Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 49-66

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 Speaker Similarity of Personalized Speech Synthesis System

王聖堯*、黃奕欽*

Sheng-Yao Wang, Yi-Chin Huang

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品質的表現，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兩個議題來建立出一套可合成多語者之文字轉語音系統，首先針對多語者的議題中，目標為透過少量樣本(Zero-Shot)來達成語者轉換，我們透過語者嵌入向量(Speaker Embedding)的引入來實作多語者語音合成系統，並比較針對不同任務所建立的語者嵌入向量的效果差異。在此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，為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，作為提升頻譜的 Post-Net 的部分，在此處我們使用了一個後置濾波器(Post-Filter)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語音品質，且合成一般長度語音句的時間約為 2 秒鐘，已接近即時合成個人化語音之成

* 國立屏東大學資訊科學研究所
Department of Computer Science, National Pingtung University
E-mail: mike456852@gmail.com; ychuangnptu@mail.nptu.edu.tw

果。未來的研究方向會加入更多資訊來幫助語者嵌入向量在 TTS 的效能上改進。

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 the speaker embedding into encoding module of the end-to-end TTS framework while using small amount of the speech data of the training speakers. For speaker embedding method, in our study, two speaker embedding methods, namely speaker verification embedding and voice conversion embedding, are compared for deciding which one is suitable for the personalized TTS system. Besides, we substituted the conventional post-net module, which is conventionally adopted to enhance the output spectrum sequence, to a post-filter network, which is further improving the speech quality of the generated speech utterance. 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, other feature such as prosody information will be incorporated to help the TTS framework to improve the performance.

關鍵詞：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, Zero-Shot, Post-Filter

1. 緒論 (Introduction)

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真且自然的語音，並且不需要太多的語音數據及訓練時間，而為了擴展到其他語者，常見的方法有語音轉換和模型自適應兩種方法：

- 語音轉換： 透過更換不同語者訊息來達成目標，有基於 GAN 的 StarGAN-VC (Kameoka *et al.,* 2018) 和 CyCleGAN-VC (Kaneko *et al.,* 2018) 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，唯一的侷限就是僅能更換語者不能更改內容。

- 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 生成對應語者的聲音，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。

基於語音轉換和模型自適應在多語者 TTS 上的不足，於是有著 (Jia *et al.,* 2018) 和 (Chien *et al.,* 2021) 等研究，將語音轉換或語者辨識這兩種方法取代模型自適應中的 Speaker ID Table 來使模型擴展到沒看過的語者。在本次研究中，我們將比較分別使用語音轉換和語者辨識這兩種任務所設計的語者嵌入向量作為我們 TTS 系統中語者的表示方式，並比較何者對於我們提出的架構更合適。

我們的 TTS 架構是基於 Google 所提出的自回歸模型 Tacotron 2 (Shen *et al.,* 2018)，它由三個神經網路區塊組成，每個區塊都有明確的目的以便我們進行改動：

- 編碼器: 將輸入的文字編碼成一種潛在表示，通常為了使模型擴展到多語者，會將文字潛在表示與語者嵌入向量串接。
- 解碼器: 於訓練期間，將文字潛在表示與目標頻譜的每個音框建立注意力對齊 (Chorowski *et al.,* 2015)，於推論期間，依據當前音框與文字潛在表示推測出下一個音框的值，直至注意力機制對齊到停止符號 (例如：文字中的句點)為止。
- Post-Net: 提升整體頻譜的品質。

Tacotron 2 的模型架構如圖 1 所示：

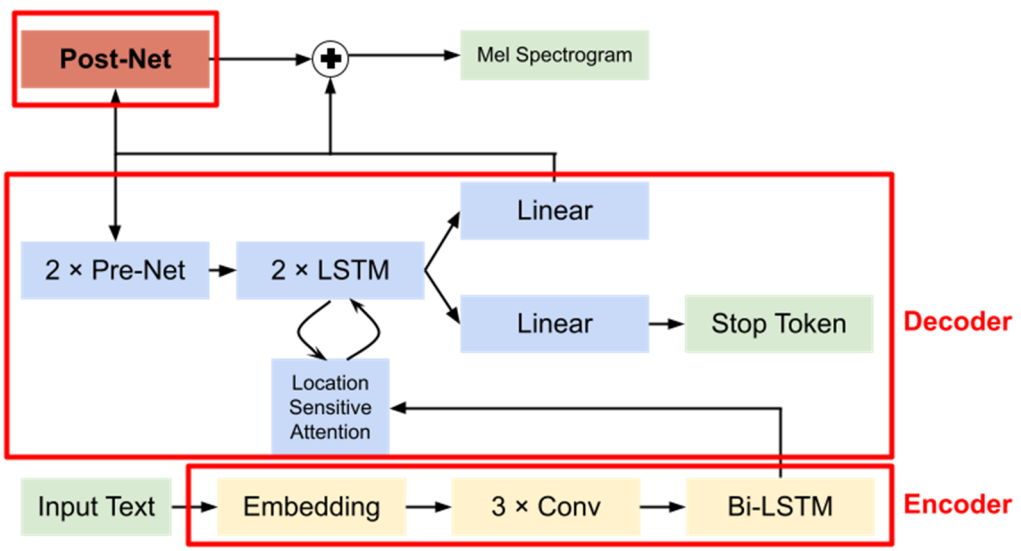

圖 1. Tacotron 2 模型架構
[Figure 1. Tacotron 2 model architecture]

Tacotron 2 整體架構對於目前神經網路的技術來說是相對舊的，隨著 Self-Attention (Vaswani *et al.,* 2017) 大量被運用於語音合成的任務上，改善了如 Tacotron 2 因使用 RNN 神經網路需要依照順序傳播的大量計算，如 Transformer TTS (Li *et al.,* 2019) 和 Fastspeech 2 (Ren *et al.,* 2020)；也有著各種注意力機制的方法被提出，以改善 Tacotron 2 舊有注意力機制訓練速度慢或是較長的句子會發生漏字或重複發音的問題，如 Forward Attention (Zhang *et al.,* 2018) 及 Dynamic Convolution Attention (Battenberg *et al.,* 2020)。因此，我們將運用近期的神經網路技術來更動 Tacotron 2 模型，期望模型訓練速度加快、合成語音品質的提升以及加強合成多語者語音的語者相似度。

我們將在第二章節闡述語者嵌入向量所用到的語音轉換及語者辨識模型，在第三章節闡述本次研究改動 Tacotron 2 的方法，第四章節闡述實驗結果，最後，在第五章節闡述本次研究的結論。

2. 語者嵌入向量 (Speaker Embedding)

2.1 語音轉換 (Voice Conversion)任務

在本次研究中，我們使用 AdaIN-VC 作為本次研究的語音轉換模型，雖然如之前所述，語音轉換的模型有很多種，但並不是都能提取出語者嵌入向量，如 StarGan 及 CycleGAN 等 GAN 模型雖然也是語音轉換，但它們是透過在訓練期間判別器 (Discriminator) 的約束，使得生成的語音接近內部語者，這無法提取出語者嵌入向量；屬於 AutoEncoder 模型的 AutoVC 也無法提取語者嵌入向量，因它是利用編碼層將語者訊息去除，並在解碼層加入 Speaker ID Table 來進行轉換的，而 AdaIN-VC (Adaptive Instance Normalization-Voice Conversion) 是一種將圖片風格轉換的技術套用到語音轉換上的 VAE 模型，它透過兩個編碼層將語音編碼成語者潛在表示及內容潛在表示，並透過解碼層組合兩者後生成轉換後的語音，我們可以藉由更換語者潛在表示來達到語音轉換的效果，其模型架構如圖 2 所示：

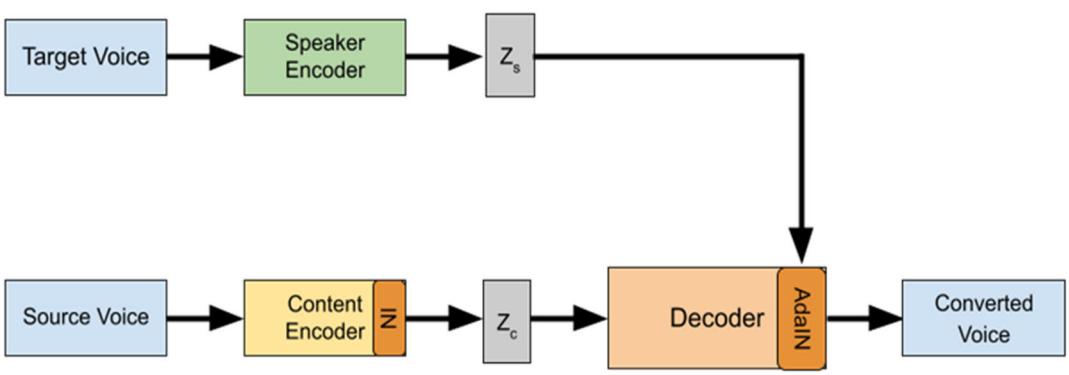

圖 2. AdaIN-VC 模型架構
[Figure 2. AdaIN-VC model architecture]

2.2 語者辨識 (Voice Verification)任務

我們使用 Learnable Dictionary Encoding (Cooper *et al.,* 2020) 簡稱 LDE，作為本次研究的語者辨識模型，它是基於 X-Vector (Snyder *et al.,* 2018) 所做的改進，並且在語者辨識的任務上以及多語者 TTS 系統上皆是優於 X-Vector 的。

X-Vector 的運作方式是將整個語音分成數個片段並透過數層卷積計算其輸出特徵，再將所有特徵取平均與標準差通過線性轉換來計算該語者的嵌入向量。LDE 與 X-Vector 不同的地方是 LDE 引入了數個 Dictionary Clusters，這些 Clusters 是需要透過神經網路去學習的，它們代表某些說話人的特徵，LDE 使 X-Vector 得到的輸出特徵與所有 Clusters 計算彼此差距的平均值與標準差來判斷該語音接近哪一個 Clusters，然後再進一步讓神經網路判斷該語者的嵌入向量，其模型架構如圖 3 所示：

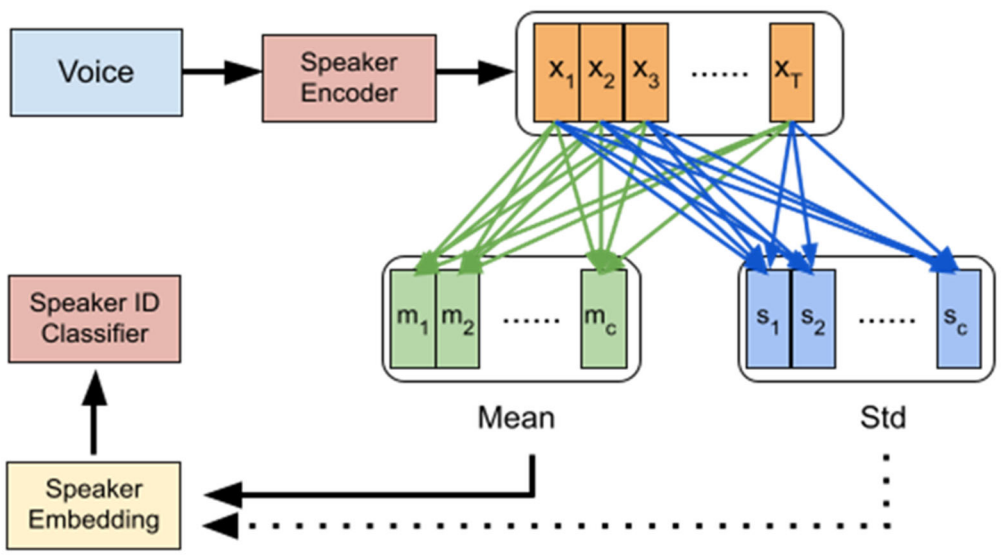

圖 3. LDE 模型架構
[Figure 3. LDE model architecture]

3. 研究方法 (Research Method)

3.1 編碼器 (Encoder)

基於原本的 Tacotron 2 架構，我們將 LSTM 的輸出降維降至 128 維並通過 Self-Attention 當作另一個輸出，Self-Attention 會將 LSTM 輸出的潛在表示進行全域相關性的連接，這些資訊將在解碼層幫助注意力機制更快的對齊，我們將原 LSTM 輸出稱為內容資訊 (Content Information)，另一個通過 Self-Attention 的輸出稱為長距離內容資訊 (Long-distance Content Information)，同時，為了使模型能夠合成多語者的語音，我們在這兩個潛在表示後方串接了語者嵌入向量，詳細架構如圖 4 所示：

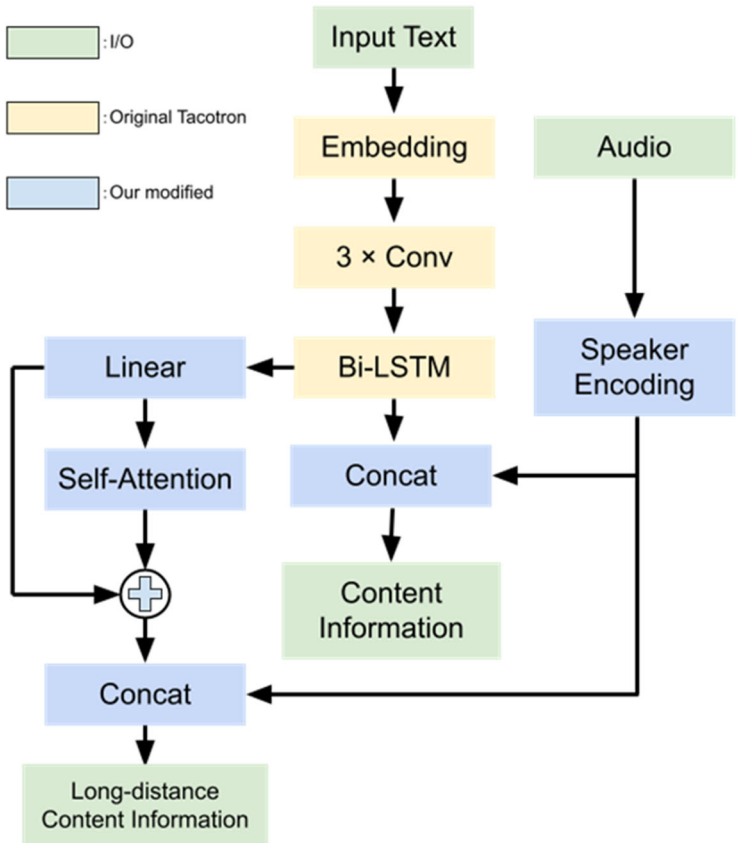

圖 4. 編碼器架構
[Figure 4. Encoder architecture]

3.2 解碼器 (Decoder)

解碼器部份我們做了較多的改動，首先，由於編碼層有兩個輸出，因此我們分別引入了兩個不同的注意力機制，我們為內容資訊引入了 Forward Attention 取代 Tacotron 2 舊有的注意力機制，它可以更快地引發對齊，並且能改善因長句所引發的重複發音或漏字的問題；長距離內容資訊則引入了 Bahdanau Attention (Bahdanau *et al.,* 2014)，它是一個傳統的 Additive Attention，因其架構較為簡單，可以快速地得到某些頻譜與文字的關係，這將能夠幫助 Forward Attention 更快地引發對齊，並且因為低維度的關係，它不會與 Forward Attention 競爭文字與頻譜間的對齊，在實驗結果會有更詳細地說明。

此外，為了加強語者嵌入向量對於模型的作用，我們在 Pre-Net 層加入了語者嵌入向量，透過神經網路的學習，能夠使模型更看重語者嵌入向量。

最後，在通過 LSTM 解碼後，我們又再一次引入 Self-Attention 將頻譜潛在表示的資訊進行全域相關性的連接，以幫助後續線性轉換更快地優化，其詳細架構如圖 5 所示：

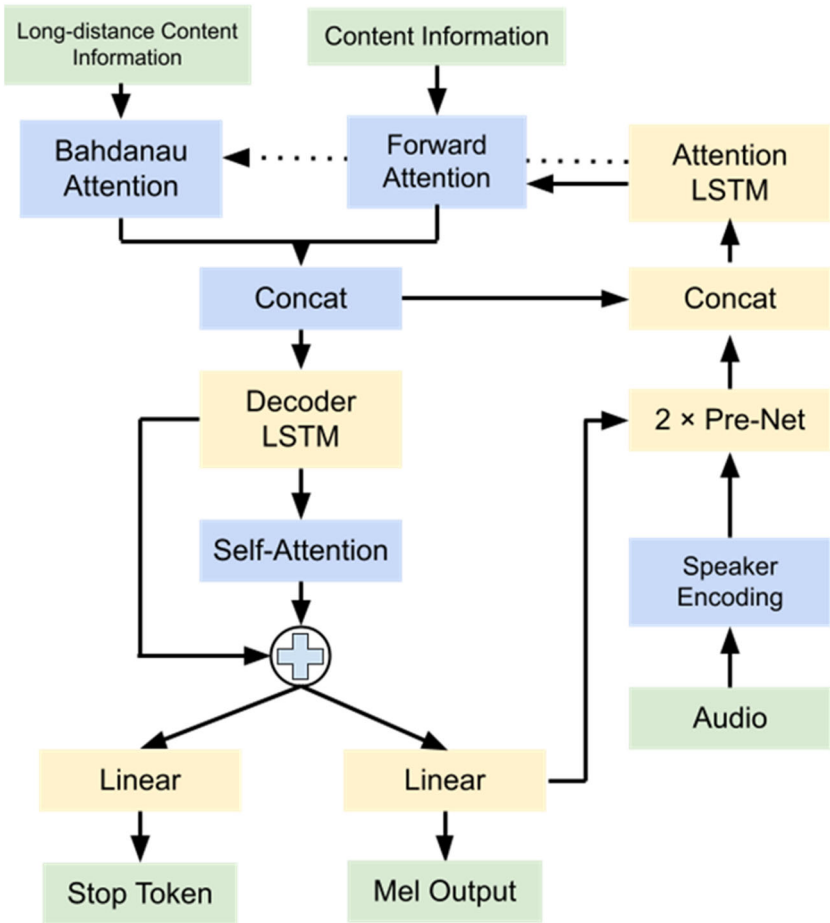

圖 5. 解碼器架構
[Figure 5. Decoder architecture]

3.3 Post-Net

原本 Post-Net 目的是為了改善頻譜重構的品質，在 Tacotron 2 的論文裡提到，有 Post-Net 的 MOS 評分是比較高的。

在本次研究中，我們額外引入了另一個架構 Diffwav (Kong, Z. *et al.,* 2020) 作為 Post-Filter 來與 Post-Net 比較。Diffwave 是 Nvidia 於 2020 年推出的 Vocoder，能夠將頻譜轉換成波形訊號，它的基礎理論是 Denoising Diffusion Probabilistic Model (Ho *et al.,* 2020)，簡稱 DDPM。DDPM 是一個馬可夫鍊 (Markov Chain) 模型，透過指定步數為目標添加高斯噪音直至目標變成高斯亂數，再透過朗之萬動力學 (Langevin Dynamics) 反向還原至目標。

我們利用上述的原理，將 Diffwave 修改成頻譜間的轉換的 Post-Filter，期望透過添加噪音能使生成的頻譜有著更多的細節，其運作流程如圖 6：

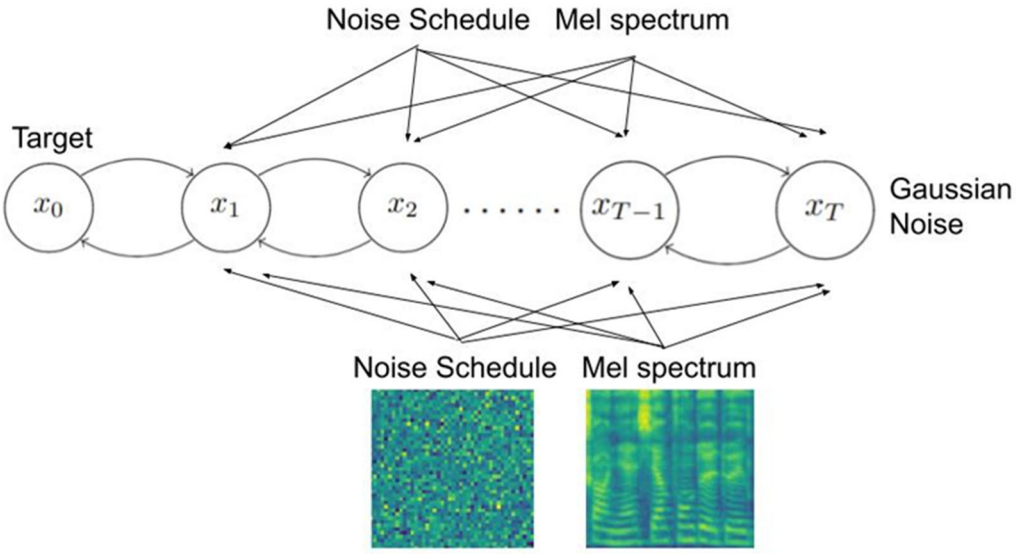

圖 6. Diffwave 流程
[Figure 6. Diffwave process]

如圖 6 所示，Diffwave 透過模型反覆運作並以噪音表 (Noise Schedule，強度由小到大的噪音) 與梅爾頻譜作為輸入條件使模型在訓練期間學習到如何透過輸入條件來添加噪音分佈破壞輸入目標；由於模型已經學得如何依照輸入條件添加噪音分佈，在推論期間，運用反函式的作法，將添加的噪音分佈除去，使輸入的高斯噪音逐漸還原成目標，其模型架構如圖 7 所示：

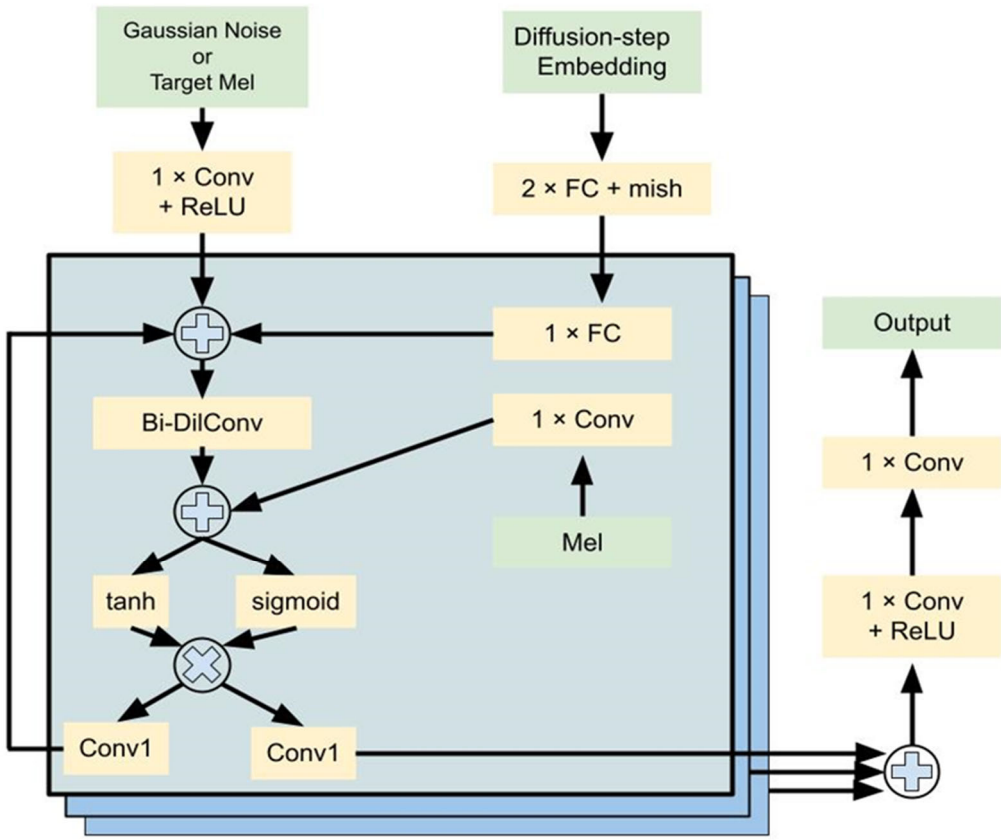

圖 7. Diffwave 模型架構
[Figure 7. Diffwave model architecture]

4. 實驗 (Experiments)

4.1 資料集 (Dataset)

我們使用 AISHELL-3 高保真中文語音數據庫作為本次實驗的資料集，共有 88035 個音檔，218 位語者，採樣率為 44.1kHz，16bit。我們將所有音檔下採樣至 22050Hz，並從中提取出 173 位語者(約佔整體語者 80%)，每位語者隨機取 100 句音檔作為訓練集，共 17300 個音檔，其餘 45 為語者當成未看過語者測驗模型合成外部語者的性能。

4.2 實驗設置 (Experimental Setups)

首先，我們使用 HiFiGAN (Kong, J. *et al.,* 2020) 作為本次實驗的 Vocoder，沒有重新訓練也沒有進行參數的微調，僅使用原作者實現的 Github 中所提供的預訓練模型。接著，我們利用資料集的音檔分別對於語音轉換的 AdaIN-VC 和語者辨識的 LDE 模型訓練，使其

生成 128 維度的語者嵌入向量。在我們提出改動的 Tacotron 2 模型架構之中，編碼層的輸出 Content Information 輸出維度仍維持 512 維，串接上語者嵌入向量後為 640 維；Long-distance Content Information 輸出維度為 128 維，串接上語者嵌入向量後為 256 維 。 在解碼層中，我們把語者嵌入向量升維至 256 維並以 Softsign 激活函數激活，於 Pre-Net 層中與頻譜相加，其餘設置皆按照原 Tacotron 2。

我們提出的 TTS 模型是在 Pytorch 神經網路框架上運行，並以 Nvidia GeForce RTX 2070 GPU 訓練，批量大小 (Batch Size) 設為 8，共訓練 208,000 個 Steps，約為 96 個 Epochs。

4.3 實驗結果 (Results)

4.3.1 語音品質 (Speech quality)

首先，我們使用客觀評測 (MOS) 來證實實驗結果，分別合成語音轉換和語者辨識所訓練的 TTS 系統各 10 個內部語者的音檔來比較品質，另外再合成各 10 個內部語者的音檔比較語者相似度，其結果如表 1：

表 1. 語音轉換和語者辨識的 MOS
[Table 1. MOS for Voice Conversion and Speaker Verification]

	Quality	***Similarity***
Tacotron 2 with VC	***2.67 ± 0.35***	***2.70 ± 0.41***
Tacotron 2 with SV	***2.54 ± 0.37***	***2.31 ± 0.18***

根據表 1 可以發現語音轉換提取出來的語者嵌入向量對於我們的 TTS 系統效果較好，因此進一步使用語音轉換的語者嵌入向量來比較 Post-Filter 與原始 Post-Net 的效果，結果如下表：

表 2. 比較 Post-Filter 與 Post-Net 的效果
[Table 2. Compare the effects of Post-Filter and Post-Net]

	Quality	***Similarity***
Post-Filter	***3.75 ± 0.35***	***3.75 ± 0.71***
Post-Net	***2.67 ± 0.71***	***2.50 ± 0.30***

接著我們使用 Mel Cepstral Distortions (MCD) 作為客觀評測的方法，隨機從內部語者與外部語者各挑選 5 個男性與女性語者，每個語者合成 10 個音檔來計算 MCD 值，結果如下表：

表 3. 計算 Post-Filter 與 Post-Net 的 MCD，值越小越好。
[Table 3. Calculate the MCD of Post-Filter and Post-Net, the smaller the value, the better.]

	Inside		***Outside***	
	Men	***Women***	***Men***	***Women***
Post-Filter	***6.99***	***7.30***	***8.15***	***8.65***
Post-Net	***7.31***	***7.98***	***9.20***	***9.11***

我們還使用 Resemblyzer 分析器計算不同性別在語音轉換上的語者空間，Resemblyzer 是一個透過神經網路來比較或分析語音的 Python 套件。研究中， 男性與女性每位語者皆合成 10 句 Post-Filter 和 Post-Net 個音檔與原語者比較，其結果如圖 8 和圖 9，我們可以從這兩張圖中發現，在內部語者中，合成的女性音檔都很接近原音檔，在男性音檔中則可以發現 Diffwave 較 Post-Net 接近原始音檔，不管是語音合成或語者辨識效果皆相似；在外部語者中，可以發現語音合成的語者空間較為集中，而語者辨識的語者空間較為發散，並且 Diffwave 比 Post-Net 稍微接近原始音檔。我們可以斷定語音合成任務以及針對 Post-Net 所提出的 Diffwave 架構對於我們的多語者 TTS 系統來說是更有幫助的。

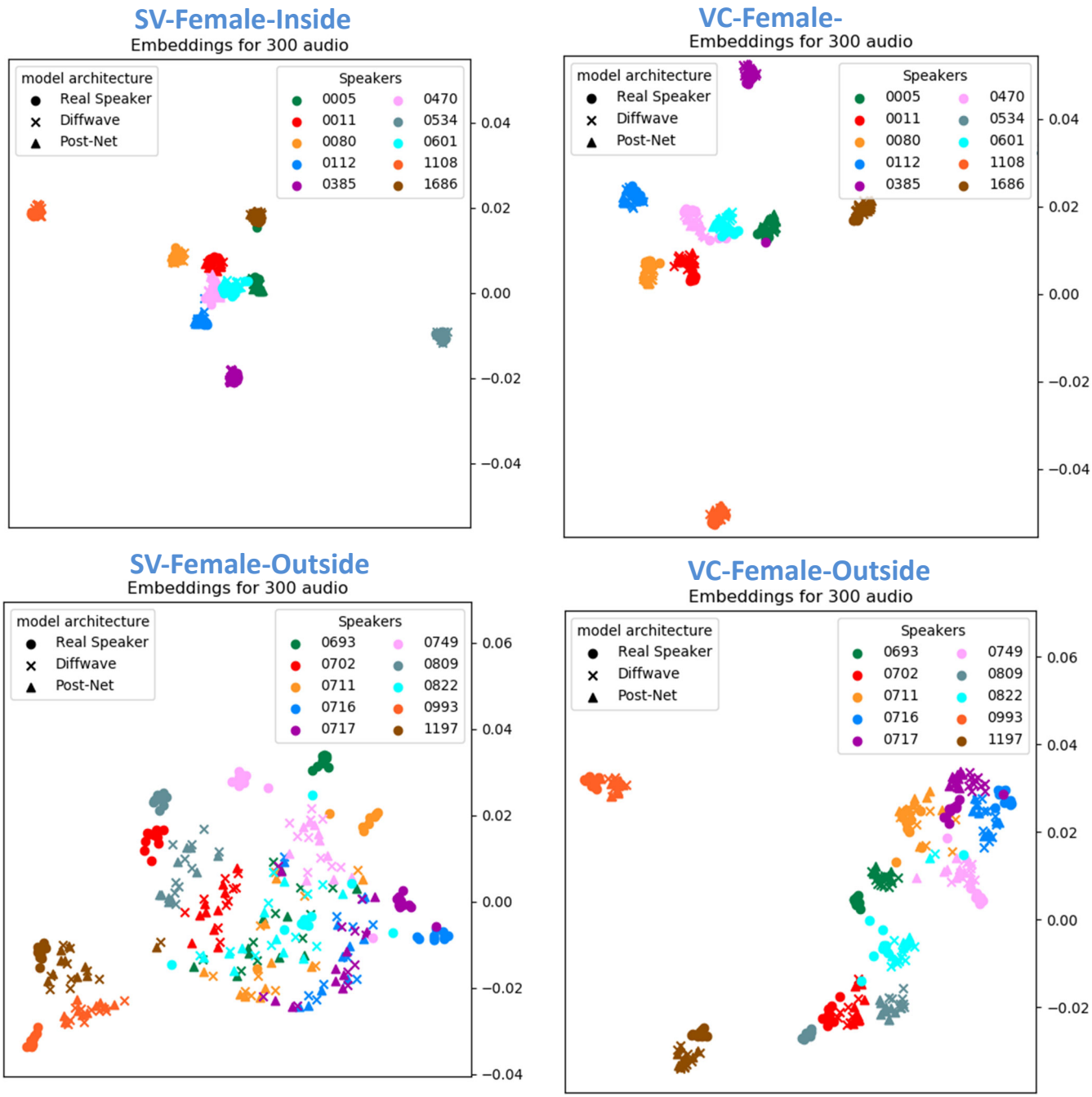

圖 8. 內外部女性語者的語者空間。

[Figure 8. Speaker space for inside and outside female speakers.]

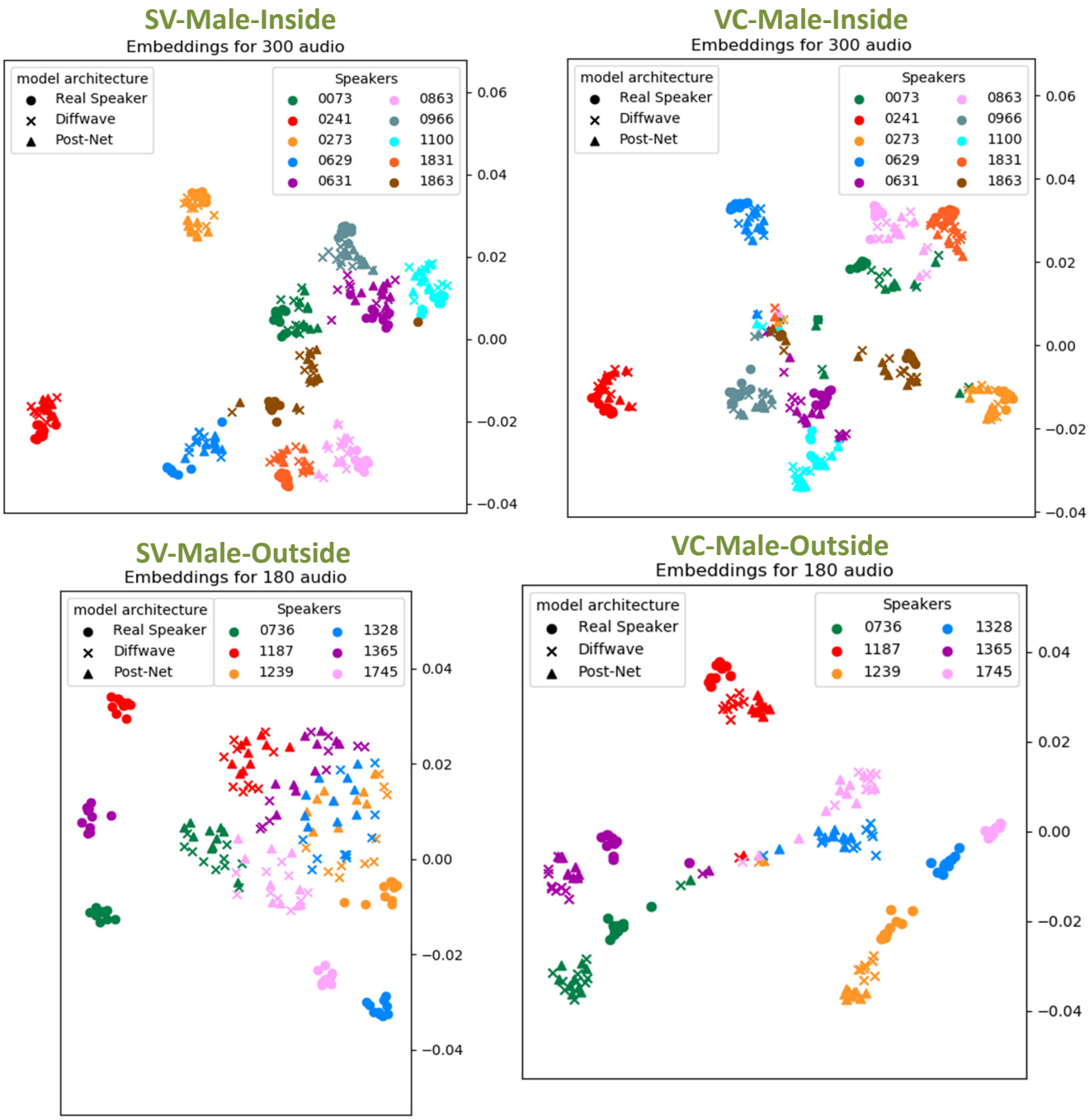

圖 9. 內外部男性語者的語者空間。

[Figure 9. Speaker space for inside and outside male speakers.]

4.3.2 注意力機制的改動 (Change in Attention mechanism)

在我們所提出的架構中，Decoder 層引入了兩個注意力機制，分別為 Forward Attention 及 Bahdanau Attention，以 “今天天氣很好。(jin1 tian1 tian1 qi4 hen2 hao3.)” 為例子，對齊圖如下：

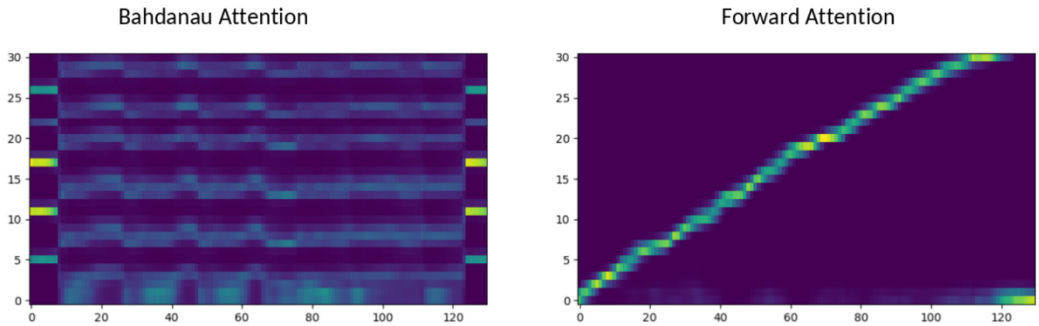

圖 10. 注意力對齊圖
[Figure 10. Attention alignment figure.]

可以看到圖 10 顯示出 Forward Attention 是斜對角的對齊圖，而 Bahdanau Attention 則包含了一些具有規律性的資訊，我們進一步針對 Bahdanau Attention 的對齊圖所顯示的資訊研究：

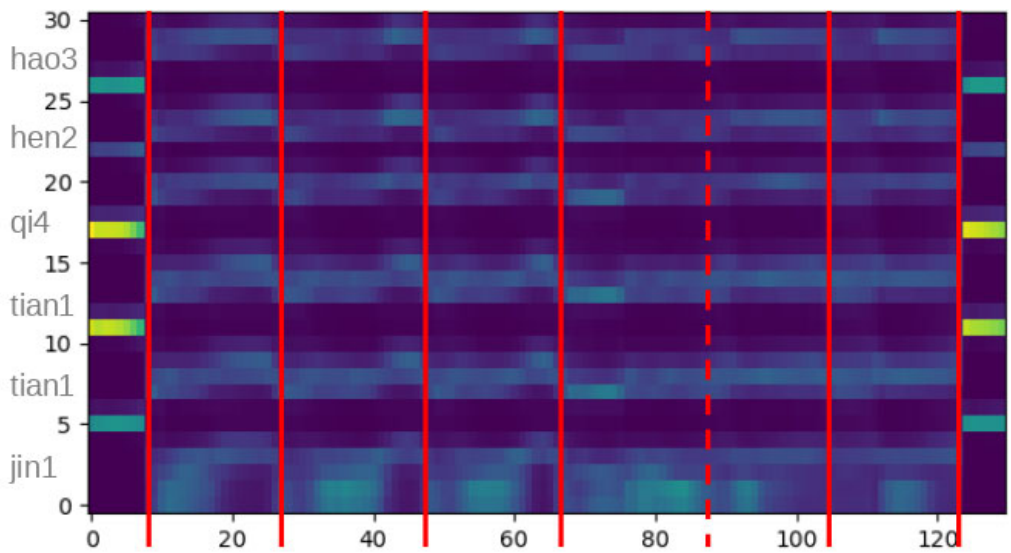

圖 11. 解析 Bahdanau Attention
[Figure 11. Parsing Bahdanau Attention.]

從圖 11 的紅線分段處，我們發現 Bahdanau Attention 提供了每段語音大概的音框範圍，圖中虛線左右處分別是” qi4” 跟” hen2” 的發音，由於它們主要都是氣音，導致分段沒有很明顯，而最左側及最右側對稱性的條紋可以判斷為空格資訊，即該片段是靜音的。

既然 Bahdanau Attention 夾帶每段語音大概的音框範圍資訊，那這些資訊是否能幫助模型快速建立對齊呢?下圖將顯示有無 Bahdanau Attention 的差異：

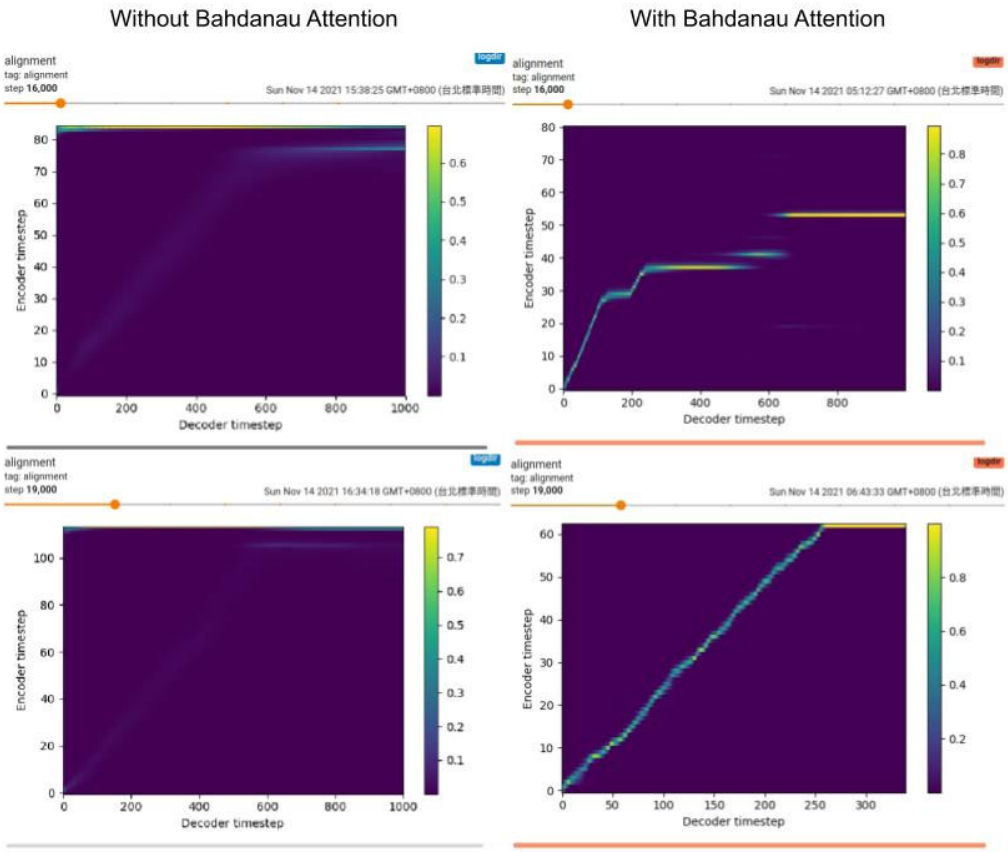

圖 12. Bahdanau Attention 能否幫助模型快速對齊？
[Figure 12. Can Bahdanau Attention help the model to align quickly?]

從圖 12 得知，模型訓練到 16000 個 Steps 時，儘管雙方都無法建立良好的對齊，但有 Bahdanau Attention 的對齊是優於沒有 Bahdanau Attention 的，在 19000 個 Steps 時，有 Bahdanau Attention 已經能建立對齊了，另一個則隱約有對齊線而已，因此可得知，Bahdanau Attention 加上 Forward Attention 的架構是能夠幫助模型快速地建立對齊。可以於我們的網站上聆聽樣本：https://babaili.github.io/rocling2021_demo/

5. 結論 (Conclusion)

在本次研究中，我們改進了多語者 Tacotron 2 的架構，透過加入語者嵌入向量便可合成未知語者的語音，並且比較語音轉換與語者辨識這兩個不同任務的語者嵌入向量用於 TTS 的成效，由實驗結果得知語音轉換的效果是優於語者辨識的，使用 Post-Filter 來提

升合成語音的語者相似度以及語音品質皆優於原始的 Post-Net，最後，於解碼層中添加第二個注意力機制有助於模型快速引發注意力對齊。未來的研究方向會加入更多資訊來幫助語者嵌入向量在 TTS 的效能上改進，例如：音韻(Prosody)資訊、發音(Articulation)資訊。

參考文獻(References)

Battenberg, E., Skerry-Ryan, R. J., Mariooryad, S., Stanton, D., Kao, D., Shannon, M., & Bagby, T. (2020). Location-relative attention mechanisms for robust long-form speech synthesis. In *Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6194-6198. https://doi.org/10.1109/ICASSP40776.2020.9054106

Chien, C. M., Lin, J. H., Huang, C. Y., Hsu, P. C., & Lee, H. Y. (2021). Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8588-8592.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. arXiv preprint arXiv:1506.07503.

Chou, J. C., Yeh, C. C., & Lee, H. Y. (2019). One-shot voice conversion by separating speaker and content representations with instance normalization. arXiv preprint arXiv:1904.05742.

Cooper, E., Lai, C. I., Yasuda, Y., Fang, F., Wang, X., Chen, N., & Yamagishi, J. (2020). Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6184-6188. https://doi.org/10.1109/icassp40776.2020.9054535

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. arXiv preprint arXiv:1806.04558.

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT)*, 266-273. https://doi.org/10.1109/SLT.2018.8639535

Kaneko, T., & Kameoka, H. (2018). Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO)*, 2100-2104. https://doi.org/10.23919/EUSIPCO.2018.8553236

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. arXiv preprint arXiv:2010.05646.

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761.

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019, July). Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6706-6713. https://doi.org/10.1609/aaai.v33i01.33016706

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,4779-4783. https://doi.org/10.1109/ICASSP.2018.8461368

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)*, 5329-5333. https://doi.org/10.1109/ICASSP.2018.8461375

Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019). Autovc: Zero-shot voice style transfer with only autoencoder loss. In *Proceedings of the 36th International Conference on Machine Learning(PMLR)*, 5210-5219.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS'17)*, 5998-6008.

Zhang, J. X., Ling, Z. H., & Dai, L. R. (2018, April). Forward attention in sequence-to-sequence acoustic modeling for speech synthesis. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4789-4793. https://doi.org/10.1109/ICASSP.2018.8462020