The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

整合語者嵌入向量與後置濾波器於提升個人化合成語音之語者相似度
Incorporating speaker embedding and post-filter network for improving speakersimilarity of personalized speech synthesis system

Sheng-Yao Wang, Yi-Chin Huang
National Pingtung University
mike456852@gmail.com, ychuangnptu@mail.nptu.edu.tw

摘要

近年來在語音合成的研究之中，單一語者的合成系統已經有著高音質的表現，但對於多語者系統來說，合成語音的品質與語者相似度仍是一大挑戰，本研究針對合成語音的品質與語者相似度，透過兩個面向來建立出一套可合成多語者之文字轉語音系統，首先針對於多語者的議題，並盡量達成透過少量樣本 (zero-shot) 來達成語者轉換，我們透過探討兩類語音嵌入向量 (speaker embedding) 對於多語者語音合成系統較爲合適，我們比較了用於語者辨識 (speaker verification) 以及單純用於語音轉換 (voice conversion) 的語者嵌入向量。接著，爲了提升合成的語者相似度以及語音品質，我們嘗試置換類神經網路架構中，作爲提升頻譜的 post-net 的部分，在此處我們使用了一個後置濾波器的網路來取代，且比較和 post-net 所產生的頻譜差異以及探討其模型參數量之差異性。實驗結果表明，透過疊加性注意力機制來整合語者嵌入向量進入到類神經網路架構的語音合成系統的確能夠有效地產生具有目標語者的合成語音，並且在加入後置濾波器網路後能夠比傳統透過 post-net 的方式來強化合成語音的語者特性以及合成語音的語音品質，且合成一般長度語音句的時間約爲 2 秒鐘，已接近即時合成個人化語音之成果。未來將進一步探討如何產生可控制語速或情緒之個人化語音。

Abstract

In recent years, speech synthesis system can generate speech with high speech quality. However, multi-speaker text-to-speech (TTS) system still require large amount of speech data for each target speaker. In this study, we would like to construct a multi-speaker TTS system by incorporating two sub modules into artificial neural network-based speech synthesis system to alleviate this problem. First module is to add speaker embedding into encoding module for generating speech while a large amount of the speech data from target speaker is not necessary. For speaker embedding method, in our study, two main speaker embedding methods, namely speaker verification embedding and voice conversion embedding, are compared to deciding which one is suitable for our personalized TTS system. Second, we substituted the conventional post-net module, which is adopted to enhance the output spectrum sequence, to further improving the speech quality of the generated speech utterance. Here, a post-filter network is used. Finally, experiment results showed that the speaker embedding is useful by adding it into encoding module and the resultant speech utterance indeed perceived as the target speaker. Also, the post-filter network not only improving the speech quality and also enhancing the speaker similarity of the generated speech utterances. The constructed TTS system can generate a speech utterance of the target speaker in fewer than 2 seconds. In the future, we would like to further investigate the controllability of the speaking rate or perceived emotion state of the generated speech.

關鍵字：多語者語音合成、語音轉換、語者識別、少量樣本、後置濾波器

Keywords: Multi-speaker Text-to-Speech, Voice Conversion, Speaker Verification, Zeor-shot, Post-Filter

1 緒論

就單一語者的語音合成技術來看，其合成技術已經能夠合成出逼眞且自然的語音，並且不需要太多的語音數據及訓練時間，而爲了擴展到其他語者，常見的方法有語音轉換和模型自適應兩種方法，語音轉換透過更換不同的語者訊息來達成目標，有基於 GAN 的 StarGAN (Kameoka et al., 2018) 和 CyCleGAN (Kaneko et al., 2020) 等方法，也有基

於 AutoEncoder 的 AdaIN-VC (Chou et al., 2019) 和 AutoVC (Qian et al., 2019) 等方法，它們都有相當不錯的效果，唯一個侷限就是僅能更換語者而不能更改內容；而模型自適應主要是在 TTS 系統中加入 Speaker ID Table 來使模型能夠依照 Speaker ID 生成不同語者的聲音，它既能更換內容也能更換語者，但是需要大量不同語者的語音數據以及較多的訓練時間來達成目標，且無法擴展到沒看過的語者，因此，有 (Chien et al., 2021) 和 (Jia et al., 2018) 等研究，將語音轉換與模型自適應兩種方法結合，或是引入語者辨識取代模型自適應裡的語者編號以此擴展到沒看過的語者。

近年來各種語音合成的神經網路模型被提出，但是自回歸模型的 Tacotron 2(Shen et al., 2018) 仍然有著很大的討論空間，它將文字轉換爲一序列的 Embedding 表示，並對每個音框的梅爾頻譜建立注意力對齊，這使每個文字與特定時間單位的頻譜建立一種映射關係，上一個音框推測出下一個音框使得生成的頻譜更有連續性。FastSpeech 2(Ren et al., 2020) 或是 Transformer TTS(Vaswani et al., 2017) 等非自回歸模型，因爲加入了 Self-Attention，使得文字或梅爾頻譜間有著跨時間單位的連接，這減輕了 Tacotron 2 使用 LSTM 作爲文字編碼輸出層或是頻譜解碼輸出層因爲序列長度所帶來的記憶門壓力，如 (Wang et al., 2019) 所述。

Tacotron 2 使用 Local Sensitive Attention 作爲文字與梅爾頻譜間建立映射關係的方法，對於現在的 TTS 系統來說，訓練速度慢且對於較長的語句可能會發生漏字、重複發音等現象，該問題已由 Google 使用 Dynamic Convolution Attention 改善 (Battenberg et al., 2020)。也有如 Glow TTS(Kim et al., 2020) 使用單調函數來限制 Attention 只能向前對齊的方式來解決訓練速度慢以及重複發音的問題。

最後，Tacotron 2 經過文字轉換成序列、序列與頻譜建立映射關係，再透過 LSTM 解碼輸出之後，還會經過 Post-net 層，其目的是要解決解碼輸出後的頻譜過於平滑的問題，現在很多 TTS 架構的 Post-Net 都是採用該架構。

鑑於 Tacotron 2 有著明確的合成步驟，使得許多研究都在其系統上完成，例如多語者 TTS 系統。目前合成語音品質較好的多語者系統是在 Tacotron 2 內部建立 Speaker ID Table，但它無法擴展到沒看過的語者，於是語者識別和語音轉換的方法被加入到語音合成系統裡。

在本研究中，我們提出了一個多語者語音合成的系統，它基於 Tacotron 2 的架構並加入語者驗證及語音轉換的方法以擴展到沒看過的語者，也加入了 Self-Attention 減輕訊息長距離的傳播所造成的負擔，最重要的是我們引入了一種後處理的方式，使我們的系統有著良好的語音品質以及語者相似度，我們也期望透過該後處理方式減輕多語者系統所需要的大量語音數據以及訓練時間。

接下來，我們在第二章節說明本研究所提出的多語者語音合成系統，並於第三章節說明實驗過程與結果，第四章則是說明我們的結論。

2 提出的系統架構

在此章節，我們將多語者語音合成分成四個部份:

- Speaker Embedding：該部份我們比較語音轉換及語者驗證所提取的 Speaker Embedding 何者對於我們的模型較有幫助。
- Text Encoding：該部份目的是將中文字轉換成一種 Embedding 表示，其輸出作爲 Decoding 的輸入。
- Decoding：該部份目的是將 Text Encoding 的輸出與梅爾頻譜建立映射關係，其輸出爲梅爾頻譜。
- Post-Net：該部份目的是增強 Decoding 輸出的梅爾頻譜的特性。

我們的系統運作的順序如 Figure 1。

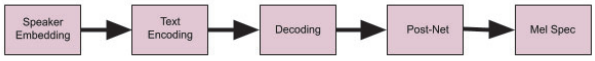

Figure 1: 系統運作順序

2.1 Speaker Embedding

我們引入語音轉換和語者驗證的方式來取代多語者 Tacotron 2 中的 Speaker ID Table，以便我們能擴展到沒看過的語者。

語音轉換的模型採用 (Chou et al., 2019) 的架構，其架構是一種 Variational AutoEncoder，它能夠將來源語音分解成語者編碼跟內容編碼，透過更換語者編碼的方式來達到語音轉換的效果，其模型能夠用於資料外的語者，這正是我們所考慮的。

語者驗證的模型採用 (Cooper et al., 2020) 的架構，其架構是一種 ResNet34(He et al.,

2016) 的改進，該模型於語者驗證中的性能可與 X-Vector(Snyder et al., 2018) 相比，在多語者語音合成系統中，是優於 X-Vector 的，因此我們選擇使用它。

2.2 Text Encoding

我們將輸入的中文字透過 pypinyin 轉換爲羅馬拼音作爲輸入，一樣經過 3 層 Conv Layer 和雙向 LSTM 作爲 Encoding，我們在這邊做了一個改動，爲了減輕 LSTM 必須依照順序傳播訊息所產生的高負擔，我們實現了 (Cooper et al., 2020) 與 (Wang et al., 2019) 等人提出的改進，在雙向 LSTM 輸出後降維並加入 Self-Attention 作爲另一個 Encoding 輸出，透過 Self-Attention 可以連接遠處的訊息狀態這點特性，有效的減輕 LSTM 的負擔，且在 Decoding 的部份能夠更快的與梅爾頻譜建立映射關係。

因此，Text Encoding 將會有兩個輸出，我們分別取名爲 Text Information 及 Long-distance Text Information，同時這兩個個輸出都會與語音轉換所提取的語者嵌入向量串接，架構如 Figure 2所示：

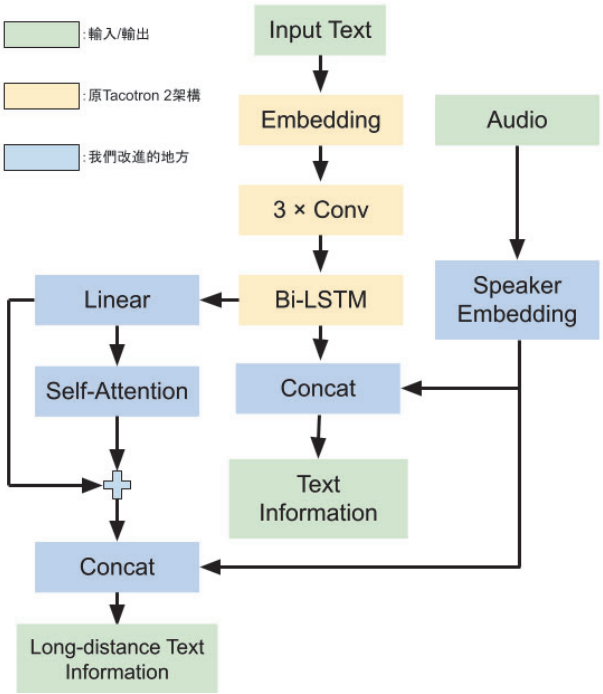

Figure 2: Text Encoding 架構

2.3 Decoding

由於 Text Encoding 層有兩個輸出，因此我們引入了兩個 Attention 機制：Dynamic Convolution(DCA) 與 Bahdanau Attention。

Text Information 使用 DCA 與梅爾頻譜建立映射關係，改善原本 Tacotron 2 因舊有的 Attention 導致漏字、重複發音現象，且 DCA 訓練速度上也比較快，如 (Battenberg et al., 2020) 所述；而 Long-distance Text Information 則使用 Bahdanau Attention 與梅爾頻譜建立映射關係，這是早期的一種 Additive Attention(Bahdanau et al., 2014)，是 Local Sensitive Attention 的簡化版，使用原因是爲了讓 Self-Attention 所學得的長距離訊息能夠簡單幫助 DCA 快速建立映射關係，該 Attention 無須與梅爾頻譜建立良好的映射關係。

另外，爲了加強語者嵌入向量的效果，這邊同樣實現 (Cooper et al., 2020) 和 (Wang et al., 2019) 等人提出的改進，在梅爾頻譜通過 Pre-Net 層之前與語者嵌入向量相加，這能夠有效的使語者嵌入向量影響梅爾頻譜。

以及，最後一層 LSTM 解碼輸出的時候，同樣引入 Self-Attention 幫助 LSTM 減輕訊息傳播的負擔，架構如 Figure 3所示：

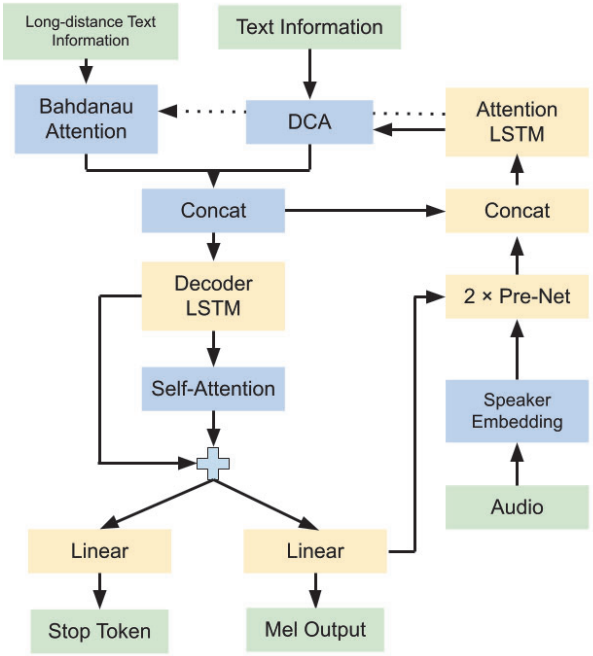

Figure 3: Decoding 架構

2.4 Post-Net

原本的 Post-Net 是爲了改善頻譜過度平滑導致生成的聲音過於低沉，但我們研究發現該 Post-Net 架構仍無法有效解決頻譜平滑的問題，我們受到 (Kaneko et al., 2017a) 和 (Kaneko et al., 2017b) 的啓發，設置了一個 Post-Filter，透過添加噪音的方式補齊與目標頻譜間的差距，我們發現這種方式增強了合成聲音的品質之外，還增強了語者的相似度，使得生成音檔更接近目標語者，並且，這種修改

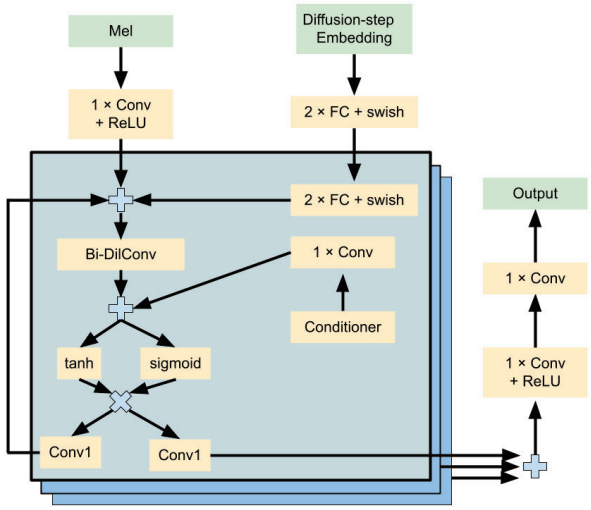

Figure 4: Diffwave 架構

並不會使模型參數有所增加。

爲了改善 Tacotron 2 的 Post-Net，我們使用 Diffwave(Kong et al., 2020) 模型作爲一個 Post-Filter 取代 Post-Net。

Diffwave 是基於降噪機率擴散 (Ho et al., 2020) 所提出的一種聲碼器模型，該擴散概率透過指定步驟的馬可夫鏈與可控條件逐漸將白噪音轉換爲波形訊號，其結構如 Figure 4所示。

我們將輸入改爲與 Decoding 輸出的頻譜相同形狀的噪音，可控條件爲 Decoding 輸出的頻譜，透過指定的步驟來將噪音逐步消除，這裡值得注意的是，與聲碼器需要生成出乾淨的波形訊號不同，我們適當減少了步驟的次數，期望該 Post-Filter 生成的頻譜保留部份噪音來減少與眞實頻譜的差距。

3 實驗過程與結果

3.1 實驗設置

我們使用 AISHELL-3(Shi et al., 2020) 作爲本次實驗的中文語料庫，該語料庫具有 218 位語者，取 173 位語者每位語者 100 個音檔作訓練集（語者數約佔整體 75%，其餘語者作爲 Unseen 測試模型性能。），並將所有音檔下採樣至 22050Hz，並取出 80 維梅爾頻譜作爲訓練 Tacotron 2 與語音轉換的輸入，其音框長度爲 1024 個採樣點，音框移動的距離爲 256 個採樣點，音高的頻率範圍爲 20Hz 至 8000Hz。

爲了比較我們所提出的方法優劣，我們訓練了四個多語者 Tacotron 2，分別爲：

- Tacotron 2 + 語者驗證
- Tacotron 2 + 語音轉換
- Tacotron 2 + 語者驗證 + Self-Attention + Post-Filter (Our propose A)
- Tacotron 2 + 語音轉換 + Self-Attention + Post-Filter (Our propose B)

另外，爲了方便撰寫，在後續的內文中如果沒有提起語者驗證或語音轉換，則代表兩者性能是一致的。

3.2 實驗過程

我們使用 AISHELL-3 的語料庫來訓練語音轉換與語者驗證的模型，其訓練設置皆依照 (Chou et al., 2019) 和 (Cooper et al., 2020) 所提供，取出的語者嵌入向量維度數皆設置 128，且我們將各個語者提取的語者嵌入向量取平均，依照 AISHELL-3 的語者數我們共提取出 218 個語者嵌入向量。

再來是多語者 Tacotron 2，在 Text Encoding 的部份，我們將語者嵌入向量做相同維度的 Linear 與 ReLU，使它與我們的模型匹配，另外，我們在通過 Bi-LSTM 之後還添加了 Linear + Self-Attention 作爲另一個輸出，其輸出維度爲 128。

Decoding 的部份只有 Pre-Net 需要注意，我們將上一個音框降維度至 256，語者嵌入向量則升維至 256 並以 Softsign 激活後才做相加，其輸出才通過 Pre-Net。

最後的 Post-Net 部份，我們將 Decoding 輸出的頻譜直接通過 Diffwave Post-Filter 取代了原本 Tacotron 2 的 Post-Net，並且其輸出結果不需要再做殘差連接。

Diffwave Post-Filter 的模型是於 Tacotron 2 訓練完後才訓練的，因爲我們需要 Tacotron 2 的輸出作爲 Post-Filter 的訓練集。

我們在下方的 Table 1 中提供了模型的訓練時間作爲參考，由於多語者 TTS 僅訓練 72 個小時，因爲訓練時間不夠長的關係導致其輸出音質還未達到最佳狀態，但我們透過 Post-Filter 增強了音質與語者相似度，可以在下方連結試聽我們的樣本[1]。

Model	Batch Size	Total Step
語音轉換	32	1M
語者驗證	32	1M
Tacotron 2	64	99k
Our propose A/B	64	99k
Post-Filter	16	320k

Table 1: 模型訓練參數

[1] https://babaili.github.io/tacotron2_samples/

3.3 實驗結果

本來於上方有提到我們訓練了四個系統，但由於 Tacotron 2 沒有 Self-Attention 的幫助，訓練了 72 小時仍未有良好的對齊線，因此在下方僅討論 Our propose A/B 的結果。

我們比較了原音檔與 Our propose A/B 的頻譜，請看 Figure 6，依圖像來看，我們可以發現 Post-Filter 距離原音檔的頻譜仍有一大段距離，最明顯變化的地方是雜訊的部份多了橫向的雜訊。我們將上述頻譜拿去做語者相似度分析，採用 Resemblyzer 分析器[2]來投影語者空間，該分析器是 (Wan et al., 2018) 的實現，其結果如 Figure 5，由於 AISHELL-3 大部分語者的性別是女性（佔整體語料庫 80%）的關係，所以對於合成男性聲音區分度沒有如女性一樣明顯。

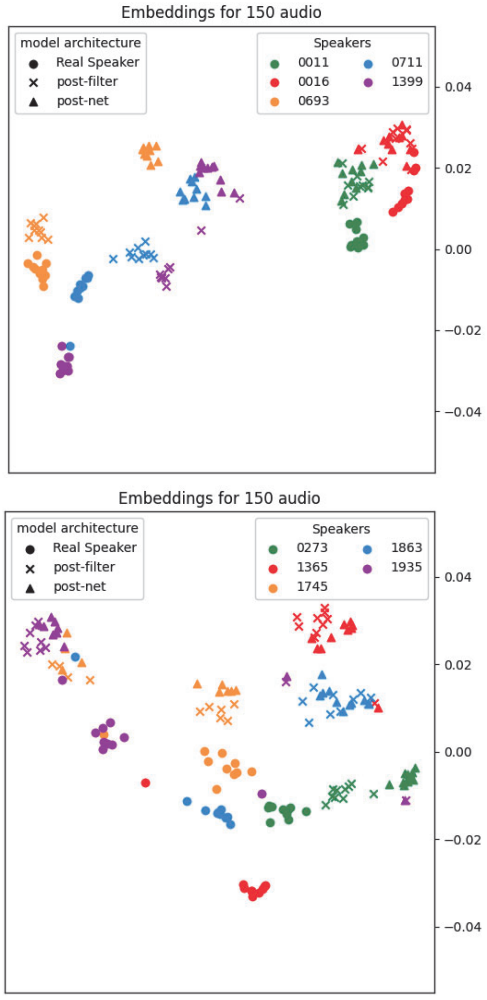

Figure 5: 透過 Resemblyzer 分析，上圖爲女性，下圖爲男性。

接著我們使用客觀評測與主觀評測來證實實驗結果，首先，使用 Mel Cepstral Distortions (MCD) 作爲客觀評測的方法，我們隨機取 5 個語者各 10 個眞實音檔，共 50 個眞實音檔，並用四個系統分別合成對應的 50 個音檔，總共 250 個音檔，接著將音檔用 World 分析器取出頻譜資訊，再透過 SPTK 工具轉換成梅爾廣義倒頻譜後才計算 MCD，結果如 Table 2。

Model	MCD
VC-Post-Filter	9.62±0.42
VC-Post-Net	10.16±0.53
SV-Post-Filter	9.29±0.85
SV-Post-Net	10.29±0.72

Table 2: MCD 客觀評測，數值越低越好。VC: 語音轉換，SV: 語者驗證

Model	Quality
VC-Post-Net	2.67±0.35
VC-Post-Filter	3.75±0.71

Table 3: MOS 主觀評測，比較 Post-Net 和 Post-Filter 音質。

再來是使用 Mean Opinion Score(MOS) 作爲主觀評測的方法，我們針對語音轉換和語者辨識做音質的比較，兩者皆用 Post-Net 進行 MOS 主觀評測，我們在語音轉換上獲得了 2.67 的分數，在語者辨識上獲得了 2.54 的分數，我們發現語音轉換比語者驗證還來得好，於是我們接著比較語音轉換在 Post-Net 和 Post-Filter 的音質差異，其結果如 Table 3，接著我們對語者相似度進一步的比較，其結果如 Table 4。

我們也提供 Tacotron 2、Post-Net 及 Post-Filter 的模型參數，如 Table 5，也大略測試了生成頻譜、Post-Filter 推論和頻譜轉換成波形的時間，如 Table 6。

實驗結果證明，Post-Filter 確實可以取代 Tacotron 2 的 Post-Net，我們可以看到 Post-Filter 的參數甚至比 Post-Net 還小，這樣更換所付出的成本僅僅是多了 6 個小時的訓練時間。在推理合成音頻上，因爲 Diffwave 模型是透過大量迴圈來進行降噪的，它的推理時間比 Tacotron 2 還多了大概 0.2 秒，這樣的付出我們認爲是值得的，如同我們所展示的音頻，Post-Filter 對於音質和語者相似度的提昇是可見的。

4 結論

本研究提出了一種多語者的 Tacotron 2 架構，修改了 Encoding、Decoding 層的架構，使得語者嵌入向量的效果更加明顯，同時，我們

[2] https://github.com/resemble-ai/Resemblyzer

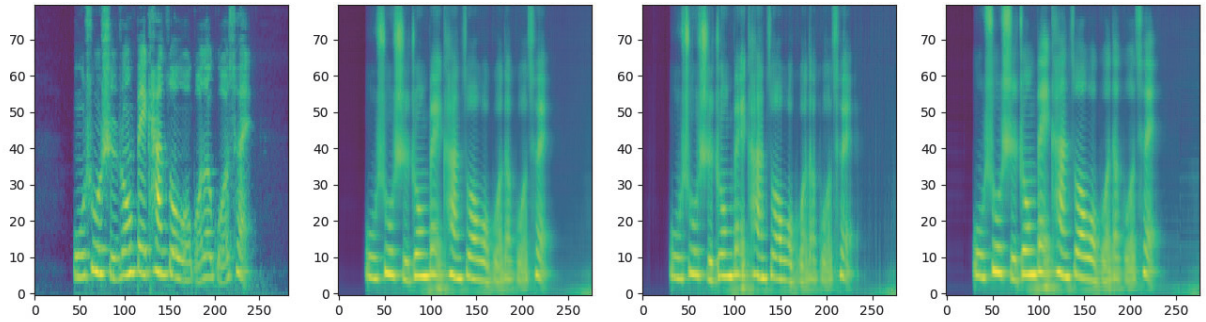

Figure 6: 第 1 張：原始頻譜、第 2 張：Tacotron 2 Linear、第 3 張：Tacotron 2 Post-net、第 4 張：Tacotron 2 Post-Filter，以『武術始終被看作我國的國粹』爲例。

Model	Similarity
VC-Post-Net	2.70±0.41
VC-Post-Filter	3.75±0.71
SV-Post-Net	2.31±0.18
SV-Post-Filter	3.51±0.32

Table 4: MOS 主觀評測，VC 和 SV 的語者相似度

Model	Parameters
Baseline	29M
Our propose	45M
Post-Net	4M
Diffwave Post-Filter	2M

Table 5: 模型參數量

Model	Time
Text Encoding + Decoding	0.4 s
Post-Filter	0.6 s
Diffwave Vocoder	0.9 s

Table 6: 模型推論所花費時間，以『武術始終被看作我國的國粹』爲例，總合成時間約爲 2 秒。

也使用 Diffwave 作爲 Post-Filter 取代原本 Tacotron 2 的 Post-net，我們實驗證明，這種取代不但能夠提昇合成的音質，還加強了語者的特性，使得合成出來的頻譜與目標語者更爲接近，且模型的參數也不會因爲這種改動而大幅度增加。

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. 2020. Location-relative attention mechanisms for robust long-form speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198. IEEE.

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. 2021. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592. IEEE.

Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6184–6188. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*.

Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*.

Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE.

Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino. 2017a. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4910–4914. IEEE.

Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2020. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. *arXiv preprint arXiv:2010.11672*.

Takuhiro Kaneko, Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi. 2017b. Generative adversarial network-based postfilter for stft spectrograms. In *Interspeech*, pages 3389–3393.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *arXiv preprint arXiv:2005.11129*.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Xin Wang Wang, Junichi Yamagishi Yamagishi, Yusuke Yasuda Yasuda, and Shinji Takaki Takaki. 2019. Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language.