# 真假音之聲學特徵分析與轉換

# Feature Analysis and Conversion between Chest voice and Falsetto

專業領域：資工組
組別：A183
指導教授：劉奕汶
組員：蘇東城、鄭璟珽、黃凱琳

## Abstract

歌聲合成(SVS)的已是現今十分熱門的研究對象，然而，歌聲合成系統通常對於其合成之歌聲的音色上其實有相當大的不可控性，合成出的聲音通常一次也只有一種音色。而若能控制音色上的改變，不僅能提升歌唱作品的完整度，也能讓合成出的歌聲有更多樣的變化。

此研究主要著重在流行樂中常見的音色—真音與假音之間的轉換。我們此次的研究是聲碼器 WORLD VOCODER[8]與語音轉換模型 StarGAN 的融合應用。我們利用 WORLD VOCODER 抽取音檔的重要參數：F0、SP、AP，將真音與假音的 SP 與 AP 作為 StarGAN 的輸入，進行轉換後再由 World Vocoder 重新合成為目標音檔。資料庫的部分，我們也自行錄製了演唱流行歌曲的真音與假音共約 30 分鐘。在得到最終轉換結果後，我們對大眾進行了聽感測驗，幫助我們以客觀角度分析此次的研究成果。

經過聽感測驗調查後我們發現，轉換之音檔在音色上有較接近轉換目標，但在音質上些許退步。除此之外，我們也發現除了 AP、SP，F0 的調整也會對也會影響聽感音色。由於合成結果與預期仍有落差，我們對 World Vocoder 提取各聲學特徵的方式以及人類真假音的發聲機制都也進行進一步的論文研究，做了多個次要實驗，提出合理推論與解釋。

# Introduction

## 1. 聲學特徵定義

由於此次研究我們要進行音色間的轉換，我們需要先選擇以何種聲學特徵去描述一段歌聲。基於 WORLD VOCODER，一段聲音的組成可分成音高(F0)、頻譜包絡線(Spectral envelope, 簡稱 SP)、非週期性參數(Aperiodic parameter，簡稱 AP)，其中，F0 為一隨時間變動的曲線，SP 為一張包絡線的時頻圖，AP 則為訊號雜訊能量比例的時頻圖。

## 2. WORLD VOCODER 介紹

如前述所言，這次研究我們使用 WORLD VOCODER 作為此次研究的聲碼器。WORLD VOCODER 是一種參數型的架構，它將 F0、SP 和 AP 作為合成聲音的三種基礎參數，能夠抽取輸入音檔的參數以及使用參數合成音檔，擁有雙向的分析與合成的功能。

在抽取基礎參數時，WORLD VOCODER 分別使用了架構 Harvest[3]、CheapTrick[4]、D4C[9]來抽取一段聲音的 F0、SP 和 AP，其中

Harvest 在抽取 F0 時，分成

a. 候選 F0 的尋找　　b. F0 路徑的優化

CheapTrick 在抽取 SP 時，分成

a. 窗函數的使用　　b. 功率頻譜密度的平緩化　　c. 倒濾波器的過濾

D4C 在抽取 AP 時，則分成

a. 時域穩定參數的尋找　　b. 週期性成分的抽取　　c. AP 的最終計算

在合成聲音時，WORLD VOCODER 是先將 impulse chain 和白噪在各頻帶以特定比例混和，最後經過 SP 濾波後合成出音訊。完整架構如下圖。

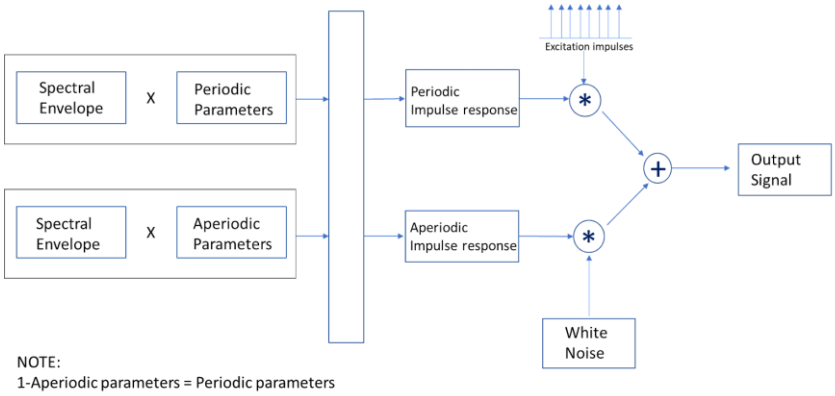
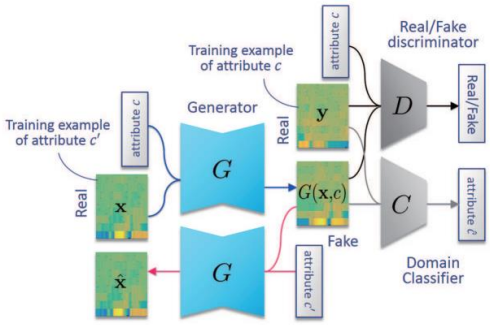



圖 1 WORLD VOCODER 合成聲音流程圖

## 3. StarGAN-VC

StarGAN[6]是一個不需要平行資料且可以達到多對多轉換的聲音轉換系統，它是以生成對抗網路作為基礎架構來進行音色間的轉換，完整架構如下。




[6]

圖 2 StarGAN 模型架構圖

由上方架構圖可看出，StarGAN 是由生成器(G)、鑑別器(D)和一個分類器(C)所組成，他們的功能分別為

a. 鑑別器：能夠區分實際此類別的特徵圖與生成器產生的特徵圖。
b. 生成器：產生的特徵圖要讓分類器能夠正確辨認，且讓鑑別器無法區分與訓練資料的差異。
c. 分類器：從訓練資料中學習如何區分各類別的資料。

## 4. 研究流程

a. 以手動客製化編輯調整歌聲真假音音色為目標，進行這次的專題。
b. 錄製以真假音音色做分類的華語流行歌聲資料庫，約 7 分鐘的真音，7 分鐘的假音，以及 6.4 分鐘的真假音(標註真假音切換時間)。
c. 基於 GAN 已經在圖片畫風轉變上已經有顯著的成效，提出使用 StarGAN 針對真
假時頻圖特徵進行轉換。
d. 以 SP 做為主要轉換的特徵，為了驗證此做法的可行性，我們先使用 LPC 以及 world vocoder 將平行聲訊的 SP 特徵取出後交換並重新合成，確定成功轉換 SP 就能成功轉換歌聲真假音音色。
e. 進行第一次訓練後發現音質不錯，但音色轉換效果不佳，因此進行以下推論:
   (1) 真假音音域有一定的落差，可能要取音域重疊的音檔進行訓練。
   (2) 真音資料有些並未相當的「真」，SP 的特徵可能不夠明顯。
   (3) 可能需要將 AP 部分也進行合成。首先，AP 也會控制 source-filter 模型中 excitation 訊號；此外，基於真假音發聲原理[2]，真假音出自於聲帶(glottal source)振動模式的不同，由此可知，轉換 AP 特徵可能是需要的。
   (4) 資料量可能不均：真音資料遠大於假音資料。
f. 經過一番實驗與對於 vocoder 的研究，我們對以上問題做了一些探討與解釋：
   (1) 我們發現當將真音轉換成假音後，有時候會因為音高太低導致音色還是偏真音。對此我們利用 vocoder 調高音高，發現音色聽起來會變假。
   (2) 實驗測試只取真音資料中真音特色較明顯的音檔來訓練，訓練出來的模型在

音色轉換上確實有較明顯。

(3) 嘗試將 AP 與 SP 連接後進行訓練，但模型轉換效果並未明顯提升。我們推測 WORLD 使用的 excitation 訊號並非我們想像中的 glottal source 訊號 (e.g. EGG)，WORLD 使用的 excitation 訊號為 impulse chain 與白噪訊號以某種比例混合。

(4) 我們發現資料量不均為主要原因。假音資料第一次僅收集到共約 3 分鐘的音檔，進行第二次資料蒐集後，將假音資料補足至約 7 分鐘，音色轉換效果明顯上升。

g. 用新的資料重新訓練 StarGAN 模型，並得到最終實驗結果。

## 5. 系統設計

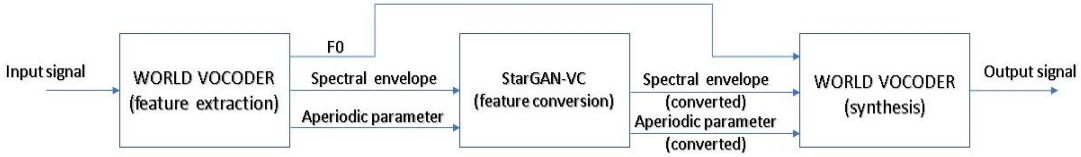



圖 3 真假音轉換系統設計圖

我們先利用 world vocoder 對輸入音檔進行聲學特徵(SP,AP,F0)的抽取，由於我們希望轉換後音高不要轉變，故僅會將 AP 與 SP 進行編碼後連結起來放入 StarGAN 模型進行轉換，取得轉換後的 AP 和 SP 後會再將其輸入 world vocoder，合成出目標音色的聲音。

## 6. 實驗結果

a. 聲學特徵轉換結果

透過觀察轉換前後的特徵時頻圖，我們發現假音轉真音時 MCC 特徵紋路會往高維延伸，真音轉假音 MCC 高維部份會被模糊掉。

下圖為假音轉真音的結果。縱軸為 frame number，橫軸為 56 維的 MCC(Mel-cepstrum coefficient)，後面連接四個一維的 coded-AP(複製四份)。

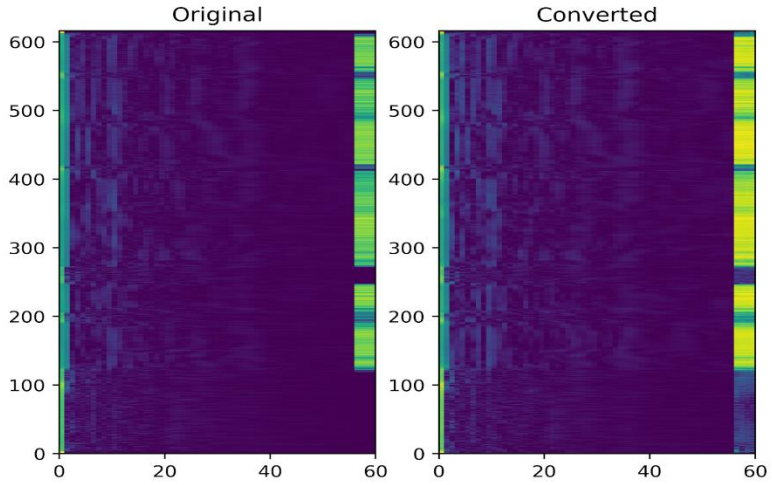


圖 4 假音轉真音前後對比圖

b. 轉換效果與音質評估

我們將合成結果隨機選出幾個真音轉假音與假音轉真音的範例，混合其他未轉換的真假音音檔，以及某些真轉假音檔調高兩個半音以及假轉真音檔調低兩個半音，共十個測試音檔及兩個參考音檔，希望實驗看看調整音高對於真假轉換是否會影響聽感。此測驗總共邀請 22 人在不知道音檔的來源的狀況下，評估這些音檔的真假音程度與音質。

評估結果：

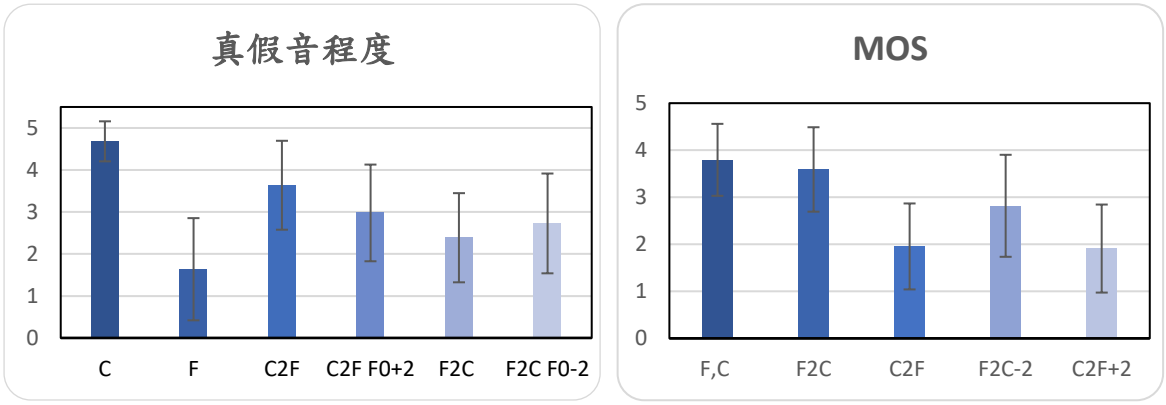



圖 5 真音程度評分圖和音值評分圖

C：真音 copy synthesis

F、C：真假音原始音檔 copy synthesis

F：假音 copy synthesis

C2F：真音轉假音

C2F F0+2：真音轉假音並調高兩個半音

F2C：假音轉真音

F2C F0-2：假音轉真音並調低兩個半音

註: 誤差棒範圍為上下各一個標準差

**7. 結論**

由上面的結果可以發現，我們轉換音色的效果是可以被人辨別的，或許音色仍無法達到原該音色音檔的水準，但音色的部分轉假轉真已經可以讓人產生混淆的效果。此外可從結果部分得知，利用此 vocoder 調整音高確實會改變到聲音的音色聽感，調高會讓音色變假，調低則會讓聲音變真。音質的部分，假轉真的部分音質較佳，真轉假的音質則相對較差。

基於音高 F0 會對音色造成影響，則其的關聯性將會是未來要探討的目標，由這次的專題實驗結果我們推論在不同 F0 下需要生成不同類型的 SP,AP 特徵才能讓真音的特色維持，讓人持續定義它是真音。而底下隱含的 SP,AP 特徵是否有與 F0 之間有一定規則? 這個規則是否能用機器學習的方式預測? 或者是用非機器學習的方式就能實踐? 這些會是我們下個需要解決的問題。

# 心得

一開始，我們對深度學習、歌聲合成等都沒有相關專業知識，因此，實作專題的第一學期都在努力充實背景知識。到了實作專題第二學期，一開始決定做真假音轉換時其實只有一個模糊的想法，到最後有一個完整的架構，有很大一部份要感謝學長和教授給予了我們許多中肯實用的建議，像是 acoustic model 和聲音轉換系統的選擇以及針對實驗結果進行討論等，讓我們知道要如何處理這些難題。而經過這次的專題，我們了解到自我學習的重要性，也學習到如何進行團隊溝通與分工，我們相信這些都會是使我們成長的重要養分。

# 參考文獻


1. Morise, Masanori & zawa, Kenji. (2017). Low-Dimensional Representation of Spectral Envelope Without Deterioration for Full-Band Speech Analysis/Synthesis System. 409-413. 10.21Miyashita, Genta & O437/Interspeech.2017-67.
2. Roubeau, Bernard & Henrich Bernardoni, Nathalie & Castellengo, Michèle. (2008). Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited. Journal of voice : official journal of the Voice Foundation. 23. 425-38. 10.1016/j.jvoice.2007.10.014.
3. Morise, Masanori. "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals." INTERSPEECH. 2017.
4. Morise, Masanori. "CheapTrick, a spectral envelope estimator for high-quality speech synthesis." *Speech Communication* 67 (2015): 1-7.
5. Choi, Yunjey, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
6. Kameoka, Hirokazu, et al. "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.
7. P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967-5976, doi: 10.1109/CVPR.2017.632.
8. M. Morise, F. Yokomori, and K. Ozawa: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016.
9. Morise, Masanori. "D4C, a band-aperiodicity estimator for high-quality speech synthesis." Speech Communication 84 (2016): 57-65.