The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 96-110

WaveNet 聲碼器及其於語音轉換之應用

WaveNet Vocoder and its Applications in Voice Conversion

黃文勁*、羅振州*、黃信德*、曹昱**、王新民*

Wen-Chin Huang*, Chen-Chou Lo*, Hsin-Te Hwang*, Yu Tsao**, Hsin-Min Wang*

*中央研究院資訊科學研究所

Institute of Information Science

Academia Sinica

**中央研究院資訊科技創新研究中心

Research Center for Information Technology Innovation

Academia Sinica

摘要

多數語音轉換模型仰賴以傳統來源濾波器模型(source-filter model)為基礎之聲碼器(vocoder)對語音訊號進行語音參數抽取以及合成語音。然而，受限於傳統聲碼器的諸多理論與假設，以傳統聲碼器為架構進行語音轉換所生成的語音，其自然度以及與目標語者的相似度均無法進一步提升。在深度學習(deep learning)領域中，WaveNet 是現階段最成功的語音生成技術之一，能產生與過去方法相比自然度更高的語音。WaveNet 聲碼器為 WaveNet 的一個延伸，具備產生超越傳統聲碼器的高品質語音的能力，並已逐漸被從事語音轉換研究之國外團隊所採用。過去，國內研究團隊所開發的語音轉換模型多以傳統聲碼器為基礎進行語音轉換，本論文試圖將 WaveNet 聲碼器引入國內幾個新近提出的語音轉換模型，以評估 WaveNet 聲碼器在這些語音轉換模型上的應用潛力。於實驗中，我們比較了三種語音轉換模型分別使用傳統聲碼器與 WaveNet 聲碼器所得到的結果。其中，所比較的語音轉換模型包括 1)變分式自動編碼器(variational auto-encoder, VAE)、2)結合生成式對抗型網路之變分式自動編碼器、以及 3)跨特徵領域變分式自動編碼器(cross domain VAE, CDVAE)。實驗結果顯示，三個語音轉換模型在使用 WaveNet 聲

碼器後，與目標語者的相似度均獲得顯著的改善。在自然度方面，則僅有以 VAE 為基礎之語音轉換模型在使用 WaveNet 聲碼器後有顯著的提升。

關鍵詞：WaveNet，聲碼器，語音轉換，變分式自動編碼器。

Abstract

Most voice conversion models rely on vocoders based on the source-filter model to extract speech parameters and synthesize speech. However, the naturalness and similarity of the converted speech are limited due to the vast theories and constraints posed by traditional vocoders. In the field of deep learning, a network structure called WaveNet is one of the state-of-the-art techniques in speech synthesis, which is capable of generating speech samples of extremely high quality compared with past methods. One of the extensions of WaveNet is the WaveNet vocoder. Its ability to synthesize speech of quality higher than traditional vocoders has made it gradually adopted by several foreign voice conversion research teams. In this work, we study the combination of the WaveNet vocoder with the voice conversion models recently developed by domestic research teams, in order to evaluate the potential of applying the WaveNet vocoder to these voice conversion models and to introduce the WaveNet vocoder to the domestic speech processing research community. In the experiments, we compared the converted speeches generated by three voice conversion models using a traditional WORLD vocoder and the WaveNet vocoder, respectively. The compared voice conversion models include 1) variational auto-encoder (VAE), 2) variational autoencoding Wasserstein generative adversarial network (VAW-GAN), and 3) cross domain variarional auto-encoder (CDVAE). Experimental results show that, using the WaveNet vocoder, the similarity between the converted speech generated by all the three models and the target speech is significantly improved. As for naturalness, only VAE benefits from the WaveNet vocoder.

Keywords: WaveNet, Vocoder, Voice Conversion, Variational Auto-Encoder.

一、緒論

語音轉換(voice conversion)泛指在不改變來源語音的說話內容的條件下，將來源語音轉換成目標語音。語音轉換技術的應用廣泛，包括將聲音品質較差的窄頻語音訊號(narrowband speech)轉換成聲音品質較好的寬頻語音訊號(wideband speech) [1]、作為文字轉語音 (text-to-speech)系統之後處理 [2]、或是將發聲受損病患的語音轉為正常語音 [3] 等。其中，最基本及常見的應用為語者語音轉換(speaker voice conversion) (文獻中常泛稱語者語音轉換為語音轉換)，其目標為不改變來源語者(source speaker)的說話內容下，將來源語者的語音轉換成目標語者(target speaker)的語音 [4]。

多數的語音轉換方法是在聲碼器(vocoder)的框架下來實踐語音轉換，亦即需要先將語音波形進行分析(或稱參數化)，以求得聲碼器所需要的參數，例如：頻譜(spectrum)、韻律(prosody)、激發源(excitation)等特徵參數(通稱為語音參數)。接著，再對這些語音參數進行轉換，並將轉換過後的語音參數透過聲碼器還原回語音波形。簡而言之，一個典型的語音轉換系統可被拆解為三個部分：分析、轉換、合成。轉換後語音品質的好壞及與目標語者語音的相似程度，除了與語音轉換方法有密切關係外，也與聲碼器在進行分析與還原回語音波形的過程有關。過去，聲碼器的設計主要是基於來源濾波器模型(source-filter model)的理論，包括早期的線性預估編碼器 [5] 以及現今語音轉換與語音合成(文字轉語音)領域中廣被使用的高品質聲碼器，例如 STRAIGHT [6] 與 WORLD [7] 聲碼器。然而，基於來源濾波器模型理論所設計的聲碼器對人類發聲的過程有諸多的假設與簡化，使得聲音品質、自然度等無法進一步提升。

近年來，以深度類神經網絡(deep neural network, DNN)為基礎的技術被廣泛使用於語音生成相關主題上。其中，WaveNet [8] 是現階段最成功的語音生成技術之一。應用於文字轉語音，此網絡架構可從給定的語言參數及語音參數，直接產生出接近真實人聲的高品質語音波形 [8]。此外，基於 WaveNet 的聲碼器技術也已經被提出 [9, 10]。WaveNet 聲碼器為一種以資料驅動(data driven)之聲碼器，即透過大量訓練語料學習所得到的聲碼器。其扮演的角色主要用以取代傳統聲碼器在還原語音波形的過程。確切地說，給定傳統聲碼器中抽取而得的語音參數作為 WaveNet 聲碼器的輸入，則語音波形

可直接透過 WaveNet 聲碼器的輸出端解碼得到。由於 WaveNet 聲碼器可直接產生語音波形(包含了頻譜與相位資訊)，避免了傳統聲碼器使用訊號處理的技術分開估測頻譜與相位資訊後再組合還原回語音波形的過程中所導致的誤差問題，因而可以產生比傳統聲碼器自然度更高的語音。

國內已有許多相當成熟的語音轉換技術[11, 12]，但多數仍使用傳統聲碼器，使得轉換的聲音品質無法繼續提升。WaveNet 聲碼器問世後，各國頂尖的語音轉換團隊紛紛開始使用。於 2018 年語音轉換挑戰賽(Voice Conversion Challenge 2018, VCC2018) [13] 中獲得前兩名的隊伍便是使用 WaveNet 聲碼器於他們的語音轉換系統中 [14, 15]，進而大幅提升轉換後語音的自然度及與目標語者語音相似度。本論文旨在對 WaveNet 聲碼器做一介紹，並展示其與國內研究團隊所開發的語音轉換技術結合後所帶來之正面效益，提供國內語音轉換、甚至其他語音生成任務研究者參考。

本論文的章節安排如下。於第二章，我們介紹 WaveNet 聲碼器。於第三章，我們介紹由國內團隊所開發的三種語音轉換技術。第四章則為實驗結果與分析。此篇論文的總結與未來展望則呈現於第五章。

二、WaveNet 聲碼器

2.1 網絡架構

WaveNet [8] 是一個深層自迴歸網絡，藉由以下條件機率式來逐點取樣，可以生成品質極高的語音波形：

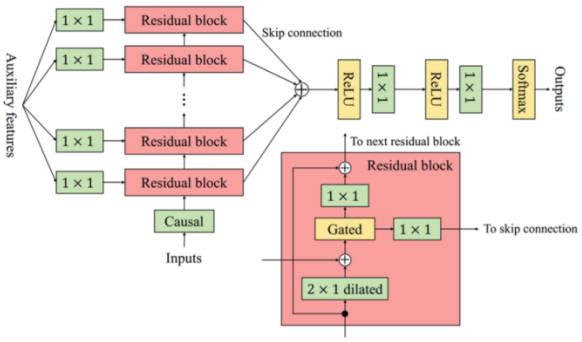
$$P(X|h) = \prod_{n=1}^{N} P(x_n|x_{n-r}, \dots, x_{n-1}, h)\,, \tag{1}$$

其中n為當前取樣的樣本點編號，r為感知域大小，x_n為當前要生成的語音訊號樣本點，且此樣本點是以整數、有限的離散位階來表示(例如 16 bits量化位階，即$2^{16} = 65536$個整數值)，h則為輔助特徵向量。若將WaveNet作為聲碼器(例如 [9, 10])，則輔助

特徵向量包含如傳統聲碼器STRAIGHT [6] 或World [7] 所抽取的語音參數等資訊，即頻譜特徵(spectral feature)、基頻(fundamental frequency)及描述激勵訊號的非週期性參數(aperiodicity)。給定輔助特徵向量，根據式(1)，WaveNet便可以估計出對應的語音樣本序列。如圖一所示，一個標準的WaveNet聲碼器包含了許多的殘差區塊(residual block)，每個殘差區塊中有一個2×1的空洞因果卷積層(dilated causal convolution)、一個門控激活函數(gated activation function)、以及一個1×1的卷積層。空洞因果卷積層可以擴大網路的感知域，而門控激活函數則被定義為：

$$\tanh\left(W_{f,k} * x + V_{f,k} * h\right) \odot \sigma\left(W_{g,k} * x + V_{g,k} * h\right), \tag{2}$$

其中x為空洞因果卷積層的輸出，W和V為可被訓練的卷積核，$*$代表卷積運算，$\odot$代表點對點相乘，$\sigma(\cdot)$代表sigmoid函數。值得一提的是，WaveNet的訓練是將每個語音樣本點產生的過程視為分類問題，並使用交叉熵(cross entropy)目標函式來訓練網路。以16 bits量化位階的語音波形為例，每個語音樣本點可以表示成65536種有限的類別，而WaveNet的訓練目標即為正確分類每個語音樣本點。為了減少計算量與避免類別過多造成的分類錯誤，輸入的波形會先藉由傳統訊號處理編碼方法(例如μ-law編碼法)進一步量化至較少位元的量化位階(例如8 bits，即每個語音樣本點僅有256個類別)。



圖一、WaveNet 聲碼器示意圖。

2.2 多語者 WaveNet 聲碼器之語者調適

最早的 WaveNet 聲碼器是一種語者依賴(speaker dependent)的聲碼器，即使用單一語者

的語料作為訓練資料集 [9]。由於訓練 WaveNet 聲碼器需要大量的訓練語料，而收集單一語者大量的語料顯得困難且不切實際，多語者 WaveNet 聲碼器(multi-speaker WaveNet vocoder)因而被提出，即在訓練階段使用多名語者的語料來訓練 WaveNet 聲碼器 [10]。將多語者 WaveNet 聲碼器應用於語音轉換時，必須針對目標語者進行語者調適(speaker adaptation)，亦即使用目標語者的語料，對多語者 WaveNet 聲碼器進行調適訓練。實驗證實，與多語者 WaveNet 聲碼器相比，此一語者調適的技巧可以有效提升輸出語音的品質及與目標語者的相似度 [14, 15, 16]。

三、語音轉換模型

3.1 基於變分式自動編碼器的語音轉換

變分式自動編碼器(variational auto-encoder, VAE) [17, 18] 可以將輸入至其編碼器(encoder)(或稱編碼函數)的語音音框(speech frame) x 編碼成包含語音內容的隱藏編碼(latent code)，以z表示。此編碼器被假設為語者獨立，亦即希望其所輸出之隱藏編碼z僅包含語音內容，如音素等語音資訊，而不包含語者資訊。隱藏編碼z與語者表示法y串接後，透過解碼器(decoder)(或稱解碼函數)來還原回對應語者表示法相關的語音訊號，便可達到語音轉換的目的。VAE 的目標函數如下三式所示，而在 VAE 學習與建模過程之最大化目標函式即為最大化式(3)：

$$log\, p_\theta(x) \le -J_{vae}(x) = -J_{obs}(x) - J_{lat}(z)\,, \tag{3}$$

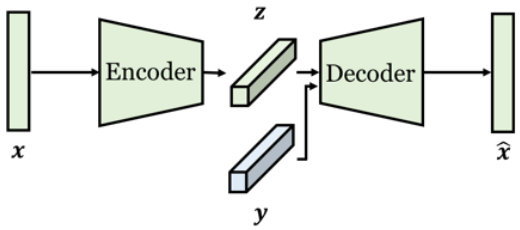
$$J_{lat}(\phi; z) = D_{KL}(q_\phi(z|x)||p_\theta(z))\,, \tag{4}$$

$$J_{obs}(\phi, \theta; x) = -E_{q_\phi(z|x)}[log\, p_\theta(x|z, y)]\,, \tag{5}$$

其中ϕ, θ分別為編碼器、解碼器參數，$D_{KL}(\cdot\,||\,\cdot)$為 KL 散度(Kullback-Leibler divergence)。

VAE 架構如圖二所示，其中x表示語音音框，z為該音框的隱藏編碼，y為語者表示法。在模型建模學習階段，VAE 對輸入語音拆解其語音內容然後還原，並且同時最佳化語者表示法，將此自我編碼與解碼還原過程應用於眾多語者的語音資料後，即可習得所有語者通用的編碼器、解碼器及各個語者各自之語者表示法。在使用階段係將輸入來源

語者之每一語音音框x經由編碼器編碼得到隱藏編碼z後，結合建模學習階段所習得之目標語者的語者表示法y，透過解碼器來將該來源語者語音轉換成目標語者語音。由於 VAE 是藉由將語音內容z結合對應語者表示法y來達到語音轉換，將來源語音轉換成建模學習語料中的任意語者，表示編碼器在建模學習過程會習得整合不同的分佈特性，最後再透過不同語者表示法來形成對應的分佈。相較於傳統語者轉換方法需要不同語者的平行語料(即語者們的訓練文本相同)及需將不同語者之相同語句進行音框對齊，VAE 在模型訓練階段是透過自我還原的方式來習得編碼器、解碼器及語者表示法，故不需要平行語料，亦毋需對齊語句。換言之，VAE 是一種可以應用於無平行語料條件下的語音轉換技術。



圖二、基於變分式自動編碼器的語音轉換示意圖。

3.2 引入生成式對抗型網路的語音轉換

生成式對抗型網路(generative adversarial network, GAN)的概念是透過讓生成模型(generator)以及鑑別模型(discriminator)互相對抗，而讓生成模型的輸出分佈最佳化[19]。在訓練過程中，生成模型會依據不同的輸入資訊以及模型架構，來產生近似真實分佈的生成輸出分佈；鑑別模型則負責區分當前輸入的分佈究竟是屬於生成模型所生成的分佈，還是真實資料的分佈。在基於 VAE 的語音轉換中，其解碼函數所解碼生成的語音決定了最終轉換後語音的品質，而藉由引入 GAN 的目標函數，原 VAE 輸出的語音品質得以進一步被增強 [12]。

在諸多 GAN 的變形中，Wasserstein GAN(W-GAN)以訓練穩定為特色而被廣泛使用[20]。W-GAN 是以推土機距離(earth mover's distance，或稱 Wasserstein distance)作為目標函數的一種 GAN 架構。於文獻 [12] 中，作者藉由計算模型所近似之分佈與原始分佈

的推土機距離來衡量語音轉換演算法的好壞，並透過 Discriminator 來增強原始 VAE 的輸出語音品質。為了讓語音轉換在非對齊語料條件下實現，目標函數被定義為推土機距離的對偶形式(Kantorovich-Rubinstein duality)，其定義如下式：

$$W\left(p_t^*, p_{t|s}\right) = \sup_{|D|_L \leq 1} \left(E_{x\sim p_t^*}[D(x)] - E_{x\sim p_{t|s}}[D(x)]\right)， \tag{6}$$

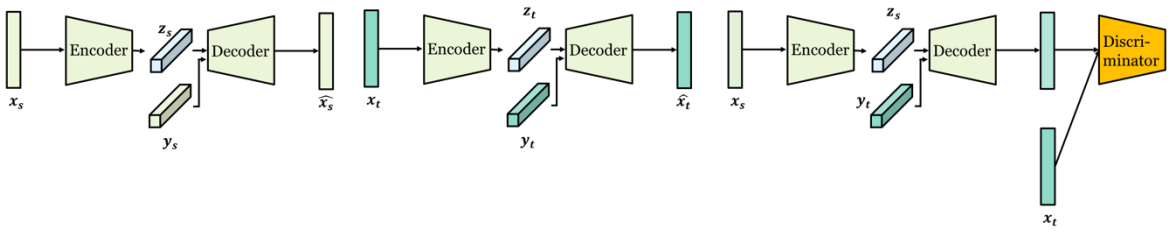
其中，D 為符合 1-Lipschitz continuity 的評分函數(critic function)，等價於 GAN 架構中的鑑別模型，p_t^*為目標語者的真實機率分佈，$p_{t|s}$為轉換後語音的機率分佈。被用來近似實現這個函數的深度神經網路為D_ψ，其中的真實分佈p_t^*可藉由取樣目標語者的語音求得，而$p_{t|s}$則是透過原始的 VAE 來獲得，亦即由轉換後的語音輸出求得。在語音轉換的應用上，推土機距離可寫為下式：

$$E_{x\sim p_t^*}\left[D_\psi(x)\right] - E_{z\sim q_\phi(z|x_s)}\left[D_\psi(G_\theta(z, y_t)\right]， \tag{7}$$

其中，G_θ為 VAE 的解碼函數。與上所述之 GAN 架構相同，評分函數D_ψ與解碼函數G_θ(等價於 GAN 架構中的生成模型)的目標是互相對抗的。在此需額外說明的是，這個架構是讓評分函數對於真實目標語音以及 VAE 解碼函數所生成的轉換語音來做鑑別，希望給真實的語音較高的分數，而給生成的語音較低的分數。然而，解碼函數的首要任務即為讓解碼生成之語音的分佈可以盡量接近目標語者真實的語音分佈，亦即減少評分函數對於真實語音與轉換語音的分數差值。透過結合 VAE 生成特性與 W-GAN 的對抗架構來達到增強 VAE 轉換語音的品質，此演算法被稱作 VAW-GAN，如圖三所示。其目標函數如下式：

$$J_{VAWGAN} = -D_{KL}(q_\phi(z|x)||p\_\theta(z)) + E_{q_\phi(z|x)}[log\, p_\theta(x|z, y)$$

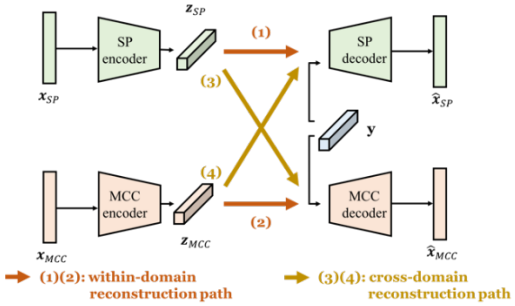
$$+E_{x\sim p_t^*}\left[D_\psi(x)\right] - E_{z\sim q_\phi(z|x_s)}\left[D_\psi(G_\theta(z, y_t)\right]。 \tag{8}$$



圖三、引入生成式對抗型網路的 VAE 語音轉換示意圖。

3.3 跨特徵領域變分式自動編碼器的語音轉換

在VAE語音轉換中，假設編碼器能萃取出與語者資訊無關的語音內容。跨特徵領域變分式自動編碼器(cross domain VAE, CDVAE)的語音轉換之中心思想，即為充分利用從同一音框中抽取的、擁有不同屬性的頻譜特徵參數，來讓編碼器萃取出更純粹的語音內容 [21]。我們使用STRAIGHT編碼器 [6] 所求得的頻譜波封(簡稱STRAIGHT spectrum, SP)及梅爾倒頻譜係數(mel cepstral coefficients, MCCs) [22] 作為兩種不同的頻譜特徵參數。如圖四所示，CDVAE為一系列自編碼器的集合，每種特徵參數有其各自的一組編碼器及解碼器。對於每種特徵參數所對應的編碼器與解碼器組，除了要可以重建輸入特徵(圖四中的路徑(1)與(2))，編碼器還要能找出一個能讓其他特徵解碼器也可以成功重建的表示法；同樣的，解碼器要能從任意編碼器所編碼出的表示法重建出該特徵(圖四中的路徑(3)與(4))。我們更進一步在目標函數中加上一個相似誤差，讓不同特徵參數的編碼器所找出的表示法互相趨近(即z_{SP}與z_{MCC}要接近)。



圖四、跨特徵領域的 VAE 語音轉換示意圖。

四、實驗

4.1 實驗設定

我們使用 Voice Conversion Challenge 2018 (VCC2018)比賽中大會所提供的語料，其中包括了 12 位專業英語語者的語音，每位語者錄製了 81 句訓練用語料以及 35 句測試用語料，取樣頻率為 22050 赫茲 [13]。我們使用 WORLD 聲碼器 [7] 作為抽取語音參數。語音訊號首先被切成長 25 毫秒、重疊 5 毫秒的音框。接著，我們從每個音框中抽取語音參數，包括 513 維的頻譜波封(spectral envelope)、513 維的非週期性訊號及基頻；35 維

的梅爾倒頻譜係數(包括第 0 維的音框能量)則從頻譜波封中抽出。

在WaveNet聲碼器方面，我們以Hayashi 等人 [10] 所提出WaveNet聲碼器還原聲音波形，實作上也使用其提供的WaveNet聲碼器開源版本[1]。此外，所有模型及訓練參數都和此開源版本預設值相同。我們使用VCC2018中的所有訓練語料，總數為972句，總長度約為54分鐘來訓練多語者WaveNet聲碼器。WaveNet聲碼器所使用的輔助特徵向量(即輸入WaveNet聲碼器的語音參數)包括去除第0維的34維的梅爾倒頻譜係數、壓縮為2維的非週期性訊號、基頻，以及清音/濁音二元特徵。為了將這些輔助特徵向量使用在WaveNet聲碼器中，實作上採用了時間解析度調整(time resolution adjustment)的技巧 [9, 10]，即簡單地複製這些聲學特徵，使得他們和輸入的語音波形有同樣的時間解析度(亦即取樣頻率)。多語者WaveNet聲碼器的訓練迴圈數為200000，語者調適則是透過進一步使用目標語者的語料進行5000個迴圈的訓練完成。

在語音轉換模型的部分，我們使用公開的VAE及VAW-GAN語音轉換模型開源版本[2]，所有模型和訓練參數都和預設值相同。我們提出的CDVAE尚未開源，但其實作概念上相當於兩個VAE，因此我們使用兩個相同的VAE模型，訓練參數也仿照開源版本中的設定。我們同樣使用VCC2018的所有訓練資料。我們對頻譜波封取對數，並進行unit-sum的能量正規化，並將能量在訓練及轉換階段取出。在轉換階段，頻譜參數由語音轉換模型進行轉換，其在使用能量還原後，與不更動的非週期性訊號，以及在對數空間進行線性轉換後的基頻，一同輸入WORLD聲碼器或WaveNet聲碼器進行語音合成。詳細VAE、VAW-GAN及CDVAE關於語音參數、類神經網路架構與參數等設定可參考 [12, 18, 21]。

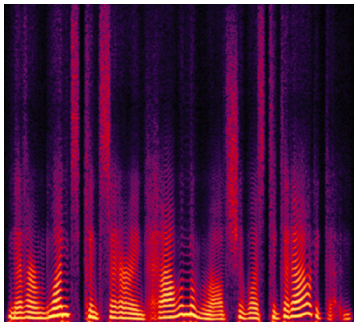
4.2 實驗結果與討論

我們針對 VCC2018 中 SF1 to TF1 這一個同性別轉換對，隨機從測試語料中選取 10 句語音，先利用 VAE、VAW-GAN、CDVAE 三種語音轉換模型進行語音轉換後，將其輸出特徵分別使用 WORLD 聲碼器以及 WaveNet 聲碼器合成語音。

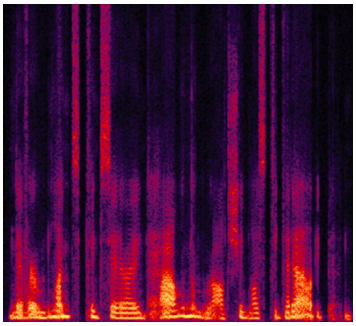
[1] https://github.com/kan-bayashi/PytorchWaveNetVocoder
[2] https://github.com/JeremyCCHsu/vae-npvc

4.2.1 頻譜圖

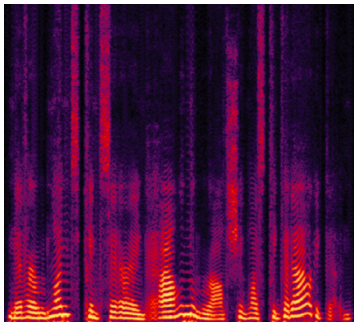
圖五為轉換語音之頻譜圖。從圖中可發現，使用 WaveNet 聲碼器所生成之語音頻譜在中高頻的結構較接近真實語音(較為隨機)；而低頻的結構則過於模糊、雜訊較多，缺乏真實語音該有的共振峰結構(formant structure)。因此，WaveNet 聲碼器所產生的語音雖然較自然，但有較多的雜音。



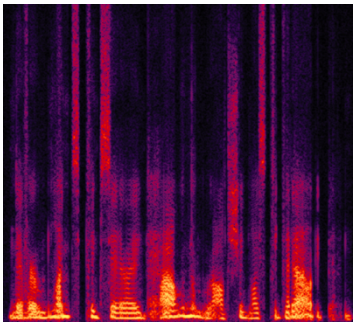
(a) VAE WORLD



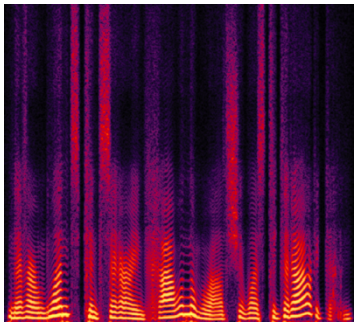
(b) VAE WaveNet



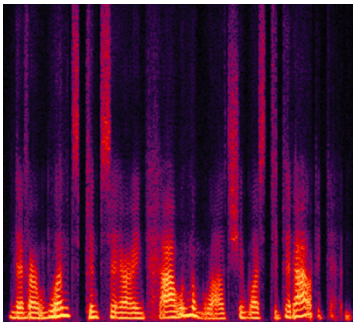
(c) VAW-GAN WORLD



(d) VAW-GAN WaveNet



(e) CDVAE WORLD



(f) CDVAE WaveNet

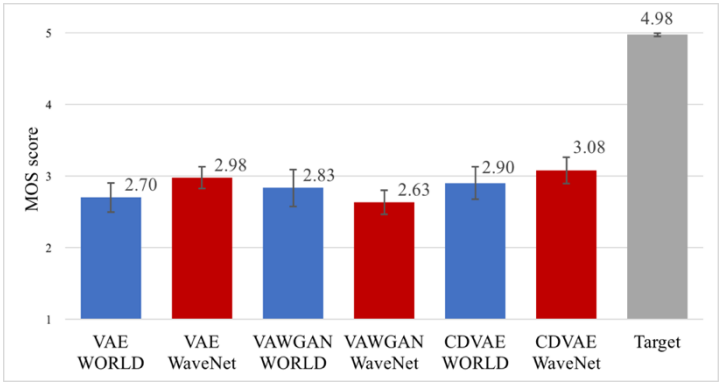
圖五、使用不同聲碼器生成之語音頻譜圖。

4.2.2 聽測實驗

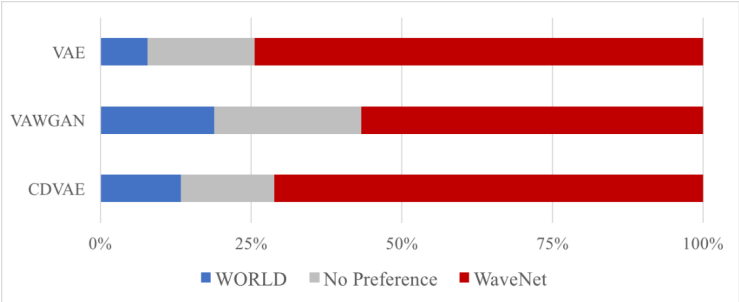
此實驗邀請 9 位受試者來進行語音自然度及與目標語者相似度的聽測實驗。圖六為受試

者針對轉換語音自然度所評估的平均意見分數(mean opinion score, MOS)之結果(包含目標語者之真實語音)。綜合平均分數及信心區間結果，可發現 WaveNet 聲碼器相對 WORLD 聲碼器在 VAE 語音轉換模型中有顯著的優勢，在 CDVAE 語音轉換模型中有些許但不顯著的優勢，在 VAW-GAN 中則是 WORLD 聲碼器相對於 WaveNet 聲碼器有些許但不顯著的優勢，這和 [23] 的結果類似。我們觀察到，WORLD 聲碼器傾向產生穩定但較糊的聲音，而 WaveNet 聲碼器雖然可以產生與 WORLD 聲碼器相比較清晰自然的語音，但有時所產生的聲音聽起來較為不穩定，含有抖音及些許雜音。我們認為可能的原因有二：其一為，在 VCC2018 語料庫中的語者彼此語音波形走勢(waveform trajectory)相差較大，造成 WaveNet 聲碼器難以有效攫取正確的走勢；其二為，WaveNet 聲碼器是以從真實語音所抽取之語音參數進行多語者訓練及語者調適，但實際進行語音轉換時卻是輸入轉換後的語音參數，這兩者間可能有落差(mismatch)。未來若能減少訓練及轉換間的落差，便可能產生自然度及品質更高的語音。

圖七為受試者針對轉換語音與目標語音之相似度的 ABX 測試結果。從圖中可以清楚地看到，相較於 WORLD 聲碼器，使用 WaveNet 聲碼器來合成語音可以有效的提高與目標語者間的相似度。綜合圖六與圖七之結果，我們可以發現 WaveNet 聲碼器所生成的語音在自然度上與 WORLD 聲碼器相似，但在與目標語者之相似度上則皆較優。值得注意的是，當把 WaveNet 聲碼器接在 VAW-GAN 後，不僅生成語音之自然度與 WORLD 聲碼器相較之下降低，在與目標語者之相似度所獲得的提升也不如 VAE 和 CDVAE。這將是未來需要被探討的。



圖六、語音自然度之平均意見分數。誤差線代表 95%信心區間。



圖七、與目標語者相似度之偏好。

五、結論

在本論文中，我們對 WaveNet 聲碼器做了一簡單的介紹，並將其應用於數個國內研究團隊所研發之語音轉換模型。從聽測實驗的結果，WaveNet 聲碼器展現了其與傳統聲碼器相比之下，所能對現有語音轉換模型產生之正面效益。目前看到的效益主要在提升轉換後語音與目標語者語音之間的相似度，在語音自然度方面則與傳統 WORLD 聲碼器有相近的表現。在未來，我們希望能針對 WaveNet 聲碼器所合成語音之自然度進行優化，例如加大訓練語料，或者探討更有效的語者調適技巧，以縮小輸入 WaveNet 聲碼器之語音參數在訓練階段與實際轉換階段之間的落差等等。我們也期望國內語音研究者因為我們的引介，能方便快速地開始使用 WaveNet 聲碼器工具，讓生成語音的品質提升。

參考文獻

[1] W. Fujitsuru, H. Sekimoto, T. Toda, H Saruwatari, and K. Shikano, “Bandwidth Extension of Cellular Phone Speech Based on Maximum Likelihood Estimation with GMM,” Proc. NCSP2008

[2] C. C. Hsia, C. H. Wu, and J. Q. Wu, “Conversion Function Clustering and Selection Using Linguistic and Spectral Information for Emotional Voice Conversion” *IEEE Trans. on Computers*, 56(9), pp. 1225–1233, September 2007.

[3] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, K. Shikano, “Alaryngeal Speech Enhancement Based on One-to-many Eigenvoice Conversion,” *IEEE/ACM Trans. on Audio, Speech, and*

Language Processing, 22(1), pp. 172–183, January 2014.

[4] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[5] B. S. Atal and S. L. Hanauer : “Speech analysis and synthesis by linear prediction of the speech wave”, in J. Acoust. Soc. America , vol. 50, no. 2, pp.637–655, Mar. 1971.

[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3, pp. 187 – 207, 1999.

[7] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. Syst., vol. E99- D, no. 7, pp. 1877-1884, 2016.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *CoRR*, vol. abs/1609.03499, 2016.

[9] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” Proc. INTERSPEECH, pp. 1118–1122, 2017.

[10] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, “An investigation of multi-speaker training for WaveNet vocoder,” Proc. ASRU, 2017.

[11] J. Chou, C. Yeh, H. Lee, L. Lee, “Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations,” Proc. INTERSPEECH, pp. 501-505, 2018.

[12] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks,” in *Proc. Interspeech*, 2017, pp. 3364–3368.

[13] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” in *Proc. Odyssey*, 2018, pp. 195–202.

[14] L. Liu, Z. Ling, Y. Jiang, M. Zhou, L. Dai, “WaveNet Vocoder with Limited Training Data for Voice Conversion,” Proc. INTERSPEECH, pp. 1983-1987, 2018.

[15] P.L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "NU voice conversion system for the voice conversion challenge 2018," in *Proc. Odyssey* 2018, pp. 219-226.

[16] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, “The NU Non-Parallel Voice Conversion System for the Voice Conversion Challenge 2018,” in *Proc. Odyssey*, 2018, pp. 211–218.

[17] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *CoRR*, vol. abs/1312.6114, 2013.

[18] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from non-parallel corpora using variational auto-encoder,” in *Proc. APISPA ASC*, 2016, pp. 1–6.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde- Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial networks,” *CoRR*, vol. abs/1406.2661, 2014.

[20] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” *CoRR*, vol. abs/1701.07875, 2017.

[21] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, “Voice Conversion Based on Cross-Domain Features Using Variational Auto Encoders,” in Proc. ISCSLP 2018.

[22] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” *in Proc. ICASSP 1992*.

[23] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, “Statistical voice conversion with WaveNet-based waveform generation,” Proc. INTERSPEECH, pp. 1138– 1142, 2017.