# 單語者台灣腔中文即時語音合成系統

# Real-Time Single-Speaker Taiwanese-Accented Mandarin Speech Synthesis System

王奕雯 Yih-Wen Wang
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
M083040011@student.nsysu.edu.tw

陳嘉平 Chia-Ping Chen
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
cpchen@mail.cse.nsysu.edu.tw

## 摘要

本論文研究單語者台灣腔中文即時語音合成系統，架構上採用文字序列端到梅爾頻譜圖序列端的合成器，再串接一個從梅爾頻譜圖到語音訊號的聲碼器。首先，我們嘗試使用 GST Tacotron-2 合成器串接 Griffin-Lim 聲碼器，搭配不同的資料集，包括北京腔中文語料與台灣腔中文語料等等，以及不同的訓練方式，包括遷移式學習(Transfer Learning)與集成式學習(Ensemble Learning)等等，進行了三種系統設定實驗。接著我們使用 Tacotron-2 串接 Griffin-Lim 架構與中文語料，實驗是否使用預訓練模型(Pretrained Model)，再進行了兩種系統設定實驗。最後，我們從上述五種系統設定中挑選出 MOS 最高者，再將其聲碼器從 Griffin-Lim 替換成 WaveGlow，評估兩種聲碼器對 MOS 的影響。我們使用的資料集包含單人中文 12 小時的標貝語料、單人中文 4.5 小時的個人錄製語料、單人中文 2.2 小時的教育廣播電台語料，以及單人英文 24 小時的 LJSpeech 語料。最終 MOS 最高的單語者台灣腔中文即時語音合成系統為，使用標貝語料預訓練、再使用教育廣播電台語料接續訓練的 Tacotron-2 模型，並串接使用 LJSpeech 語料預訓

練、再使用標貝語料接續訓練的 WaveGlow 模型，MOS 評分可達 4.32，且該語音合成系統產生 10 秒 48kHz 的語音只須 1.3 秒，因此為即時語音合成系統。

## Abstract

In this paper, we study a real-time single-speaker Taiwanese-accented Mandarin speech synthesis system. This system uses an end-to-end sequence-to-sequence model from the text sequence to the Mel spectrogram sequence, and a vocoder to map the Mel spectrogram sequence to synthesized speech waveform. We first use the GST Tacotron-2 sequence-to-sequence model and the Griffin-Lim vocoder. The system is trained with several datasets, such as Mainland-accented Mandarin corpus and Taiwanese-accented Mandarin corpus, and with different training methods including transfer learning and ensemble learning. In this stage, three experiments were carried out. In addition, we use Tacotron-2 and Griffin-Lim with the same data sets and experimented with using model pretraining. In this stage, two experiments were carried out. Finally, the system setting with the highest MOS in the experiments is selected, and the Griffin-Lim vocoder is replaced by WaveGlow vocoder. The datasets we use include 12-hour Biaobei Mandarin corpus, 4.5-hour personal recording Mandarin corpus, 2.2-hour National Education Radio Mandarin corpus, and 24-hour LJSpeech English. At the end of day, the Real-Time Single-Speaker Taiwanese-Accented Mandarin Speech Synthesis System with the highest MOS we achieved is the system as follows: Tacotron-2 is pretrained with the Biaobei corpus, and then trained with the National Education Radio corpus, and the WaveGlow vocoder is pretrained with the LJSpeech corpus, and then trained with the Biaobei corpus. This system achieves the MOS score of 4.32 and generates 10 seconds of 48kHz speech in 1.3 seconds.

關鍵詞：Tacotron-2、GST Tacotron-2、Griffin-Lim、WaveGlow、Transfer Learning、Ensemble Learning、Pretrained Model

## 一、緒論

大數據時代的到來，深度學習成為熱門議題之一，人機互動的情況也早已普及，像是數位助理、智能導航、以及有聲書等等。在這些廣泛的應用當中，語音合成的技術就扮演了相當重要的角色。雖然語音合成的產品眾多，且能產生中文語音的技術也相當成熟，但合成的中文語音其實大多數為「北京腔調的中文語音」，會形成此結果的主要原因是因為可大量取得的中文語料，多為北京腔調的語者錄製而成。因此該研究希望透過神經網路的語音合成技術，利用端到端直接學習從文本到聲學特徵的對應關係，並且使用有限的中文語料，搭配不同的訓練方式，達到「台灣腔調的中文語音合成」。

一個完整的語音合成系統需要合成器與聲碼器，在神經網路的訓練過程中，合成器使用可訓練的神經網路 Tacotron-2[3]、GST Tacotron-2[4]，其輸入為成對的文字與音檔，透過端到端的神經網路，一次輸出一幀的梅爾頻譜圖(Mel-Spectrogram)；聲碼器則使用了演算法 Griffin-Lim[1]以及可訓練的神經網路 WaveGlow[2]，在演算法 Griffin-Lim 中，輸入是梅爾頻譜圖中幅度譜的資訊，透過六十次迭代演算輸出時域波形；而在神經網路 WaveGlow 中，輸入為成對的音檔與梅爾頻譜圖，透過多個可逆的變換函數組成序列，最後輸出時域波形。訓練完成後，於推斷時只要輸入一段欲合成的文字，透過合成器輸出梅爾頻譜圖，最後由聲碼器輸出語音訊號。本文分為四個部分：第一部分為緒論；第二部分為研究方法，介紹資料集的使用、資料前處理、合成器與聲碼器的模型架構、不同模型架構與不同資料搭配不同的訓練方式；第三部分為實驗結果的分析與評估；第四部分為結論。

## 二、研究方法

### （一）資料集

#### 1. 標貝資料集

標貝資料集是由「標貝(北京)科技有限公司」於 2018 年所開放。由一位中年女性錄音者錄製而成，時長共 12 小時，採樣頻率為 48KHz、16-bit，錄製環境為專業錄音棚環境，語料內容包含各類新聞、小說、科技等領域，詳細的資料規格如表一。後續內容提及時，將該資料集簡稱為「Biaobei 資料集」。

#### 2. 客製化資料集

該資料集為我個人製作。會製作此資料集的主要原因是希望合成的中文語音能具備台

表一、Biaobei 資料集數據規格

| | |
|---|---|
| 數據內容 | 中文標準女聲語音庫數據 |
| 語音類型 | 標準普通話 (北京腔調) |
| 錄音者 | 單一語者，中年女性 |
| 錄音環境 | 專業錄音棚環境，無背噪 |
| 錄音工具 | 專業錄音設備 |
| 有效時長 | 共約 12 小時，10000 個 wav，3~5s/wav |
| 採樣格式 | 無壓縮 wav 格式，採樣率為 48KHz、16-bit |
| 標註格式 | 文本標註為.txt 格式；音節音素標註為.interval |

表二、YW 資料集數據規格

| | |
|---|---|
| 數據內容 | 客製化中文女聲語音庫數據 |
| 語音類型 | 標準國語 (台灣腔調) |
| 錄音者 | 單一語者，青年女性 |
| 錄音環境 | 研究室，略有背噪 |
| 錄音工具 | 個人筆電麥克風 |
| 有效時長 | 共約 4.5 小時，3800 個 wav，3~5s/wav |
| 採樣格式 | 無壓縮 wav 格式，採樣率為 48KHz、16-bit |
| 標註格式 | 文本標註為.txt 格式 |

灣腔調，而目前開放免費使用的單一語者中文語料大多數為北京腔調。整份資料集的製作由我進行錄音，錄製的文本內容為 Biaobei 資料集的文本，錄製的環境為安靜無人的研究室，詳細的資料規格如表二。後續內容提及時，將該資料集簡稱為「YW 資料集」。

3. NER-Trs-Vol1 資料集

此資料集全名為「北科大教育電台廣播節目語音語料庫」，主要是大量轉寫教育電台節目，產生節目音檔逐字稿，並作人工校正與切割，形成長度約 30 秒的音檔，以建置大規模台灣腔語料庫。但該資料集的錄製者為多語者，將會導致合成的語音其語者具有隨機性、非單一性，並不符合這次的單語者語音合成目標，此外，該資料集的每個音檔時長過長，使得訓練難度提升。因此，決定從多語者中，提取出單一語者的音檔，並將每個 30 秒的音檔人工切割成約 10 秒的音檔，詳細的資料規格如表三。後續內容提及時，將該資料集簡稱為「NER-2hr 資料集」。

表三、NER-2hr 資料集數據規格

| | |
|---|---|
| 數據內容 | NER 中文女聲語音庫數據 |
| 語音類型 | 標準國語 (台灣腔調) |
| 錄音者 | 單一語者，中年女性 |
| 錄音環境 | 錄音室內或錄音室以外之場所，略有背噪 |
| 錄音工具 | 教育電台節目錄製 |
| 有效時長 | 共約 2.2 小時，1495 個 wav，5~8s/wav |
| 採樣格式 | 無壓縮 wav 格式，採樣率為 16KHz、16-bit |
| 標註格式 | 文本標註為.txt 格式 |

表四、LJSpeech 資料集數據規格

| | |
|---|---|
| 數據內容 | LJSpeech 英文女聲語音庫數據 |
| 語音類型 | 美式英文 |
| 錄音者 | 單一語者，中年女性 |
| 有效時長 | 共約 24 小時，13100 個 wav，1~10s/wav |
| 採樣格式 | 無壓縮 wav 格式，採樣率為 22050Hz、16-bit |
| 標註格式 | 文本標註為.csv 格式 |

4. LJSpeech 資料集

這是一個公開的英文語音數據集，文本在 1884~1964 年之間出版，音檔由「LibriVox」於 2016 年至 2017 年錄製。總共 13,100 個音頻，每個音頻平均長度 1~10 秒不等，總時長約 24 小時，錄製者為同一女性，內容來自七部非小說類書籍，詳細的資料規格如表四。後續內容提及時，將該資料集簡稱為「LJSpeech 資料集」。

（二）資料前處理

1. 文字

由於中文字本身有數萬個相異字，加上有許多同音異字的情況，這導致無法以窮舉的方式對神經網路進行訓練。為了解決此問題，我們使用漢語拼音作為文本的輸入，並且加上數字 1~5 來表示聲調。此外，我們也對文本進行斷詞，進而提升合成中文語音的流暢度。在訓練完成後進行合成時，首先先對輸入的文本經過 jieba 套件進行斷詞，再透過 pypinyin 套件形成漢語拼音，進行後續的語音合成，整體流程可參考表五。

2. 語音訊號

在進到神經網路訓練前，會將語音訊號進行前處理，生成「梅爾頻譜圖」作為輸入。前處理的部分是使用幀大小為 50 毫秒、幀移為 12.5 毫秒，以及漢明窗(Hanning Window)進行計算，然後通過短時傅立葉轉換(STFT)得到線性頻譜。接著使用頻率範圍在 125Hz~7.6kHz、通道數為 80 的梅爾濾波器組，對 STFT 的線性頻率進行過濾，再對函數進行壓縮，從而把 STFT 幅度轉換到梅爾刻度上。

表五、文本資料前處理流程

| | |
|---|---|
| 原始文本 | 不好意思，我找不到我想要的書。 |
| 經過 jieba 進行斷詞 | 不好意思 \| ， \| 我 \| 找 \| 不到 \| 我 \| 想要 \| 的 \| 書 \| 。 |
| 經過 pypinyin 進行漢語拼音 | bu4 hao3 yi4 si1 \| ， \| wo3 \| zhao3 \| bu2 dao4 \| wo3 \| xiang3 yao4 \| de \| shu1 \| 。 |

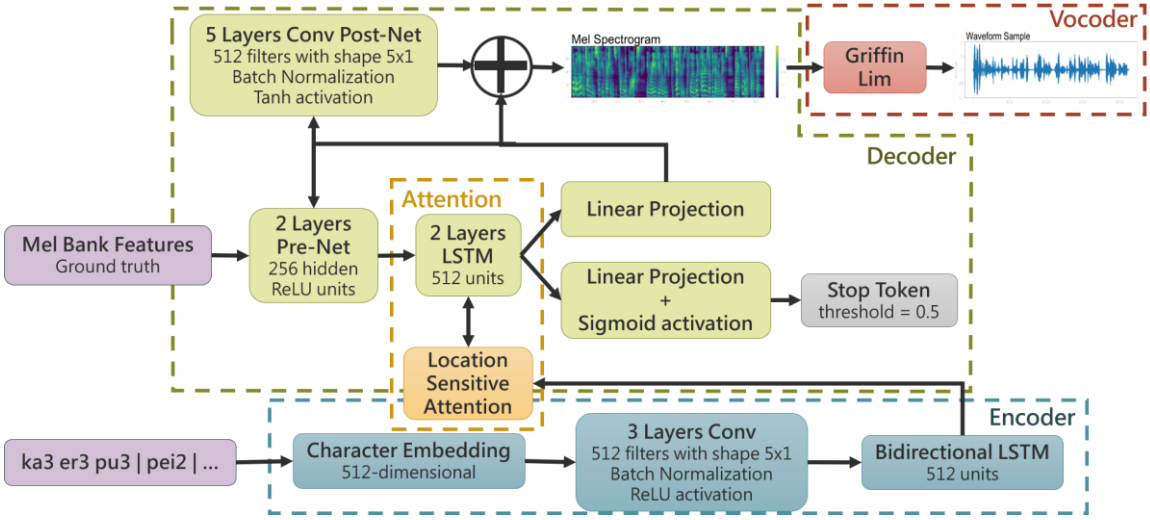
## （三）合成器

### 1. Tacotron-2

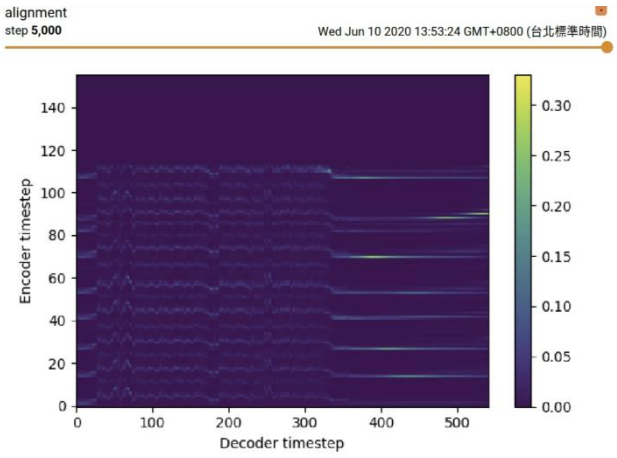
Tacotron-2 是由 Google Brain 於 2018 年提出來的一個語音合成框架[3]，模型架構如圖一，主要由三個部分組成，分別為編碼器(Encoder)、位置敏感的注意力機制(Location Sensitive Attention)與解碼器(Decoder)。編碼器將輸入的字元編碼成 512 維的字元向量(Character Embedding)，透過三層卷積(Convolution Neural Network)獲取序列中的上下文訊息、雙向長短期記憶(Bi-LSTM)將文本編碼成一個固定向量。接著透過位置敏感的注意力機制，給予解碼器在不同時間步有不同的權重。梅爾頻譜圖作為解碼器中兩層全連接預處理網路(Pre-Net)的輸入，預處理網路的輸出會與上一個時間步的上下文向量拼接送入兩層單向的長短期記憶(LSTM)，長短期記憶的輸出被用作計算本時間步的上下文向量，並且經過線性投映(Linear Projection)後，分別用來預測線性頻譜圖，每一次預測一幀，也用來計算停止符機率(Stop Token)。為了提取更高維的特徵，線性頻譜圖會經過五層的卷積後處理網路(Post-Net)來預測一個殘差，疊加到未經過後處理網路的線性頻譜圖，形成梅爾頻譜圖。而停止符機制的運作原理，是將經過線性投影的結果由 Sigmoid Activation 去預測輸出的頻譜序列是否完成的一個機率，當機率大於 0.5 時，頻譜圖的生成即停止。

在經過 Post-Net 之前，會將經過線性投影預測出的線性頻譜圖與真實頻譜圖計算一個損失；在經過 Post-Net 之後，會將經過殘差疊加後產生的梅爾頻譜圖與真實頻譜圖也計算一個損失。這兩項計算原先皆是使用 MSE Loss Function，而我們將其改成 Huber Loss Function 並比較兩者的結果。Huber Loss Function 主要是結合 MSE 與 MAE 的優點。MSE 的優點是收斂較快，因為它的梯度是隨著損失值在改變，但缺點是遇到離群值時，經過平方後計算的損失值會較大，對模型造成不好的影響。而 MAE 的優點則是對離群值較有魯棒性，損失值較低，但缺點是收斂速度慢，因為其梯度始終為 1，也因此容易錯過損失值最低的點。整個公式如(1)、(2)，使用一個超參數 $\delta$ 來控制要側重 MSE 或是 MAE，當誤差值小於 $\delta$ 時，使用 MSE，使其收斂快速；當誤差值大於 $\delta$ 時，使用 MAE，避免離群值造成較大的損失值。圖二、三分別為使用 MSE 與 Huber Loss 在訓練過程中編碼與解碼的對齊圖，可以發現同樣在步數(Step)為 5K 時，圖三已有較明顯的對齊圖，圖二則尚未。修改成 Huber Loss 後可以使訓練收斂更加快速，並且損失值越低，代表預測出的頻譜圖與真實的頻譜圖越接近，語音合成的效能也進而提升。而後續提及 Tacotron-2 的模型架構皆為修改成此損失函數的模型架構。

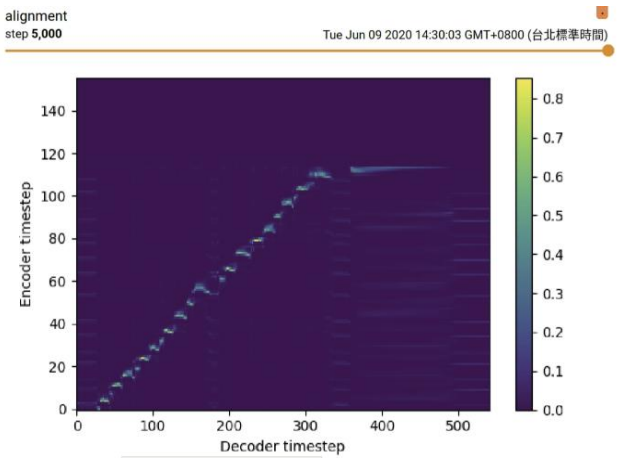
$$\text{Huber Loss}(y,\hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2, & |y-\hat{y}| \le \delta \quad (1) \\ \delta\,|y-\hat{y}| - \frac{1}{2}\delta^2, & |y-\hat{y}| > \delta \quad (2) \end{cases}$$




圖一、Tacotron-2 模型架構




圖二、Mean Square Error Loss Function




圖三、Huber Loss Function

2. GST Tacotron-2

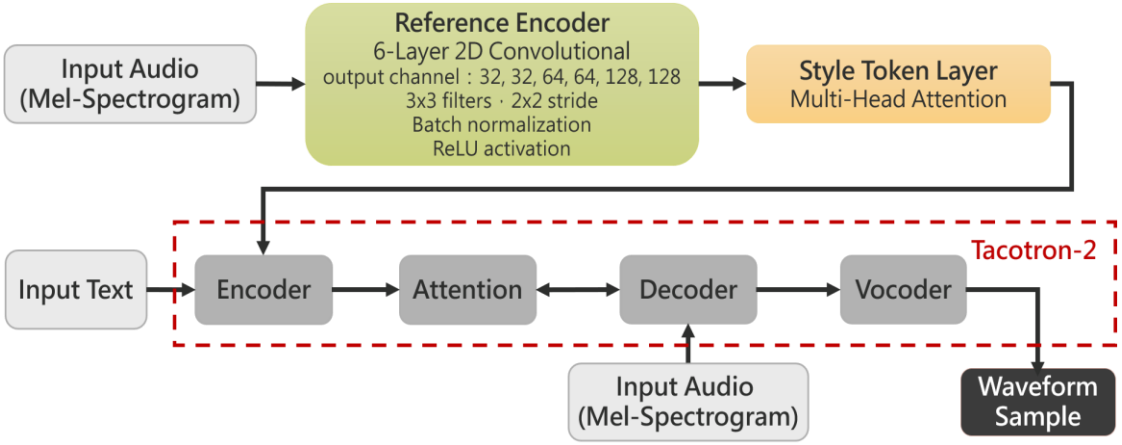
為了合成高自然度的人聲，語音合成系統必須學會對韻律建模，韻律包含語音的所有表現力因素，例如語調、節奏和重音。透過向 Tacotron-2 多增加一個注意力機制，它將語音片段中的韻律嵌入，並表達成一組基礎嵌入固定集合的線性組合，這種嵌入方式稱為，並且這些韻律無須事先標記，因此是屬於無監督式學習。基於 Tacotron-2 往上增加的模塊主要為參考編碼器(Reference Encoder)，與風格標記層(Style Token Layer)，如圖四。參考編碼器的輸入為梅爾頻譜圖，先通過六層二維的卷積(Convolution Neural Network)，再接一層單向門控循環單元(GRU)，將聲音編碼成一個

512 維的固定向量，稱為參考特徵(Reference Embedding)。風格標記層主要是由多頭注意力機制(Multi-Head Attention)構成，也就是作多次的自注意力機制(Self-Attention)，這裡設定作 8 次。模型隨機初始化一組風格特徵(Style Token Embedding)集合，而我們使用的訓練資料其風格多樣性並未很高，因此設定風格數為 4。在自注意力機制中，Q(query)為 512 維的參考特徵，K(key)與 V(value)皆為 64 維的風格特徵。運作流程是先將 512 維的參考特徵切割成 8 個 64 維的參考特徵，每一個參考特徵皆與 4 個風格特徵作點積(Dot)，進行相似度計算得到權重，將權重透過 Softmax Activation 使得權重落在 0~1 之間，接著將權重與 4 個風格特徵進行加權，得到一個 64 維的風格特徵，如公式(3)。總共進行 8 次上述的自注意力機制，最後將 8 個 64 維的風格特徵作拼接(Concat)，形成一個 512 維的風格特徵，如公式(4)。經過 GST 後得到的風格特徵，會與 Tacotron-2 中編碼器的輸出向量作拼接，進行後續解碼的動作。

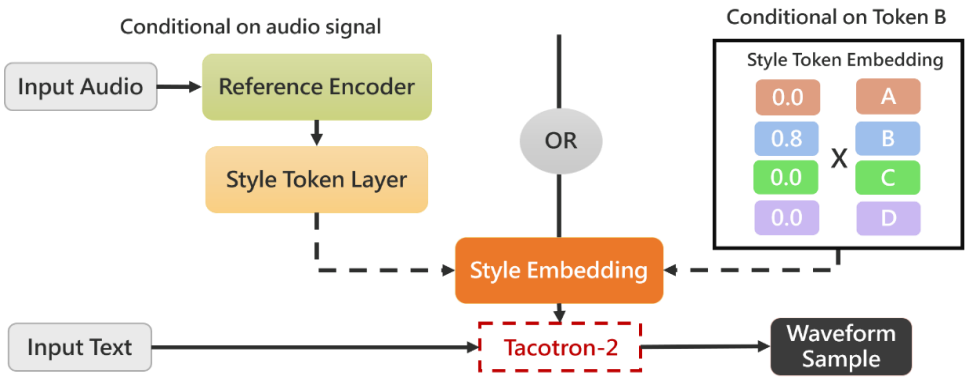
訓練完成後，推斷方法有兩種，如圖五。第一種是輸入欲合成的文本、欲合成此風格的音檔，透過參考編碼器與風格標記層得到風格特徵；第二種方法是輸入欲合成的文本，並給予一組指定風格特徵進行加權的權重，直接得到需要的風格特徵，此方法還能得知每種風格的資訊。因為在訓練過程中，模型隨機初始化一組風格特徵，我們只知道風格數量，但不會得知每種風格可能代表的是語速快慢，或是音調高低等等，那麼可以透過只給特定風格大於 0 的權重，其餘皆為 0，藉此可知該風格的資訊為何。得知每種風格的資訊後，可以更自由的進行線性組合，計算出想要的風格特徵，得到更多樣風格的語音合成。而在接續的實驗中，兩種推斷方式我們皆會嘗試，並從中比較出哪一種推斷方式以及風格參數設定，可以得到最好的合成效果。

$$Head_i = \text{Self-Attention}(QW_i^Q, KW_i^K, VW_i^V) = \sum \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (3)$$

$$\text{Multi-Head Attention} = \text{Concat}(Head_1, Head_2, \ldots\ldots, Head_i)W^{output} \quad (4)$$




圖四、GST Tacotron-2 模型架構




圖五、GST Tacotron-2 推斷流程

（四）聲碼器

1. Griffin-Lim

這是一種由演算法來合成語音的聲碼器[1]，這種聲碼器不需要事先訓練，它的輸入為合成器輸出的梅爾頻譜圖。Griffin-Lim 重建語音訊號時，需要使用幅度譜與相位譜，但在梅爾頻譜圖中是不包含相位訊息的。於是演算法第一步驟是用噪聲隨機初始化一個相位譜，第二步驟是將已知的幅度譜與初始化的相位譜經過逆傅立葉轉換(ISTFT)，得到一個初步的時域訊號。接著第三步驟，是將上一步得到的時域訊號經過傅立葉轉換(STFT)，得到新的幅度譜與新的相位譜，此時是從一個不準確的時域訊號得到幅度譜與相位譜，於是第四步驟用原先已知的幅度譜取代新的幅度譜，接著與新的相位譜再透過逆傅立葉轉換，得到一個更準確的時域訊號。如此重複上述步驟二到四，直至迭代出一個穩定的時域訊號，這裡設定的迭代次數為六十次。此演算法的整個流程可以參考表六。

2. WaveGlow

WaveGlow 是由 NVIDIA 研究小組於 2018 年提出，它是一個基於流的生成模型，透過分佈採樣生成語音，只需一個神經網路與一個最小化負對數似然(negative log likelihood)的損失函數，即可生成時域波形。主要是由多個可逆的轉換函數組成序列，

表六、Griffin-Lim 演算步驟

| |
|---|
| Step1：隨機初始化一個相位譜 |
| Step2：此相位譜與已知的幅度譜經過 ISTFT，合成新的語音 |
| Step3：對合成的語音作 STFT，得到新的相位譜與新的幅度譜 |
| Step4：丟棄新的幅度譜，用已知的幅度譜與新的相位譜，再次合成新的語音 |
| Step5：重複 Step2、3、4，直到迭代次數達到六十次，即得最終的合成語音 |

將一個簡單的分佈轉換到一個複雜的分佈，藉此模擬訓練數據的分佈，最後再透過最小化負對數似然值，進行優化。圖六為整個網路架構，首先將 8 個聲音採樣值拼接成一個向量，此動作稱為 squeeze，接著通過 12 層的 1*1 可逆卷積(Invertible Convolution)與仿射耦合層(Affine Coupling Layer)。在仿射耦合層中，只會將 x 一半的通道數 $x_a$ 作為輸入，並與梅爾頻譜圖進入 WN function。WN 是由 8 塊多種卷積層與殘差模組組成，包含單位卷積層(Pointwise Convolution)、空洞卷積層(Dilated Convolution)，以及殘差跳躍連接(Residual and Skip Connection)，此架構類似於 WaveNet[8]，計算後輸出 s 和 t，接著將剩下的另一半通道數 $x_b$，由 s 和 t 的轉換公式(9)得到 $x_b'$，並將$x_a$與$x_b'$作拼接(Concat)。此外，在仿射耦合層中，同一半部中的通道不會直接被修改，但若不跨通道的混合訊息，那會有部分的參數不會被調整，因此在每層仿射耦合層前，會添加 1*1 的可逆卷積，使得通道間的訊息可被混合。

$$\mathbf{z} \sim \mathrm{N}(\mathbf{z}; 0, \mathrm{I}) \quad (5)$$

$$\mathbf{x} = f_0 \circ f_1 \circ \dots f_k(\mathbf{z}) \quad (6)$$

$$x_a, x_b = \mathrm{split}(\mathbf{x}) \quad (7)$$

$$(\log \mathbf{s}, \mathbf{t}) = \mathrm{WN}(x_a, \text{Mel-spectrogram}) \quad (8)$$

$$x_b' = \mathrm{s} \odot x_b + \mathrm{t} \quad (9)$$

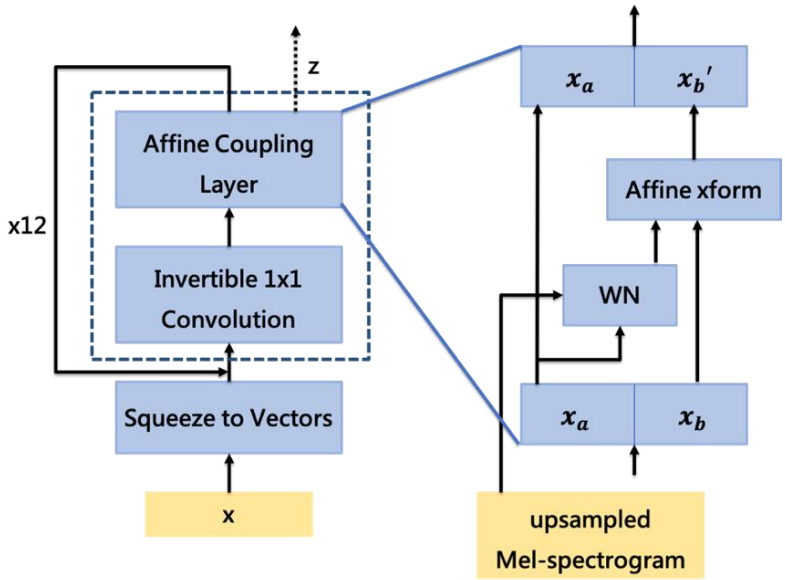
$$f_{coupling}^{-1}(\mathbf{x}) = \mathrm{concat}(x_a, x_b') \quad (10)$$

$$\log\left|\det\left(J\left(f_{coupling}^{-1}(\boldsymbol{x})\right)\right)\right| = \log|\boldsymbol{s}| \quad (11)$$

$$f_{couv}^{-1} = \mathrm{Wx} \quad (12)$$

$$\log\left|\det\left(J\left(f_{conv}^{-1}(\boldsymbol{x})\right)\right)\right| = \log|\boldsymbol{W}| \quad (13)$$

$$\log p_\theta(\boldsymbol{x}) = -\frac{z(x)^T z(x)}{2\sigma^2} + \sum_{j=0}^{\#coupling} \log s_j + \sum_{k=0}^{\#conv} \log W_k \quad (14)$$




圖六、WaveGlow 模型架構

（五）合成器－訓練方式

這部分使用五種不同的訓練過程，對合成器進行訓練，目的都是希望能使合成的中文語音具備台灣腔調。我們利用不同的資料集、凍結某部分參數的權重，或是只取出部分參數權重與另一模型結合，詳細過程會在以下說明，而以下五項實驗於聲碼器統一使用 Griffin-Lim 合成語音。

1. Tacotron-2(YW 資料集)

使用個人錄製的 YW 資料集對 Tacotron-2 進行訓練。合成的效果的確是台灣腔的中文語音，但因為個人錄製的環境及設備並非專業，使得資料集的音檔略帶雜音，且個人的發音並非完全正確清晰，語調較平淡，語速較緩慢，導致最後合成的語音品質尚須許多改善。

2. GST Tacotron-2(Biaobei 資料集)，freeze Tacotron-2，GST Tacotron-2(YW 資料集)

為了改善方法一的合成品質，我們嘗試了方法二。首先使用 Biaobei 資料集訓練 GST Tacotron-2， 當此模型的損失值已經收斂時，將模型所有參數的權重保存，接著凍結 Tacotron-2 的參數權重，目的是希望先保有 Biaobei 資料集的語音品質，最後使用 YW 資料集繼續訓練 GST Tacotron-2，去微調 GST 的參數權重，希望可因此學習到 YW 資料集的風格與腔調。訓練完畢後並進行語音合成時，於 GST 中兩種推斷方式皆嘗試：由於我們希望合成語音能具有 YW 資料集的腔調，因此首先使用 YW 資料集中未經過訓練的音檔作為參考音檔，透過參考編碼器與風格標記層得到其風格特徵；接著我們嘗試指定 4 種風格參數的權重為 style(A, B, C, D)=style(0.0, 0.4, 0.0, 1.2)。而兩者的合成效果相似，其語音的確具有 Biaobei 資料集的語音品質，但 YW 資料集的風格只學習到了語速較慢的特色。

3. GST Tacotron-2(YW 資料集)，freeze GST，GST Tacotron-2(Biaobei 資料集)

與方法二同時進行的是方法三。首先使用 YW 資料集訓練 GST Tacotron-2，接著將所有參數權重保存，凍結 GST 的參數權重，這樣的流程是希望先保有 YW 資料集的風格與腔調，最後使用 Biaobei 資料集繼續訓練 GST Tacotron-2，去微調 Tacotron-2 的參數權重，希望具有 Biaobei 資料集的語音品質。訓練完成後並進行語音合成時，同樣先使用 YW 資料集中未經過訓練的音檔作為參考音檔，接著也指定 4 種風格參數的權重為 style(A, B, C, D)=style(0.0, 0.4, 0.0, 1.2)，進行兩種推斷方式。兩者的合成結果也近似，其合成語音的確有 Biaobei 資料集乾淨清晰的品質，且有 YW 資料集的平淡語調、語速

較慢的風格，但台灣腔的成分卻仍不夠明顯，但此方法的合成效果是目前最接近實驗的目標。

4. GST Tacotron-2(Biaobei 資料集)與 GST Tacotron-2(YW 資料集)，各取部分之參數權重進行組合

嘗試過凍結參數權重並微調的方法後，我們設想是否能透過模型融合的方式達到想要的結果。首先使用 Biaobei 資料集與 YW 資料集，分別訓練 GST Tacotron-2，使得目前有兩組相同模型架構、經由不同資料集訓練後的參數權重。接著，於使用 Biaobei 資料集訓練的模型中，取出 Tacotron-2 的參數權重；使用 YW 資料集訓練的模型中，取出 GST 的參數權重，進行合併，得到一組新的 GST Tacotron-2 的參數權重。訓練完成後，使用兩種推斷方式：取 YW 資料集中未經過訓練的音檔作為參考音檔、指定 4 種風格參數的權重為 style(A, B, C, D)=style(0.0, 0.4, 0.0, 1.2)。在不作微調而是直接給定參數權重的情況下，兩者的合成效果卻仍與方法三大同小異，由此可知，方法三的參數權重已微調至與方法四的參數權重相近了。

5. Tacotron-2(Biaobei 資料集)，作為 pretrain model，Tacotron-2(NER-2hr 資料集)

方法一到四中，我們希望透過 GST Tacotron-2 這個神經網路搭配不同的訓練方式，能夠獲得具有音質好、台灣腔調的語音合成，但合成的結果總是無法完全學習到腔調的部分。因此想要透過整理出來的 NER-2hr 資料集，單純訓練 Tacotron-2，來達成目標。但因為 NER-2hr 資料集的資料量偏少，且人工切割音檔容易導致音檔的平均長度範圍偏大，致使訓練的難度提升，於是利用 pretrain model 的想法來解決此問題。首先使用 Biaobei 資料集訓練 Tacotron-2，且將 Biaobei 資料集的採樣率調至與 NER-2hr 資料集相同，當模型的損失值已收斂時，在不凍結任何參數權重的情況下，繼續訓練 NER-2hr 資料集。此方法的合成結果，順利達成我們想要的音質好、台灣腔調的中文語音合成。

### （六）聲碼器－訓練方式

透過上述五種實驗，可藉由方法五順利達成這次的目標。但上述的方法中聲碼器的部分，皆是以演算法 Griffin-Lim 來進行梅爾頻譜圖轉語音的動作，而聲碼器卻也是影響語音品質的因素之一。因此，我們將方法五搭配不同的聲碼器，希望達到更好的合成品質。

1. Griffin-Lim

Griffin-Lim 為一種迭代的演算法，並非需要透過資料集訓練的神經網路，因此不須討論資料集以及其訓練方式。而在實驗中，我們設定演算法的迭代次數為六十次，以確保穩定性。

2. WaveGlow

針對此神經網路，我們首先使用 LJSpeech 資料集進行訓練，作為一個 pretrain model，接著再使用 Biaobei 資料集接續訓練。會選擇這樣的訓練方式，是因為我們直接使用 Biaobei 資料集訓練時，其合成的語音略帶雜音，而改成使用 NER-2hr 資料集時，因為資料量偏少，即導致欠擬合的情況發生。

## （七）小結

1. YW 資料集與 NER-2hr 資料集之差別

YW 資料集為個人自行錄製的語料集，錄製環境與設備皆非專業等級，該資料集只適用於一般的個人研究上，當要建一個完整並供大多數人使用的語音合成系統時，此資料集的質量會導致語音合成的品質低落；反之，NER-2hr 資料集是透過專業設備、專業人士製作而成，其品質能有一定的保障。因此，即便起初已有客製化的台灣腔中文語料集，但仍想嘗試後續實驗，希望使用有限且質量好的資料集，藉由不同的訓練方式，來達成台灣腔中文語音合成的目標。

2. 微調(Fine-Tune)不同模型參數對合成結果之影響

在合成器的訓練方式中，方法二、三皆嘗試凍結模型中部份不須再更新的參數，並微調模型中其餘須再繼續更新的參數，而從實驗結果可以發現，方法三比方法二更接近實驗目標。我們的目標是希望藉由 GST 模組學習到台灣腔，因此直接先使用 YW 資料集對 GST 模組做訓練再凍結該參數，比起使用 Biaobei 資料集訓練 GST 模組後再微調成 YW 資料集的參數權重，台灣腔的效果能夠更加明顯。

3. 預訓練模型(Pretrained Model)對合成結果之影響

當我們使用較小資料量的語料集，重新訓練一個複雜度高的模型時，容易導致過擬合(overfitting)的情形發生，即在訓練集上能合成完整的語音，但在測試集上卻有漏字、靜音、雜音等等的問題，因此在合成器的訓練方式其方法五、聲碼器的訓練方式其方法二中，皆使用 pretrained model 的方式進行後續訓練，改善合成不佳的問題。

## 三、實驗結果

在合成器與聲碼器的各項訓練方式皆訓練完畢後，我們將生成五句與訓練資料不重複的中文文本，進行語音合成，由八位受測人員進行評估，評分標準採用 MOS(Mean Opinion Score)，在每一句子聽完後給予主觀分數，分數範圍為 1~5 分。首先針對合成器的五項實驗進行評分，越高分表示合成的語音其台灣腔調越明顯、合成音質越乾淨無噪，測試結果如表六。接著針對合成器五項實驗中，分數最高的實驗串接兩種不同的聲碼器進行評分，越高分表示合成品質越好，測試結果如表七。

## 四、結論

我們目前的研究在 Tacotron-2[3]、GST Tacotron-2[4]上嘗試了許多訓練方式，包括凍結參數的權重、組合兩個模型參數的權重，以及使用預訓練模型進行訓練，並將這些訓練結果與 Griffin-Lim[1]、WaveGlow[2]進行結合，分別評估其 MOS 的高低。透過八名受測人員評分的結過，可以發現方法五，使用 Biaobei 資料集訓練 Tacotron-2 作為預訓練模型，接著使用 NER-2hr 資料集接續訓練，並串接聲碼器 Waveglow，此實驗 MOS 為最高者。雖然最後合成台灣腔的中文語音，並非藉由 GST Tacotron-2 提取聲音韻律特徵的模型架構徹底達成，但可以發現透過不同的訓練方式，該模型能夠合成出風格多樣性的中文語音。在語音合成的領域，除了透過學習韻律使得聲音更接近人聲之外，也期望能合成更接近日常用語的語音。越來越多人使用的文字內容不再只有單一種語言，經常中英夾雜的使用，因此我們未來的方向將會朝混合語言的語音合成進行研究，了解並應用代碼轉換(Code Switching)[5]、語音克隆(Voice Cloning)[6]等技術，期望能夠合成具有韻律且混合語言的語音。

表六、合成器五項實驗之 MOS

| 方法 | 實驗方法簡述 | 於推斷時，相關參數設定 | | |
|---|---|---|---|---|
| | | MOS<br>(無須相關設定) | MOS<br>(設定參考音檔) | MOS<br>(設定風格權重) |
| 一 | Tacotron-2(YW) | 3.91 | - | - |
| 二 | GST Tacotron-2(Biaobei)，<br>freeze Tacotron-2，<br>GST Tacotron-2(YW) | - | 2.30 | 2.83 |

| | | | | |
|---|---|---|---|---|
| 三 | GST Tacotron-2(YW)，<br>freeze GST，<br>GST Tacotron-2(Biaobei) | - | 3.03 | 3.37 |
| 四 | GST Tacotron-2(Biaobei)<br>與 GST Tacotron-2(YW)，<br>各取部分之參數權重進行組合 | - | 3.10 | 3.21 |
| 五 | Tacotron-2(Biaobei)，<br>作為 pretrain model，<br>Tacotron-2(NER-2hr) | **4.25** | - | - |

表七、聲碼器兩項實驗之 MOS

| 方法 | 實驗方法簡述 | MOS |
|---|---|---|
| 一 | 表六、方法五 + Griffin-Lim | 4.25 |
| 二 | 表六、方法五 + WaveGlow | **4.32** |

## 參考文獻


[1]. Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin-lim iteration. CoRR, abs/1903.03971, 2019.

[2]. Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. CoRR, abs/1811.00002, 2018.

[3]. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predic- tions. CoRR, abs/1712.05884, 2017.

[4]. Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Batten- berg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. CoRR, abs/1803.09017, 2018.

[5]. Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng. End- to-end code-switched tts with mix of monolingual recordings. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6935–6939, 2019.

[6]. Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to speak fluently in a foreign language: Multilingual speech syn- thesis and cross-language voice cloning. CoRR, abs/1907.04448, 2019.

[7]. Bohan Zhai, Tianren Gao, Flora Xue, Daniel Rothchild, Bichen Wu, Joseph E. Gonzalez, and Kurt Keutzer. Squeezewave: Extremely lightweight vocoders for on-device speech synthesis. CoRR, abs/2001.05685, 2020.

[8]. A¨aron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Ko- ray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.