秒的音檔，相比原始合成速度非常緩慢的 WaveNet，已經可以達到即時合成的效益。

5. 結論(Conclusion)

我們提出的 C2T 機器翻譯，最好的 CER 值為 6.53%。表示 sequence-to-sequence 模型確實可以將輸入的中文文本，翻譯成正確的台羅拼音串。台語語音合成的 MOS 得分為 4.30 分，表示 Tacrtron2+WaveGlow 模型確實可以正確將台羅拼音串轉成台語語音。且系統的語音合成速度為一秒可合成約 3.5 秒之音檔，達到即時語音合成的要求。由以上實驗結果可以驗證我們中文文字轉台語語音的初步作法，確實有得到一定成效。未來將繼續增加台華平行辭典以及台語語料，進一步改善 C2T 的正確率，與合成語音的自然度。

致謝(Acknowledgements)

感謝中華電信研究院對於本論文提供的資源與協助。This work is supported partially by 中華電信 under the project “基於深度學習之台語語音合成『文字－聲學參數模型』”，Taiwan’s Ministry of Education under the project ”教育部閩南語語音語料庫建置計劃” and partially by Ministry of Science and Technology under the contract number 107-2221-E-027-102, 107-2911-I- 027-501, 107-3011-F-027-003, 108-2221-E-027-067 and 109- 2221-E-027-108.

參考文獻 (References)

Kuo, W.-C., Wang, Y.-R., & Chen, S.-H. (2004). A MODEL-BASED TONE LABELING METHOD FOR MIN-NAN/TAIWANESE SPEECH. In *Proceddings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2004*, 505-508. doi: 10.1109/ICASSP.2004.1326033

Lin, C.-J. & Chen, H.-H. (1999). A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan. *Int. J. Comput. Linguist. Chinese Lang. Process.*, *4*(1), 59-84.

Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling Neural Machine Translation. In arXiv preprint arxiv: 1806.00187

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

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., …Wu, Y. (2018). Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions. In arXiv preprint arxiv: 1712.05884

Yin, W., Kann, K., Yu, M., & Schutze, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing. In arXiv preprint arXiv:1702.01923.

Lee, M. (2019 年 6 月 17 日)。淺談神經機器翻譯 & 用 Transformer 與 TensorFlow 2 英翻中 [部落格文字資料]. 取自 https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html [Lee, M. (2019, June 17). Talk about neural machine translation & Using Transformer and TensorFlow 2 translates English into Chinese. [Web blog message]. Retrieved from https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html

朱孝國(2005 年 4 月 28 日)。語音合成 Speech Synthesize Note [部落格文字資料]。取自 https://irw.ncut.edu.tw/peterju/speech.html#categories [Chu, H.-K. (2005, April 28). 語音合成 Speech Synthesize Note [Web blog message]. Retrieved from https://irw.ncut.edu.tw/peterju/speech.html#categories]

李玟逸、李祐萱、周楚(2017 年 3 月 31 日)。年輕人不諳台語 語言傳承不能等【部落格文字資料】。取自 http://shuj.shu.edu.tw/blog/2017/03/31/年輕人不諳台語-語言傳承不能等/。[Wen-Yi Li, You-Shiuan Li, Chu Jou. (2017, March 31). Young people do not know Taiwanese. Language inheritance cannot wait. [Web blog message]. Retrieved from http://shuj.shu.edu.tw/blog/2017/03/31/年輕人不諳台語-語言傳承不能等/

趙良基 (2012)。**台語語音合成技術之研究**(碩士論文)。取自 http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22GT070060254%22.&searchmode=basic [Chao,L.-J. (2020). *The Research of Speech Synthesis Technology for Taiwanese (Master ' s thesis*). Retrieved from http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22GT070060254%22.&searchmode=basic]

潘能煌、余明興、許書豪(2011)。中文文句轉台語語音系統之連音變調預估模組。**資訊科技國際期刊**，**5**(1)，118-128。[Pan, N.-H., Yu, M.-S., & Shiu, S.-H. (20110. A Tone Sandhi Prediction Module for Chinese to Taiwanese Text-to-Speech Systems. *Int. J. Adv. Inf. Technol*., *5*(1), 118-128.

機器翻譯(2020 年 9 月 17 日)。In Wikipedia, the free encyclopedia. Retrieved November 10, 2020, from https://zh.wikipedia.org/wiki/机器翻译 [Machine Translation (2020, September 17). In Wikipedia, the free encyclopedia. Retrieved November 10, 2020, from https://zh.wikipedia.org/wiki/机器翻译]