Computational Linguistics and Chinese Language Processing
Vol. 24, No. 2, December 2019, pp. 53-62

即時中文語音合成系統

Real-Time Mandarin Speech Synthesis System

鄭安傑*、陳嘉平*

An-Chieh Cheng and Chia-Ping Chen

摘要

本論文研究與實作即時中文語音合成系統。此一系統採用文字序列到梅爾頻譜序列的轉換模型，再串接一個從梅爾頻譜到合成語音的聲碼器。我們使用 Tacotron2 實作序列到序列轉換模型，配合數種不同的聲碼器，包括 Griffin-Lim，World-Vocoder，與 WaveGlow。其中以實作可逆編碼解碼函數的 WaveGlow 神經網路聲碼器最為突出，無論在合成速度或語音品質方面，皆令人印象深刻。我們使用單人 12 小時的標貝語料實作系統。在語音品質方面，使用 WaveGlow 聲碼器的合成系統語音的 MOS 為 4.08，略低於真實語音的 4.41，而遠勝另兩種聲碼器（平均 2.93）。在處理速度方面，若使用 GeForce RTX 2080 TI GPU，使用 WaveGlow 聲碼器的合成系統產生 10 秒 48 kHz 的語音僅需 1.4 秒，故為即時系統。

Abstract

This thesis studies and implements the real time Chinese speech synthesis system. This system uses a conversion model of the text sequence to the Mel spectrum sequence, and then concatenates a vocoder from the Mel spectrum to the synthesized speech. We use Tacotron2 to implement a sequence-to-sequence conversion model with several different vocoders, including Griffin-Lim, World-Vocoder, and WaveGlow. The WaveGlow neural network vocoder, which implements the reversible codec function, is the most prominent, and is impressive in terms of synthesis speed or speech quality. We use a single speaker with 12-hour corpus implementation system. In terms of voice quality, the MOS of the synthesized system voice using the WaveGlow vocoder is 4.08, which is slightly

*國立中山大學資訊工程學系
Department of Computer Science and Information Engineering, National Sun Yat-sen University
E-mail: ajcheng@g-mail.nsysu.edu.tw; cpchen@mail.cse.nsysu.edu.tw

lower than the 4.41 of the real voice, and far better than the other two vocoders (average 2.93). In terms of processing speed, if the GeForce RTX 2080 TI GPU is used, the synthesis system using the WaveGlow vocoder produces a voice of 10 seconds and 48 kHz in 1.4 seconds, so it is a real time system.

關鍵詞：文字轉語音，Tacotron2,WaveGlow

Keywords: TTS,Tacotron2,WaveGlow

1. 緒論 (Introduction)

隨著科技的發展，人機互動的情況也越來越普及，像是 siri 語音助理、智能導航、有聲讀物等都已環繞在我們生活裡。而其中，語音合成的技術就扮演了一個非常重要的腳色。語音合成是透過機械、電子的方式產生人造語音的技術，文字轉語音技術也隸屬於語音合成。而本研究則是致力於開發出一個可合成出更快且更為逼真的文字轉語音合成系統。實現語音合成的方法有多種，其中包含參數式合成以及拼接式合成。基於參數式的語音合成系統主要是透過統計學模型，利用學習出來的語音學特徵和其聲學特徵的對應關係後，預測出相應的參數，接著聲碼器再透過這些參數合成出所期望的音頻。不過這種合成方式最大缺點乃為無法合出接近人類的自然語音，在技術上尚未有明顯的突破。拼接式語音合成系統則是透過同樣的方式去預測出這些聲學特徵，然後再到原始語音庫中找尋近似的音素，最後拼接而成。不過這種合成方法也意味著合成的音質穩定性與語音庫大小成正比。若要能合成出完善的自然語句，就必須要有齊全的資料庫，且同時為了不延遲搜尋上的效率，更必須要有個良好的演算法。而上述這些方法，除了皆有著明顯的人工痕跡之外，在專業領域上的門檻也都極高。幸運地，還有一種神經網路式的合成技術，可利用神經網路直接學習從文本端到聲學特徵端的對應關係。

2. 研究方法 (Research Methods)

2.1 資料集 (Dataset)

訓練資料選自標貝資料集，是由「標貝科技有限公司」於 2018 年所開放。由一位女性錄音者錄製而成，全長約略 12 小時，使用 48kHz 16bit 採樣頻率，錄製環境為專業錄音室及錄音軟體，語料涵蓋各類新聞、小說、科技、娛樂等領域，詳細規格如表 1。

表 1. 標貝資料集數據規格
[Table 1. Biaobei Data Specification]

數據規格	
數據內容	中文標準女聲語音數據庫
錄音語料	綜合語料樣本量:音節音子的數量、類型、音調、音連以及韻律等進行覆蓋。
有效時長	約 12 小時
平均字數	16 字
語言類型	標準普通話
發音人	女 : 20-30 歲
錄音環境	聲音採集環境為專業錄音室 1. 錄音室符合專業音庫錄製標準 2. 錄音環境和設備自始至終保持不變 3. 錄音環境的信造比不低於 35dB
錄製工具	專業錄音設備及錄音軟體
採樣格式	無壓縮 PCM WAV 格式，採樣率為 48kHz、16bit。

2.2 預測網路 (Predicted Model)

在前端預測網路部分我們重現了 Google 的 Tacotron2 (Shen *et al*., 2017)，並針對中文語音合成系統做了客製化。在架構上面使用的是一個編碼器-解碼器(Encoder-Decoder)的設置，並加入了位置敏感的注意力機制(Location sensitive attention) (Chorowski, Bahdanau, Serdyuk, Cho & Bengio, 2015)，整體架構如圖 1。而由於我們使用的是中文資料集，故在文本的內容上先進行了資料的預處理，目的是為了讓神經網路可以學習到我們中文上的韻律以及抑揚頓挫。由於漢字本身有數萬個相異字，同音異字的情況也不在少數，若以窮舉的方式來對神經網路做訓練顯然不夠明智。我們處理的方式是使用漢語拼音作為字元標註，並採用數字一到四來表示我們的聲調。雖然過程中都是以這樣的形式進行訓練，不過在合成階段時我們可以藉由”pypinyin”的套件直接透過中文輸入合成出所指定的句子。另外，為了提升中文語音的流暢度，我們也透過”jieba”斷詞系統針對文本的內容先進行斷詞，我們利用 TrieTree 的結構去生成句子中所有可能成為詞的情況，並使用動態規劃的方式找出最大機率的路徑。整個前處理的過程可參考表 2。

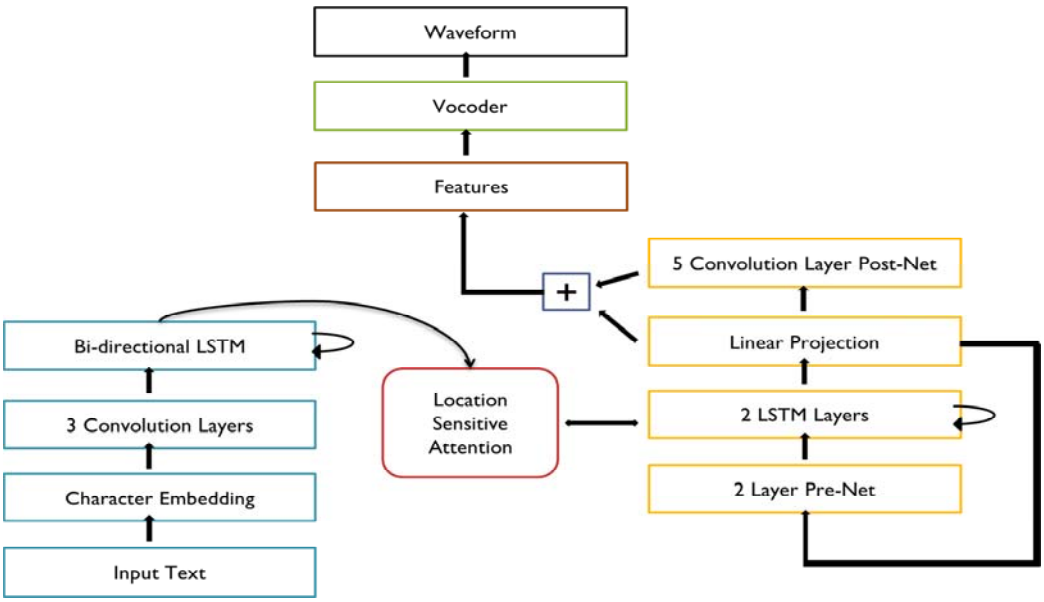

圖 1. 預測網路之模型架構
[Figure 1. The model architecture of the predicted network]

表 2. 資料前處理
[Table 2. Data preprocessing]

原始文本	可想而知，甕中捉鱉顯然比亡羊補牢更可靠更有效。
經過 jieba	可想而知|，|甕中捉鱉|顯然|比|亡羊補牢|更|可靠|更|有效|。
經過 pypinyin	ke3 xiang3 er2 zhi1|，| weng4 zhong1 zhuo1 bie1	xian3 ran2	bi3	wang2 yang2 bu3 lao2	geng4| ke3 kao4	geng4| you3 xiao4 |.

2.3 聲碼器 (Vocoder)

Tacotron2 (Shen *et al*., 2017)預設的聲碼器為 WaveNet (van den Oord *et al*., 2016)，是使用 Auto-regression 的方式生成音頻，即每在預測當前時刻的值時都是根據前一時刻的輸出結果。模型架構主要是由因果卷積(Causal Convolution)組成，而為了能在時域上獲得更廣的感知能力，模型中加入了擴大卷積(Dilated Causal Convolution) (Yu & Koltun, 2015)，如圖 2，當層數疊加，感知能力就以指數性成長。雖然這樣的模型架構能夠重現極為逼真的人類語音，也在語音合成上達到了很好的效果，但美中不足的卻是其生成速率。根據我們的評估，得花費數十秒的合成時間才能生成一秒鐘的音頻，若要作為實際應用，尚有大幅度的調整空間。而為了使系統能夠即時合成，我們根據了 Tacotron2 (Shen *et al*., 2017)的前端預測網路，並針對多種不同的聲碼器進行實測、探討。

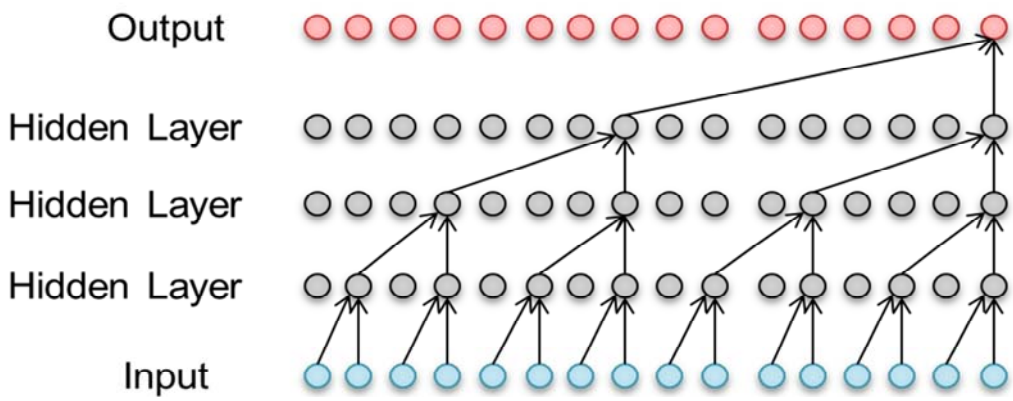

圖 2. Dilated Causal Convolution
[Figure 2. Dilated Causal Convolution]

2.3.1 Griffin-Lim

Griffin-Lim (Griffin & Lim, 1984) 是一種迭代的演算法，音頻質量雖然不如 WaveNet (van den Oord *et al.*, 2016)，但在即時系統中仍保有了競爭力，是許多即時語音合成系統比較的對象。透過迭代的次數來提升合成的音質，我們的實驗中採用了六十次的迭代以確保其穩定性。比起傳統聲碼器需要基頻和倒譜等參數而言，Griffin-Lim (Griffin & Lim, 1984) 可根據文本預測的線性頻譜圖直接重建時域波形。

2.3.2 World-Vocoder

語音是聲音的一種，是由人的發聲器官發出，具有一定語法和意義的聲音。大腦對發音器官發出運動神經指令，控制發音器官各種肌肉運動從而振動空氣而形成。整體發聲過程是空氣由肺進入喉部，經過聲帶激勵，進入聲道，最後通過嘴唇輻射形成語音。World-Vocoder (MORISE, YOKOMORI & OZAWA, 2016) 參照了此發聲原理，分別將三種聲學特徵:基頻(Fundamental Frequency)、頻譜包絡(Spectral envelope)以及非週期序列(Aperiodic parameter)對應到了人發聲機理的經典源-濾波器(source-filter)模型，最後並利用這些參數重建語音，如圖 3、圖 4。而在特徵提取的過程，我們先是透過 DIO 演算法提取出基頻，然後利用基頻中的 CheapTrick 提取包絡，最後透過 D4C 將得到後的基頻與包絡計算出非週期信號。不過為了與 Tacotron2 (Shen *et al.*, 2017)訓練做結合，這種高維度的頻譜包絡以及非週期信號我們必須先將其降維，以緩解神經網路訓練時所帶來的壓力。而透過 Merlin 工具包可幫助實現維度的轉換。以我們的實驗為例，首先將提取到的 MFCC 降至 60 維，接著將非週期序列轉變成 Band 非週期信號，此步驟可有效將維度降至 5 維，基頻部分則保持不變，最後再將上述特徵維度連接起來，輸入 Tacotron2 (Shen *et al.*, 2017)模型進而實現神經網路之訓練。

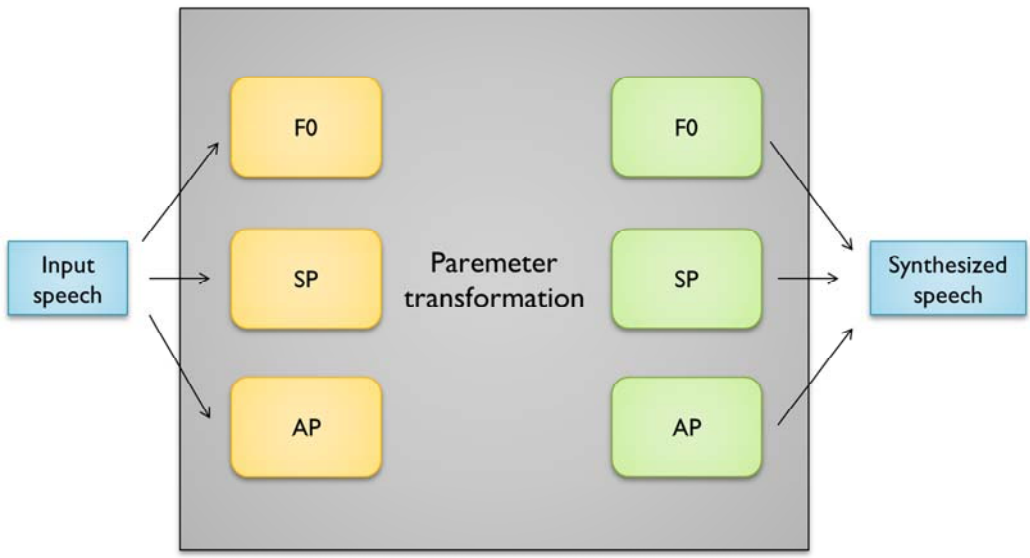

圖 3. World-Vocoder 透過 F0 、SP 、AP 重建語音信號
[Figure 3. World-Vocoder reconstructs voice signals through F0, SP, AP]

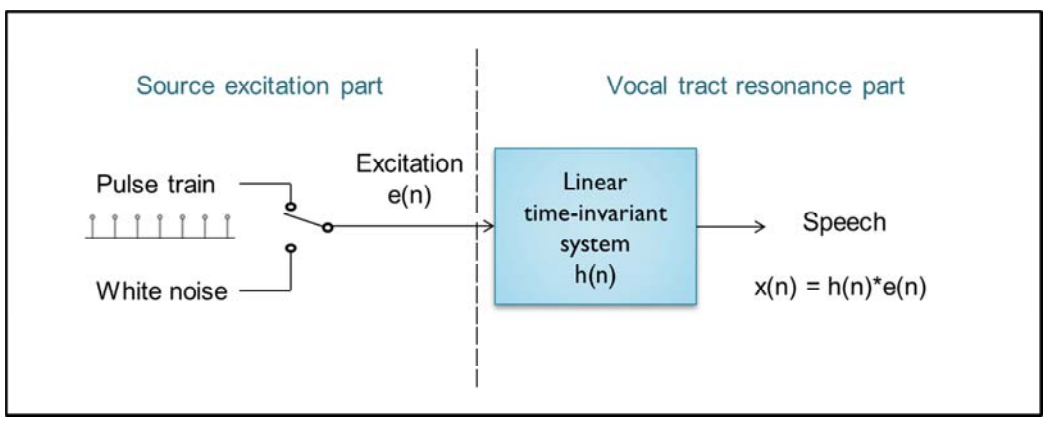

圖 4. 人發聲機理的經典源-濾波器模型
F0 、AP 、SP 分別對應到脈衝序列(Pulse train) 、非週期序列(White noise) 以及聲道諧振部分的 h(n)
[Figure 4. Classic source-filter model F0, AP and SP correspond to pulse train, white noise, and h(n) of the Vocal tract resonance part, respectively.

2.3.3 WaveGlow

隨著神經網路的發展，目前常使用到的生成模型可有:自回歸模型(Autoregressive model)、生成對抗網路(GAN)以及基於流的生成模型(Flow-based generative model) (Prenger, Valle & Catanzaro, 2018)。儘管自回歸模型在許多實驗上得到了很好的效果，但這種一次生成

一個樣本的生成方式除了需要龐大的計算資源之外，在可平行性上也受到了限制。相對而言，生成對抗網路則免除了這種煩惱，主要透過生成器與判別器不斷相互學習的迭代從而生得與真實樣本接近的分布。但一直以來生成對抗網路也不免會遇到許多問題，像是生成的多樣性不足以及訓練過程不穩定等等。不過幸運地，基於流的生成模型有效的解決了這些問題，而這樣的生成方式，也被採用在聲碼器-WaveGlow (Prenger *et al*., 2018)中。

WaveGlow (Prenger *et al*., 2018)透過分布採樣生成語音，僅需一個網路及一個最大化似然的損失函數即可生成時域波形，並且在高還原度的情況下亦能即時的合成語音。利用多個可逆的變換函數組成序列，將一個簡單的分布透過一系列的可逆函數轉換到一個複雜的分布，並藉此來模擬訓練數據的分布，最後再透過最大似然準則來進行優化。

我們使用的網路架構比照了(Prenger *et al.*, 2018)中的配置，包含 12 層的對耦映射層、12 個 1*1 的可逆卷積以及 WN 中設有 8 層的 dilated convolutions，同時根據我們的資料集調整了超參數。訓練資料為 48kHz 的音頻，我們將 160 維的梅爾頻譜作為輸入以及 FFT_size、hop_size、window_size 都設定了相符的格式以便訓練，如表 3 所示。

表 3. Biaobei 資料集超參數設定
[Table 3. Biaobei dataset hyperparameter setting]

資料集 ：Biaobei	
sample_rate(Hz)	48k
num_mels	160
FFT_size	4096
hop_size	600
window_size	2400

3. 實驗結果 (Experimental Results)

3.1 音頻質量 (Mean Opinion Score Tests)

在模型訓練完畢之後，我們從資料集裡面隨機選取了五句與訓練資料不重複的文本進行評估，受測人員一共十位。受測準則如下:每人聽測五種不同句子，每種句子各包含四個音檔，分別來自真實數據以及 World-Vocoder (MORISE *et al*., 2016)、Griffin-Lim (Griffin & Lim, 1984)、WaveGlow (Prenger *et al*., 2018)三種不同聲碼器合成的音檔。在每一句聽完後，都給予一到五分的主觀分數，總計後再平均計算。而整個過程共包含五種句子以及二十個不重複的音檔，測試結果如表 4。

表 4. Mean Opinion Scores
[Table 4. Mean Opinion Scores]

Model	Mean Opinion Score(MOS)
World-Vocoder	2.71
Griffin-Lim	3.15
WaveGlow	4.08
Ground Truth	4.41

3.2 推斷速度 Speed of Inference

一個高質量的音頻至少需要擁有 16kHz 的採樣點。而我們的實驗從前端預測網路到後端生成語音不僅都符合了標準，甚至都展現了比實時合成還要快的速度。訓練完的模型我們統一放到了 GeForce RTX 2080 TI GPU 上進行推測。以合成一個十秒且 48kHz 的音頻來說，我們分別在 World-Vocoder (MORISE *et al*., 2016) : 6 秒，Griffin-Lim (Griffin & Lim, 1984) : 1.2 秒，WaveGlow (Prenger *et al*., 2018) : 1.4 秒的時間內完成了推斷。另外，我們也整理了 Tacotron2 (Shen *et al*., 2017)預測網路搭配不同聲碼器在同一台機器上分別所佔用的資源，雖然 WaveGlow (Prenger *et al*., 2018) 在推測時間上展現了優異的合成速度，但由表 5 可看出其所佔據的資源也相當高，說明了我們的模型仍有優化的可能。

表 5. Tacotron2 結合不同聲碼器所佔用的計算資源
[Table 5. Tacotron2 combines computing resources used by different vocoders]

Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz/64GB/RTX 2080TI	World-Vocoder	Griffin-Lim	WaveGlow
GPU 使用量(GB)	0.93G	1.31G	2.5G
CPU 使用率(%)	6.75%	7.5%	14.3%
MEM 使用率(%)	2.45%	3.25%	5%

4. 結論 (Conclusions)

我們的研究目前在聲碼器上嘗試了多種可能，包含:World-Vocoder (MORISE *et al*., 2016)、Griffin-Lim (Griffin & Lim, 1984) 以及 WaveGlow (Prenger *et al*., 2018) ，並將這些合成技術都套用在我們的預測網路上。從我們的研究來看，我們發現儘管 World-Vocoder 及 Griffin-Lim 都已開發一段時間，但在音頻的還原度上仍遠不及近期興起的神經網路式合成器，且 WaveGlow (Prenger *et al*., 2018)不僅在音質的還原度亦或是合成的速度上(在 2080TI 上，一秒約莫可生成 350kHz 以上的採樣點)都給予了我們不錯的展示。但就長期而言，我們的模型還有多種可能的優化方式，像是我們使用的中文資料集規模較小，儘管經過了斷詞系統的調整，仍有部分語句無法良好的呈現人類的自然語音。日後除了收集更完整的語料庫之外，在預測網路部分加入情緒辨識作為條件，使得合成的音頻更生動更有溫度也是我們的任務之一。

參考文獻 References

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In arXiv preprint arXiv:1506.07503v1.

Griffin, D. & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236-243. doi: 10.1109/TASSP.1984.1164317

MORISE, M., YOKOMORI, F., & OZAWA, K. (2016).WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. on Information and Systems*, *E99.D*(7), 1877-1884. doi: 10.1587/transinf.2015EDP7457

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., …Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. In arXiv preprint arXiv:1609.03499v2.

Prenger, R., Valle, R., & Catanzaro, B. (2018). WaveGlow: A Flow-based Generative Network for Speech Synthesis. In arXiv preprint arXiv:1811.00002v1.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., …Wu, Y. (2017). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In arXiv preprint arXiv:1712.05884.

Yu, F. & Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. In arXiv preprint arXiv:1511.07122v3.