國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

任意對多歌唱風格轉換

Any-to-many Singing Style Conversion

許育騰

Yu-Teng Hsu

指導教授: 張智星 博士

Advisor: Jyh-Shing Roger Jang, Ph.D.

中華民國 113 年 7 月

July 2024

doi:10.6342/NTU202402100

國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

任意對多歌唱風格轉換

Any-to-many Singing Style Conversion

本論文係許育騰君（學號 R11922003）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 113 年 7 月 13 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Computer Science and Information Engineering on 13 July 2024 have examined a Master's thesis entitled above presented by HSU, YU-TENG (student ID: R11922003) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

張智星

（指導教授 Advisor）

曹昱

王新民

系主任/所長 Director: 陳祝嵩

doi:10.6342/NTU202402100

誌謝

首先，我想對我的指導教授張智星老師表達最深的謝意，感謝他在我碩士生涯中給予的細心指導，尤其是在口試準備期間提供了許多寶貴的研究建議。除此之外，我也要特別感謝王鈞右學長，沒有他的幫助和支持，我可能無法順利完成這項研究。同時，我要感謝王崇喆學長提供的重要意見，使得這項研究能夠更加完善。最後，我要衷心感謝 MIRLAB 的同學們以及主觀評量問卷的所有參與者，他們的幫忙對本論文的完成極其重要。

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

摘要

近年來，將歌聲中的歌手身份轉換成另一位歌手的任務，或稱為歌聲轉換，已經取得了巨大的成功。大多數現有的歌聲轉換系統僅考慮了歌聲的音色轉換，其他資訊則保持不變。然而，這未充分考慮歌手身份的其他方面，特別是體現在歌聲的音高曲線和能量曲線中的歌唱風格。為了解決這個問題，本論文提出了一個任意對多的歌唱風格轉換系統，將一位歌手的音高曲線和能量曲線轉換為另一位歌手的風格。為了實現這個目標，我們利用了兩個類似 AutoVC 具有信息瓶頸的自編碼器，以將歌唱風格與音樂內容區分開來。第一個自編碼器執行音高轉換，而第二個自編碼器則以音高曲線為條件執行能量轉換，以確保兩個曲線之間的一致性。考慮到顫音在歌聲表達中的重要性，我們進一步加入了強調顫音特徵的損失函數，以突顯其作用。實驗結果顯示，我們提出的模型能夠有效地在任意對多的情境下將音高和能量特徵的風格轉換為目標歌手的歌唱風格。

關鍵字：歌唱風格轉換、音高轉換、能量轉換、自編碼器、顫音學習

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

Abstract

The task of converting singer identity of a singing voice to that of another singer, or singing voice conversion (SVC), has achieved a huge success in recent years. Most existing SVC systems consider the conversion of a singing voice's timbre while leaving all other information unchanged. This, however, does not take other aspects of singer identity into consideration, particularly a singer's singing style, which is reflected in the pitch and the energy contours of a singing voice. To address this issue, this paper proposes an any-to-many singing style conversion system that converts the pitch and energy contours of one singer's style to that of another singer's style. To achieve this target, we utilize two AutoVC-like autoencoders with information bottleneck to disentangle singing style from musical contents. The first one performs pitch conversion, while the second one performs energy conversion with the condition of pitch contour to ensure a consistency between the two contours. Recognizing the crucial role of vibratos in vocal expression, we further incorporate loss functions that emphasize vibrato features to highlight their importance. Experimental results suggested that the proposed model can effectively convert the style of pitch and energy features to that of target singer in an any-to-many conversion scenario.

Keywords: singing style conversion, pitch conversion, energy conversion, autoencoder, vibrato learning

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

目次

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

圖次

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

表次

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

第一章　緒論

本章節將介紹本論文之研究內容與動機、所採用的研究方法與貢獻，以及論文的章節安排。首先，研究簡介與動機部分將說明研究的背景、問題的定義及其重要性；接著，研究方法與貢獻部分將描述研究過程中所使用的方法，以及本論文的貢獻；最後，章節概述部分將簡介本論文各章節的內容。

1.1　研究簡介與動機

歌聲轉換（singing voice conversion）旨在將一段歌聲轉換成目標歌手的聲音。其假設歌聲可以分解為「內容（content）」和「歌手身份（singer identity）」兩個部分，因此可透過歌聲轉換模型僅修改歌聲中的歌手身份，而內容的部分則保持不變。為了解決這個任務，近期大多數的歌聲轉換模型 [5, 7, 12, 16, 18, 19, 21, 22, 25, 26, 31, 38] 主要專注於轉換歌聲的音色（timbre），同時保留其他非音色的特徵。圖 1.1 展示了一個歌聲轉換模型的範例架構，可以觀察到音高的資訊直接從來源音檔中保留下來，並未經過轉換。在這樣的問題定義下，近年來提出了各種方法將音色與其他特徵區分開來，從而在音檔品質和轉換後的歌聲音色相似度方面取得了顯著的進步 [11]。

然而，這樣的問題定義將歌手身份簡化為僅包含音色的資訊，並未考慮到歌唱風格，這可能會導致整體歌手身份相似度存在一些問題。舉例來說，有些歌手在表演時傾向於加入強烈的顫音（vibrato），而其他歌手則相反。如果我們僅將帶有強烈顫音的歌手音色轉換成很少唱顫音的歌手，但保留了強烈顫音的歌唱風格，那麼轉換後的歌聲可能會顯得有些奇怪，原因在於轉換後的歌聲未能完全符合目標歌手的歌唱風格。因此，即使音色能夠完美轉換，轉換後的歌聲與目標歌手之間的整體歌手身份相似度仍有待改進。

doi:10.6342/NTU202402100

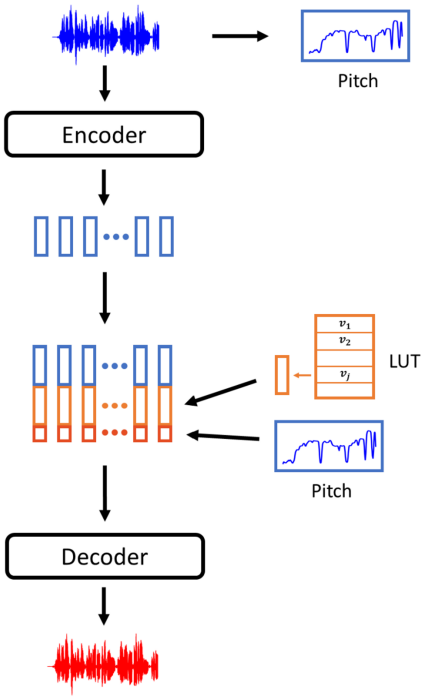


圖 1.1: 歌聲轉換模型之範例架構圖 [5]

為了解決這個問題，在本論文中，我們提出了一個新的任務，即在任意對多（any-to-many）的情境下進行歌唱風格轉換。其中，「任意」指的是來源歌手可以是任何一位歌手，並不一定是在訓練過程中見過的；而「多」則表示目標歌手必須是在訓練過程中見過的歌手。由於文獻中對於「歌唱風格」一詞的含義存在模糊性 [27]，我們首先在此提供一個更具體的定義。我們將在本論文中要轉換的歌唱風格限定為表現在音高曲線（pitch contour）和能量曲線（energy contour）中的風格，其中，音高表示歌聲的基本頻率（fundamental frequency, F0），而能量則表示音量的變化。這個定義包括了顫音、過衝（overshoot）、預備動作（preparation）[29] 以及其他可在音高曲線中辨識的歌唱技巧，同時也包含了在能量曲線中可觀察到的動態變化。換句話說，我們的重點僅在於轉換來源音檔的音高曲線和能量曲線。因此，涉及音色變化的歌唱風格，例如聲帶的使用 [27] 等，並不在本論文的研究範圍之內，儘管它們可能會被歌聲轉換模型處理，因為與音色有關。本論文的研究雖然沒有涵蓋到歌唱風格的全部範疇，然而，通過將提出的模型與歌聲轉換模型結合，我們希望轉換後的歌聲將更接近於目標歌手的歌唱風格。

doi:10.6342/NTU202402100

1.2 研究方法與貢獻

受到 AutoVC [28] 的啟發，我們採用了一個具有信息瓶頸（information bottleneck）的自編碼器（autoencoder），以將歌唱風格與音高曲線和能量曲線中的內容資訊分離開來。這裡的內容資訊包括音符音高（note pitch）和由發音不同音素（phoneme）引起的能量變化等。我們提出的歌唱風格轉換模型是一個二階段的模型。首先，第一個模型進行音高曲線轉換，接著，第二個模型根據第一個模型的輸出進行能量曲線轉換，以確保轉換後的音高曲線和能量曲線之間的一致性。此外，為了有效地捕捉歌聲中的顫音特徵，我們參考了 [17] 中提出的顫音損失函數進行訓練。

在評量方法的部分，我們進行了客觀評量和主觀評量。在客觀評量中，我們參考了語者驗證（speaker verification）的方法 [4]，計算轉換結果與目標歌手之間的歌手嵌入（singer embedding）餘弦相似度（cosine similarity）。至於主觀評量，我們則通過問卷的方式進行了比較平均意見分數（comparison mean opinion score, CMOS）[20] 評量，邀請受測者根據兩段轉換後音檔的自然度和與目標歌手歌唱風格的相似度來評分。

本論文的主要貢獻可以歸納如下：

- 提出了首個任意對多歌唱風格轉換模型。
- 將音高曲線作為能量轉換模型的輔助資訊，以確保兩者之間的一致性。
- 改良利用顫音損失函數對顫音建模的方法，並引入顫音幅度平滑損失函數。

1.3 章節概述

本論文共分為六個章節，各章節概述如下：

- 第一章：緒論
  簡介本論文之研究背景、研究動機以及研究方法。
- 第二章：文獻探討
  簡介與本論文研究目標相關之研究方法，並介紹本論文主要參考的模型。

doi:10.6342/NTU202402100

- 第三章：研究方法

  闡述本論文提出的方法和模型，包括資料處理、模型架構以及損失函數等。

- 第四章：資料集與實驗設定

  介紹本論文使用的資料集，並詳述評量方法和指標，同時闡述實驗設定。

- 第五章：實驗結果與討論

  呈現各項實驗結果，並對其進行深入分析，釐清各個結果的意義和影響。

- 第六章：總結

  綜觀各項實驗結果，歸納本論文的研究成果並探討未來可拓展的研究方向。

doi:10.6342/NTU202402100

第二章　文獻探討

據我們所知，目前尚無研究專注於歌聲中音高曲線（pitch contour）和能量曲線（energy contour）的歌唱風格轉換。然而，在其他相關領域中有一些具有相似目標的研究，例如語音轉換（voice conversion）中對語者的音高進行轉換，以及歌聲合成（singing voice synthesis）中生成情感表達豐富的歌聲，還有歌唱技巧轉換（singing technique conversion）旨在轉換歌聲中如氣泡音（vocal fry）和氣音（breathy）等歌唱技巧。以下將先依序簡介這些研究的方法，最後再介紹本論文所主要參考的模型架構 AutoVC [28]。

2.1　語音轉換中之音高轉換

語音轉換中之音高轉換旨在除了轉換語者的聲音特徵之外，還要轉換其發音時的音高。首先，我們將介紹相關研究所使用的方法，接著說明語音與歌聲在特性上的差異，這些差異的探討旨在突顯針對歌聲之音高轉換進行研究的重要性。

2.1.1　相關研究之方法

傳統的方法大多需要平行資料（parallel data）或音節標註（syllabic annotation），即不同語者說相同句子的資料，或以音節為單位進行標註。直到 [35] 提出了一種新方法，利用三音框對齊（tri-frame alignment）的方式將來源語音和目標語音的音框對齊，因此可以僅使用非平行資料進行訓練。該方法還利用高斯混合模型（Gaussian mixture model, GMM）來建模不同語者之間的聯合音高分佈，實現音高轉換。

儘管高斯混合模型可以有效地轉換語者的聲音特徵，但仍存在一些過度平滑（over-smoothing）的問題。有鑑於此，[36] 採用了神經網絡來替代高斯混合模

doi:10.6342/NTU202402100

型，以減少過度平滑的現象。不過，這個方法需要平行資料來訓練模型，而且模型架構僅包含了三層線性層。隨後，[14] 進一步提出了一個端到端（end-to-end）的語音與音高轉換模型，其模型架構包括小波核卷積編碼器（wavelet kernel convolutional encoder）和雙向生成對抗網絡（Dual-GAN）。前者用於對不同時間尺度的音高變化進行編碼，後者則用於學習語者間的聲音特徵。然而，這個方法仍然需要平行資料。

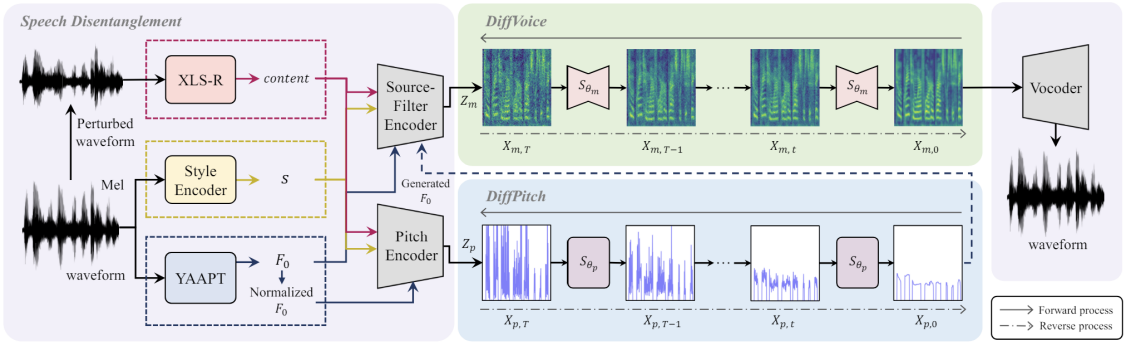


圖 2.1: Diff-HierVC 模型架構圖 [2]

此外，上述所介紹的方法基本上僅能實現一對一（one-to-one）的轉換，即來源語者和目標語者被限制為訓練模型時所使用的特定一位語者。直到 Diff-HierVC [2] 的出現，其採用了多個條件擴散模型（conditional diffusion model）的階層式框架，實現了任意對任意（any-to-any）的語音轉換和音高轉換。其模型架構如圖 2.1 所示。首先，從語音波形（speech waveform）中分離出內容、風格和音高三個部分；接著，DiffPitch 將音高轉換為目標語者的風格；最後，DiffVoice 根據轉換後的音高將來源語音轉換為目標語者的聲音特徵。

2.1.2 語音與歌聲之差異

在風格轉換方面，語音和歌聲之間存在著一些本質上的差異，特別是在音高轉換的部分。例如，不同的語者可能以不同的絕對音高表達相同的句子。然而，對於歌聲而言，這樣的音高差異通常是不被允許的，因為歌手在演唱時必須遵循樂譜的音符，所以歌聲的歌唱風格差異通常只涉及音高數值的微小變化和波動，例如顫音（vibrato）或滑音（glissandi）。儘管這樣的波動通常只在一個或兩個半音之內，但人們仍然可以根據這些細節來辨識不同歌聲之間的歌唱風格差異。因此，我們認為設計一個專門用於歌聲的音高轉換模型是值得關注的。

doi:10.6342/NTU202402100

2.2 表現性歌聲合成

表現性歌聲合成旨在透過樂譜合成出具有目標歌手音色的歌聲，且這段歌聲需富有情感表達能力。而最近，有幾篇關於歌聲合成的研究嘗試透過預測音高和能量來生成具有表現力的歌聲。

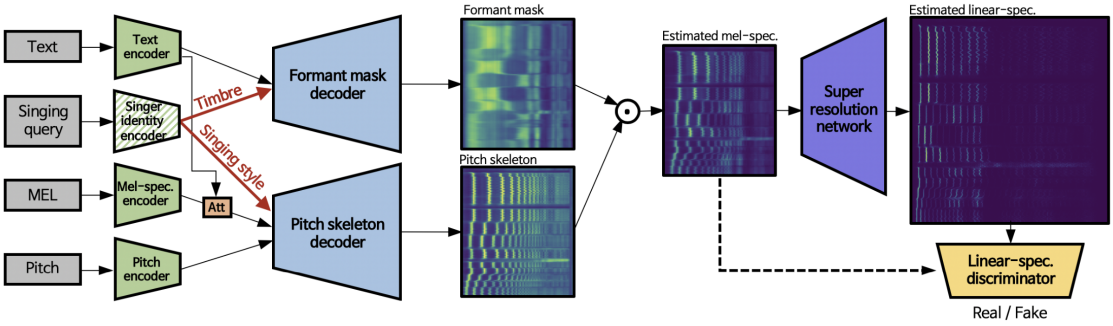


圖 2.2: J. Lee 提出之模型架構圖 [15]

[15] 提出了一個多歌手的歌聲合成模型，能夠獨立模擬不同歌手的音色和歌唱風格。其將歌手身份（singer identity）定義為包括音色和歌唱風格，並使用兩個獨立的解碼器（decoder）來模擬這兩個部分。圖 2.2 展示了其模型架構。首先，透過歌手身份編碼器生成歌手嵌入（singer embedding）；接著，Formant mask decoder 和 Pitch skeleton decoder 根據歌手嵌入分別產生對應的音色和音高特徵；最後，計算這兩者的哈達瑪乘積（Hadamard product）即可得到時頻譜（spectrogram）。

另一項研究是由 [30] 提出的歌聲合成模型，該模型能夠從包含音符和音素（phoneme）序列的樂譜中預測音高和能量。其模型架構如圖 2.3 所示。在音高方面，他們將音高曲線分解為語調（intonation）和顫音。語調表示了樂譜的音符，而顫音包含相位（phase）、深度（depth）、速度（rate）和可能性（likeliness）等四個參數，音高模型則根據這些資訊來學習顫音的特徵。而在能量方面，以往的方法通常將能量壓縮成一維向量，但作者認為這樣可能導致大量的資訊損失。因此，他們利用自編碼器（autoencoder）的架構來產生潛在能量表示（latent energy representation），以減輕壓縮時的資訊損失。

此外，在音高預測器的訓練中，[17] 運用了基於對音高曲線進行短時距傅立葉變換（short-time Fourier transform, STFT）的損失函數，以強調顫音的重要性。圖 2.4 呈現了其對顫音建模的方法。圖 2.4（a）為一段範例音高曲線。首先，對音高曲線進行了一階差分（first-order difference），以突顯高頻變化，如圖 2.4（b）

doi:10.6342/NTU202402100

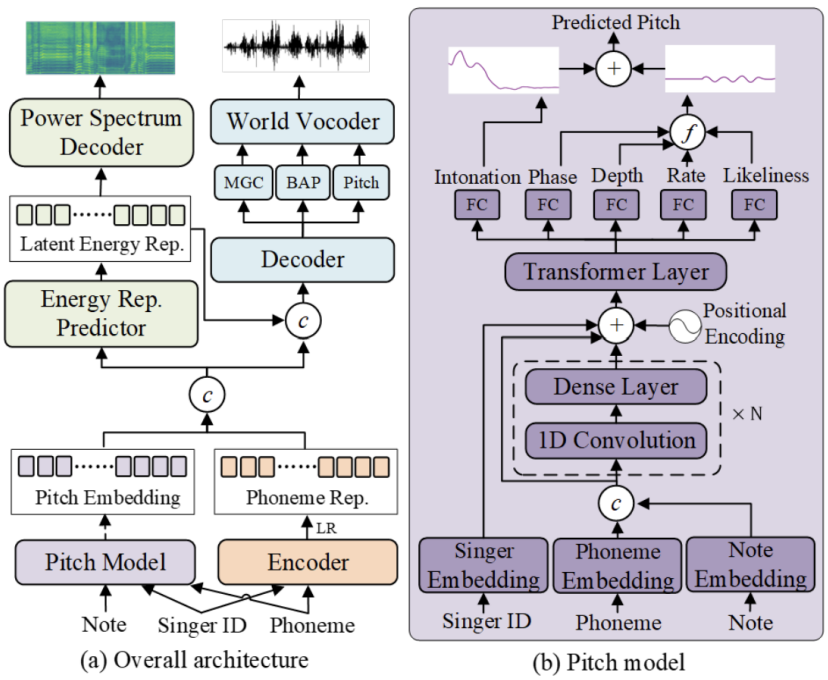


圖 2.3: ExpressiveSing 模型架構圖 [30]

所示；接著，將一階差分的結果轉換為功率時頻譜（power spectrogram），以突顯曲線中週期性的部分，如圖 2.4（c）所示；最後，考慮到顫音頻率通常在 5 Hz 至 8 Hz 之間，因此計算該範圍內功率時頻譜的最大值，即可得到代表顫音程度的顫音幅度曲線（vibrato extent contour），如圖 2.4（d）所示。

在這些研究的啟發下，我們提出了一個模型，以解決一個稍有不同的任務，即將來源音檔的歌唱風格轉換為目標歌手的風格，而非透過樂譜合成歌聲。

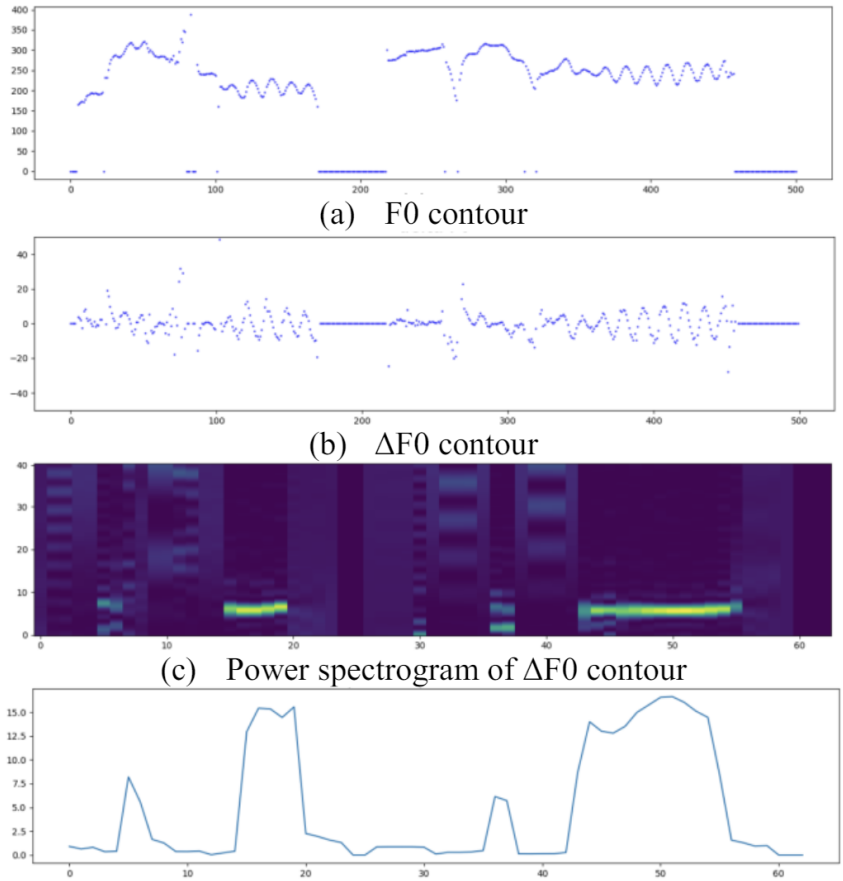
2.3 歌唱技巧轉換

歌唱技巧轉換的目的在於不影響歌手身份、音樂結構和歌詞內容的前提下，進行歌唱技巧的轉換。首先，我們將介紹相關研究所採用的方法，接著說明歌唱技巧與歌唱風格之間的差異，以強調本論文的獨特性和與現有研究的不同之處。

2.3.1 相關研究之方法

[23] 提出了首個多對多（many-to-many）歌唱技巧轉換模型，該模型通過將歌手身份和歌唱技巧分離開來的方法，同時實現了歌聲轉換（singing voice

doi:10.6342/NTU202402100



(a) F0 contour

(b) ΔF0 contour

(c) Power spectrogram of ΔF0 contour

(d) Maxima of ΔF0 power spectrogram between 5~8Hz

圖 2.4: R. Liu 提出之顫音建模方法說明圖 [17]

conversion）和歌唱技巧轉換。其模型架構如圖 2.5 所示。首先，利用特徵抽取網路（feature extraction network, FEN）從時頻譜中提取特徵；接著，分別通過歌手編碼器（圖 2.5 中藍色部分）和歌唱技巧編碼器（圖 2.5 中紅色部分）萃取對應的特徵，並藉由兩個相應的分類器（classifier）確保學習到的特徵與歌手或歌唱技巧相關；最後，透過編碼器（圖 2.5 中綠色部分）和精煉網路（refinement network）來重建和精煉時頻譜。

[27] 將轉換情境進一步擴展，提出了一個基於 AutoVC [28] 架構的任意對任意歌唱技巧轉換模型。關於 AutoVC [28] 架構的詳細介紹可參見 2.4 小節。該模型主要將 AutoVC [28] 中用於對語者身分進行編碼的語者編碼器替換為作者提出的歌唱技巧編碼器，該編碼器是經由訓練分類器的方式而得到的。

doi:10.6342/NTU202402100

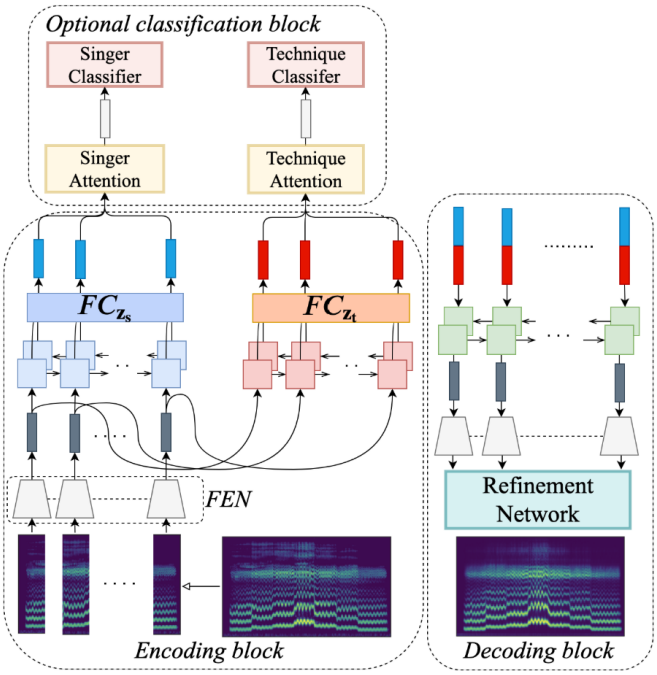


圖 2.5: Y.-J. Luo 提出之模型架構圖 [23]

2.3.2 歌唱技巧與歌唱風格之差異

雖然歌唱技巧轉換的任務看似與歌唱風格轉換相似，但實際上兩者之間存在幾個差異。歌唱技巧轉換專注於轉換一組與歌手身份無關的歌唱技巧，例如，將具有氣音技巧的歌聲轉換為氣泡音技巧。另一方面，歌唱風格轉換則著重於將歌聲的歌唱風格轉換為目標歌手的風格，而非特定歌唱技巧，該目標歌手可能在表演中採用各種歌唱技巧，或反之，可能沒有特殊的歌唱技巧。此外，上述研究中討論的部分歌唱技巧與音高曲線或能量曲線無關，而本論文則涵蓋了一些與能量動態變化相關的歌唱風格。

2.4 AutoVC 架構簡介

AutoVC [28] 是一個用於任意對任意語音轉換任務的模型，其架構基於自編碼器，並具有信息瓶頸（information bottleneck）以分離內容和語者資訊。圖 2.6 展示了其模型架構，主要分為三個部分：內容編碼器如圖 2.6（a）、語者編碼器如圖 2.6（b）以及解碼器如圖 2.6（c）。而圖 2.6（d）則為時頻譜反轉器（spectrogram inverter），用於將時頻譜轉換回語音波形。這個模型的特點在於其架構簡單，並且在訓練時僅需自重建損失（self-reconstruction loss）即可實現語音風格轉換的效

doi:10.6342/NTU202402100

果。以下將依序介紹內容編碼器、語者編碼器和解碼器的架構。

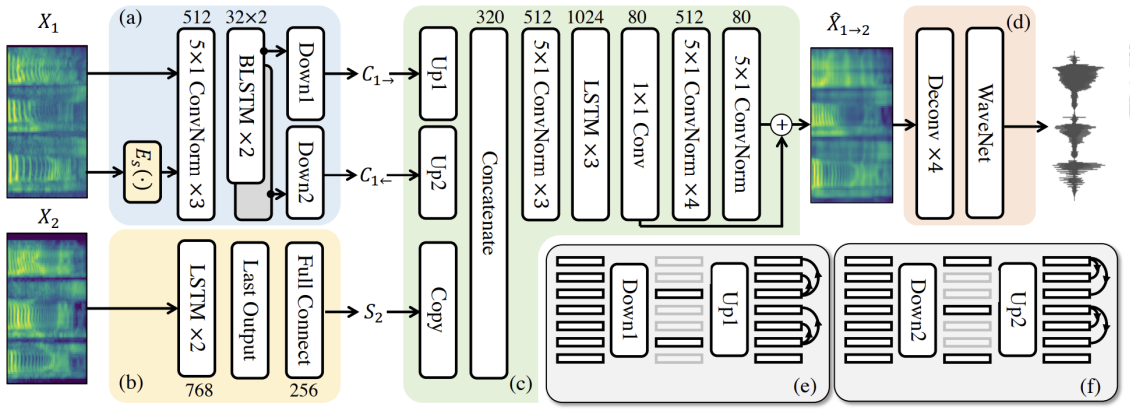


圖 2.6: AutoVC 模型架構圖 [28]

2.4.1 內容編碼器

內容編碼器主要由卷積（convolution）層和雙向長短期記憶網路（bi-directional long short-term memory, BiLSTM）所組成。其功能在於保留輸入時頻譜中的內容資訊。因此，在內容編碼器的最後，模型會對前一層的輸出進行下採樣（downsampling），以形成信息瓶頸，過濾掉語者相關的資訊。由於前一層 BiLSTM 的輸出包含正向和反向兩個方向的資訊，因此兩者的下採樣方法也有所不同。正向輸出的部分如圖 2.6（e）所示，其中下採樣因子為 3，表示每 3 個時間點只會輸出一個樣本。由於正向輸出中越往後的時間點涵蓋的資訊範圍越大，故在下採樣時會選擇較後面的時間點，因此第 3、6、9⋯⋯ 個時間點將會被保留。而反向輸出的下採樣方法則相反，如圖 2.6（f）所示。隨著時間點往前，所涵蓋的資訊範圍越大，故在下採樣時會選擇較前面的時間點，因此第 1、4、7⋯⋯ 個時間點將會被保留。

2.4.2 語者編碼器

語者編碼器由 LSTM 和線性層所組成，其目的是提供與語者相關的資訊，並確保同一語者的不同語音能產生相同的語者嵌入（speaker embedding）。在一般的多對多轉換情境中，使用語者的獨熱編碼（one-hot encoding）就足夠了；然而，AutoVC [28] 的使用情境是任意對任意，因此對於訓練過程中未見過的語者，需要透過語者編碼器來產生對應的語者嵌入。

doi:10.6342/NTU202402100

2.4.3 解碼器

解碼器由卷積層和 LSTM 所組成，首先將內容編碼器輸出的內容嵌入（content embedding）和語者編碼器輸出的語者嵌入進行上採樣（upsampling），以還原至原始輸入的時間長度，然後再將兩者串接在一起。其後半部分則包含後置網路（post network）的結構，將對初始輸出進行微調，以更好地重建時頻譜中的細節部分。而在訓練過程中，初始輸出和最終輸出都會用於計算自重建損失。

doi:10.6342/NTU202402100

第三章　研究方法

圖 3.1 描繪了本論文提出方法的整體流程。首先，我們從來源音檔中抽取音高和能量；接著，這些資料將分別通過音高轉換（pitch conversion）模型和能量轉換（energy conversion）模型進行風格轉換，這兩個模型均需一個目標歌手 ID 的輸入，以指示欲轉換至哪位目標歌手的歌唱風格；最後，模型將輸出具有目標歌手歌唱風格的音高和能量。其中，音高轉換模型輸出的音高曲線將會作為能量轉換模型的輸入，以確保音高和能量之間的一致性。以下將詳細說明音高轉換和能量轉換的方法。

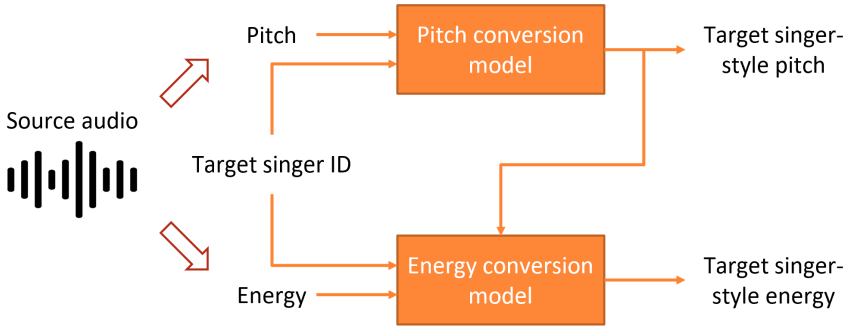


圖 3.1: 本論文提出方法之整體流程圖

3.1　音高轉換

首先，我們將介紹音高資料處理的方法，以及提出的音高轉換模型架構；接著，由於顫音（vibrato）是一種相對容易辨識的歌唱風格，我們將對其進行建模，以幫助模型更好地學習，同時進行平滑處理以避免不規則的波動；最後，我們將介紹訓練音高轉換模型時所使用的損失函數。

doi:10.6342/NTU202402100

3.1.1 資料處理

資料前處理的部分，首先，我們將來源音檔讀入後進行下採樣（downsampling），以獲得 16 kHz 的音檔；接著，我們使用 CREPE [13] 抽取音高，抽取過程中，音框跳距（hop size）為 5 毫秒，即音框率（frame rate）為 200 Hz，得到的音高曲線（pitch contour）$\mathbf{p} \in \mathbb{R}^T$ 包含 T 個音框；之後，我們先將音高範圍定義為 C1（32.7 Hz）到 B6（1975.5 Hz）之間，共跨越了 6 個八度（72 個半音），再將音高曲線 $\mathbf{p}$ 轉換為音高嵌入（pitch embedding）$\mathbf{p}_\mathrm{e} \in \mathbb{R}^{T\times 72}$，其中每個音高會以一個 72 維的向量表示，這個向量的每個維度對應到 C1（MIDI 編號 24）至 B6（MIDI 編號 95）之間的一個整數音樂數位介面（musical instrument digital interface, MIDI）編號；最後，我們使用線性插值（linear interpolation）來計算音高嵌入的數值。

例如，為了表示 MIDI 編號為 69.42 的音高，首先，我們找到 MIDI 編號 69 和 MIDI 編號 70 所對應的維度，分別是第 46 維和第 47 維；之後，根據線性插值的結果，將第 46 維嵌入數值設為 0.58，第 47 維嵌入數值設為 0.42，其餘則設為 0。這樣做的目的是確保音高嵌入的每個數值都介於 0 和 1 之間，並且每個音高嵌入向量 $\mathbf{p}_\mathrm{e}^{(i)}$ 的數值之和為 1，其中 i 代表音框索引（index），因此可以視為一種機率分佈（probability distribution）。

資料後處理的目的則是要將模型輸出的音高嵌入 $\hat{\mathbf{p}}_\mathrm{e} \in \mathbb{R}^{T\times 72}$ 轉換回輸出的音高曲線 $\hat{\mathbf{p}} \in \mathbb{R}^T$，我們採用了 CREPE [13] 的方法，將輸出音高嵌入 $\hat{\mathbf{p}}_\mathrm{e}$ 的每個數值作為權重，並計算這些權重與對應 MIDI 編號的加權平均值（weighted average），即可得到代表的音高。

3.1.2 模型架構

音高轉換的模型架構如圖 3.2 所示，其為一種類似於 AutoVC [28] 的自編碼器（autoencoder），並且具有信息瓶頸（information bottleneck）以過濾風格相關的資訊，主要包含編碼器（encoder）、解碼器（decoder）和可訓練的歌手嵌入（singer embedding）查詢表（lookup table, LUT）三個部分。其中，ConvNorm 代表一層一維卷積（convolution）後接著組正規化（group normalization）[34]，Down 和 Up 則分別表示下採樣和上採樣（upsampling），而括弧內表示的是該層的維度。

編碼器將音高嵌入 $\mathbf{p}_\mathrm{e}$ 作為輸入。首先，經過一層卷積核大小（kernel size）為

doi:10.6342/NTU202402100

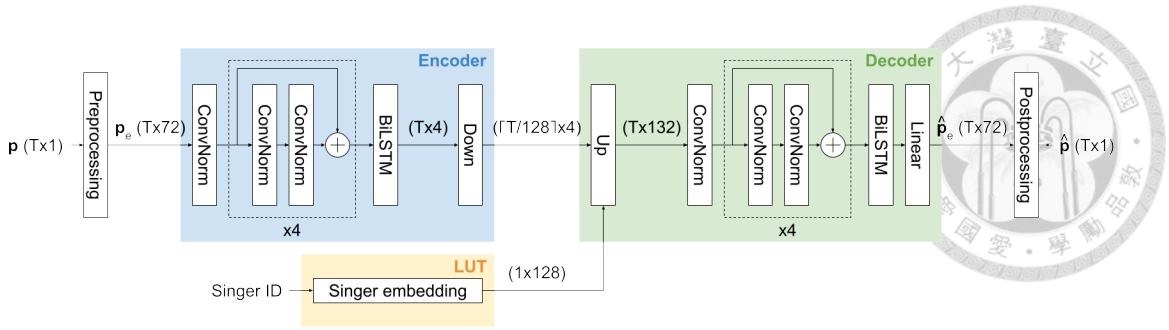


圖 3.2: 音高轉換模型架構圖

11 的一維卷積，以捕捉更長範圍的特徵；接著，通過 4 組殘差塊（residual block）[8]，每組包含 2 層卷積核大小為 5 的一維卷積；之後，通過每個方向輸出維度為 2 的雙向長短期記憶網路（bi-directional long short-term memory, BiLSTM），總共輸出 4 維；最後，與 AutoVC [28] 一樣，我們使用下採樣以塑造信息瓶頸，下採樣因子（downsampling factor）設為 128，即每 128 個時間點只輸出一個樣本，即使無法被 128 整除，仍會從餘下的部分輸出一個樣本。而 BiLSTM 的正向輸出和反向輸出的下採樣方法略有不同，以正向輸出為例，隨著時間的推移，模型所捕捉的資訊範圍逐漸擴大，因此在下採樣時傾向選擇後面時間點的輸出，換言之，輸出為第 128、256、384⋯⋯ 個時間點，而反向輸出則相反，因為越早的時間點所含資訊越多，因此輸出為第 1、129、257⋯⋯ 個時間點。由於原始的音框率為 200 Hz，因此每秒的信息瓶頸為 $4 \times 200 \div 128 = 6.25$ 維。

歌手嵌入查詢表儲存了每位歌手的特徵，並且可以在訓練過程中自動調整，由於我們的目標是任意對多（any-to-many），因此不需要像 AutoVC [28] 那樣具有歌手編碼器，來根據來源音檔產生對應的歌手嵌入。對於每位歌手，只需一個歌手嵌入即可。而歌手嵌入的維度設為 128 維。

解碼器將編碼器的輸出和歌手嵌入作為輸入。首先，將這兩組輸入進行上採樣，以匹配原始音高曲線的長度（T），方法是沿著時間軸重複這兩組輸入特徵，並將這些特徵串接在一起；接著，與編碼器類似，經過一層卷積核大小為 11 的一維卷積、4 組殘差塊和一層 BiLSTM 來處理串接後的特徵；最後，通過一層線性層產生最終的輸出音高嵌入 $\hat{\mathbf{p}}_e$。輸出音高嵌入 $\hat{\mathbf{p}}_e$ 的維度和音高嵌入 $\mathbf{p}_e$ 相同，即 $\hat{\mathbf{p}}_e \in \mathbb{R}^{T \times 72}$。

doi:10.6342/NTU202402100

3.1.3 顫音建模

為了促進模型學習與顫音相關的特徵，我們參考了 [17] 的方法，該方法首先對音高曲線進行一階差分（first-order difference），接著將其進一步轉換為功率時頻譜（power spectrogram）和顫音幅度曲線（vibrato extent contour），以凸顯顫音的特徵。更詳細的內容請參見 2.2 小節。

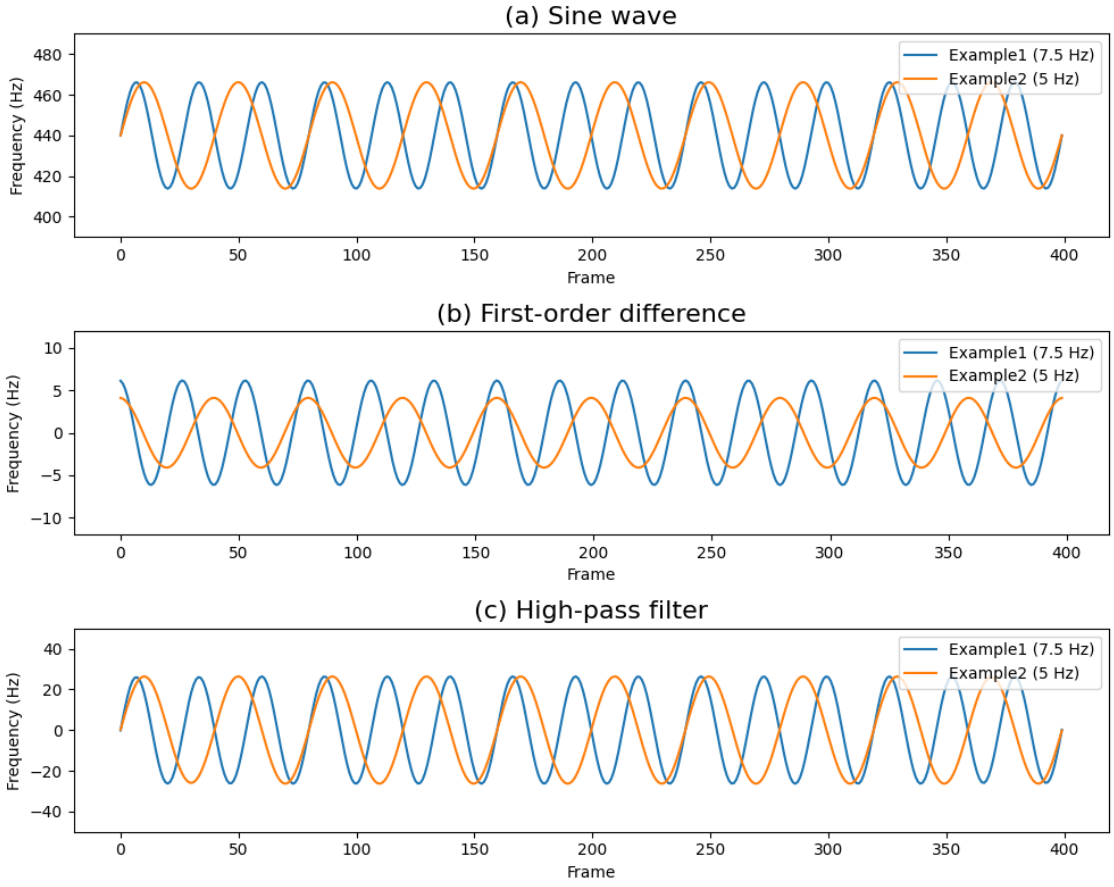


圖 3.3: 一階差分和高通濾波器之比較圖

然而，使用一階差分計算可能會引發一些問題。因為一階差分實際上類似於微分的概念，對一個正弦波進行微分時，內部與頻率相關的係數將被提取出來，乘上振幅的係數後影響整體曲線。這可能導致最終的顫音幅度曲線不僅與顫音振幅相關，還受到顫音頻率的影響。為了確保顫音幅度曲線代表的僅是顫音振幅，我們將音高曲線視為一種波形，並運用數位訊號處理（digital signal processing）的方法，通過高通濾波器（high-pass filter）強調曲線中相對高頻的成分。結果如圖 3.3 所示，圖 3.3（a）展示了兩段振幅相同但頻率不同的正弦波，其中頻率分別為 7.5 Hz 和 5 Hz，圖 3.3（b）為一階差分的結果，可以發現兩段曲線的振幅並不相

doi:10.6342/NTU202402100

同，而圖 3.3（c）為高通濾波器的結果，顯示兩段曲線的振幅相同。具體計算方法如式 3.1 所示：

$$\begin{aligned} \mathbf{p}_{\text{sharp}} &= \mathbf{p} - \text{sinc} * \mathbf{p}, \\ \mathbf{p}_{\text{FT}} &= \text{STFT}(\mathbf{p}_{\text{sharp}}), \\ \mathbf{p}_{\text{VibExt}} &= \text{vib\_frame\_max}(\mathbf{p}_{\text{FT}}), \end{aligned} \tag{3.1}$$

其中 $*$ 表示卷積運算，sinc 為 sinc 函數，其截止頻率（cutoff frequency）為 2 Hz，STFT 為短時距傅立葉變換（short-time Fourier transform, STFT），vib_frame_max 為一種音框間最大化運算，僅應用於頻率成分在 5 Hz 至 8 Hz 之間的頻率箱（frequency bin），這是顫音常見的頻率範圍。

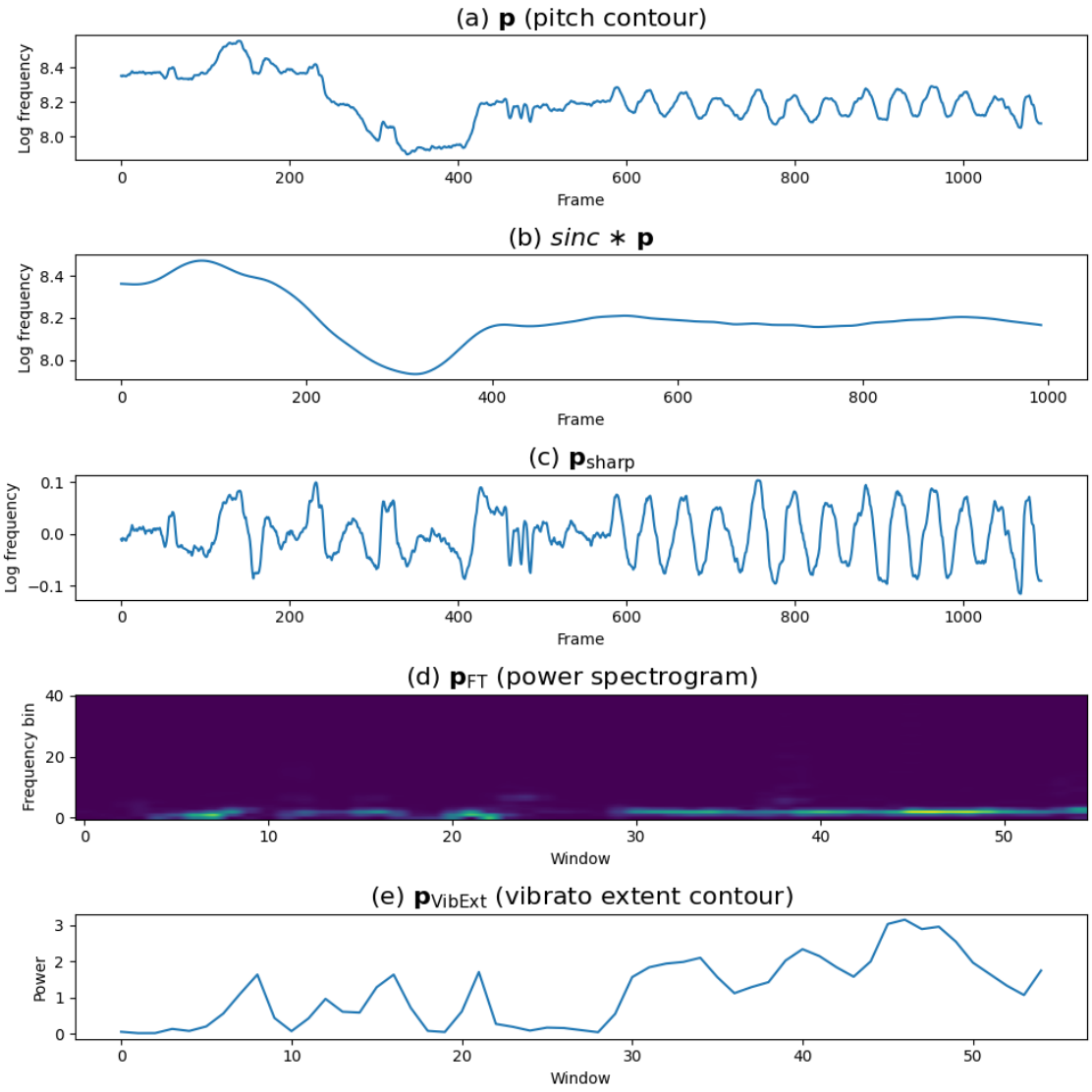


圖 3.4: 顫音建模方法說明圖

doi:10.6342/NTU202402100

圖 3.4 是顫音建模的方法說明圖。其中，圖 3.4（a）展示了一段範例音高曲線 $\mathbf{p}$。首先，將此曲線與 sinc 函數進行卷積運算，得到平滑的音高曲線，如圖 3.4（b）所示；接著，從原始音高曲線中減去卷積的結果，得到保留相對高頻成分的曲線 $\mathbf{p}_{\text{sharp}}$，如圖 3.4（c）所示。這是因為 sinc 函數在頻域（frequency domain）上相當於矩形函數，因此上述的操作相當於通過一個高通濾波器，過濾掉低頻成分。其餘的步驟與 [17] 的方法一樣，即先通過 STFT 得到功率時頻譜 $\mathbf{p}_{\text{FT}}$，如圖 3.4（d）所示，之後再經由 vib_frame_max 運算得到顫音幅度曲線 $\mathbf{p}_{\text{VibExt}}$，如圖 3.4（e）所示。其中，STFT 的快速傅立葉變換大小（FFT size）、窗口大小（window size）和音框跳距分別設為 80、80、20。

3.1.4 顫音幅度平滑

為了避免輸出的音高曲線中顫音振幅的不規則波動，以達到聽感上的自然流暢，我們將對輸出的顫音幅度曲線 $\hat{\mathbf{p}}_{\text{VibExt}}$ 進行平滑處理。計算方法如式 3.2 所示：

$$\begin{aligned} \hat{\mathbf{p}}_{\text{IsVib}} &= \text{Thresholding}(\hat{\mathbf{p}}_{\text{VibExt}}, \text{thres}_{\text{p}}), \\ \hat{\mathbf{p}}_{\text{VibDiff}} &= \text{Diff}(\hat{\mathbf{p}}_{\text{VibExt}}) \times \hat{\mathbf{p}}_{\text{IsVib}}, \end{aligned} \tag{3.2}$$

其中 Thresholding 表示一個閾值函數，會對 $\hat{\mathbf{p}}_{\text{VibExt}}$ 中的每個元素（除了最後一個）進行操作。對於索引 i，如果第 i 個元素（$\hat{p}^{(i)}_{\text{VibExt}}$）和第 $i+1$ 個元素（$\hat{p}^{(i+1)}_{\text{VibExt}}$）同時大於給定的閾值 thres_{p}，則輸出為 1，否則為 0。我們將閾值設置為 0.75，用於判斷一個音框是否包含顫音。而 Diff 則表示一階差分。

圖 3.5 是顫音幅度平滑的方法說明圖。其中，圖 3.5（a）展示了一段範例輸出音高曲線 $\hat{\mathbf{p}}$，其對應的顫音幅度曲線 $\hat{\mathbf{p}}_{\text{VibExt}}$ 如圖 3.5（b）所示。首先，將此曲線經由 Thresholding 運算得到 $\hat{\mathbf{p}}_{\text{IsVib}}$，如圖 3.5（c）所示；接著，將 $\hat{\mathbf{p}}_{\text{VibExt}}$ 進行一階差分運算，結果如圖 3.5（d）所示；最後，再將一階差分的結果和 $\hat{\mathbf{p}}_{\text{IsVib}}$ 相乘，即可得到輸出顫音幅度變化量 $\hat{\mathbf{p}}_{\text{VibDiff}}$，如圖 3.5（e）所示。換言之，只有當此變化量對應到的兩個音框皆包含顫音時才會被保留，否則會變為 0。因此，剩餘的變化量均屬於輸出音高曲線中包含顫音的部分，我們希望這些變化量越接近 0 越好，以達到平滑的效果。

doi:10.6342/NTU202402100

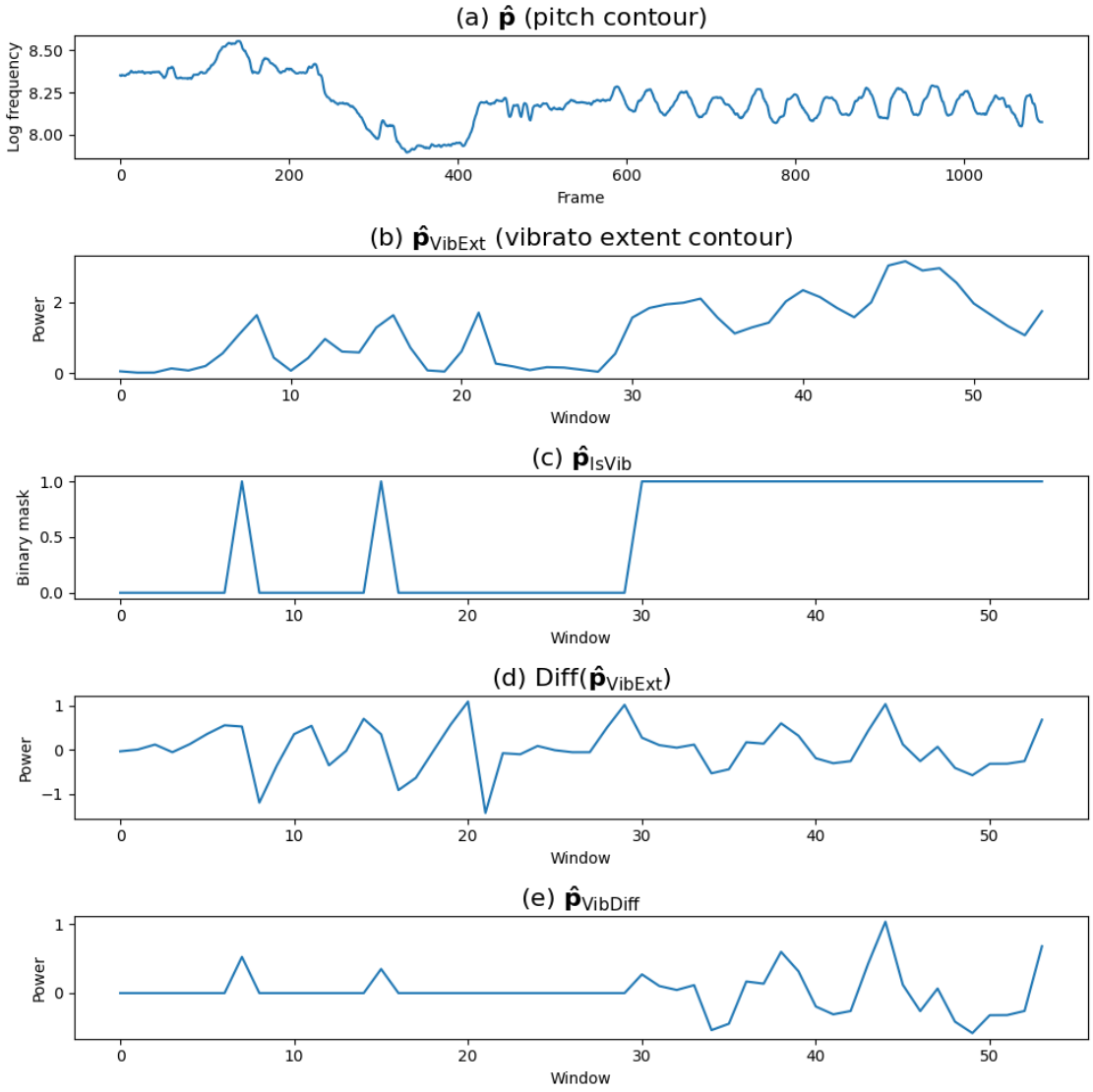


圖 3.5: 顫音幅度平滑方法說明圖

3.1.5 損失函數

與 AutoVC [28] 類似，在訓練過程中，模型透過編碼器輸出的內容相關資訊，以及歌手 ID 對應的歌手相關資訊，重建出輸入的音高曲線 $\mathbf{p}$ 和音高嵌入 $\mathbf{p}_e$，而在測試時，只需改變輸入的歌手 ID 就能達到風格轉換的效果。訓練時所採用的損失函數包含兩個部分。

第一部分為音高重建損失（reconstruction loss）$\mathcal{L}^{p}_{\text{Recon}}$，以幫助模型重建原始

doi:10.6342/NTU202402100

的輸入，如式 3.3 所示：

$$\begin{aligned}\mathcal{L}^{\mathrm{p}}_{\mathrm{ReconE}} &= \mathrm{BCE}(\mathbf{p}_{\mathrm{e}}, \hat{\mathbf{p}}_{\mathrm{e}}),\\ \mathcal{L}^{\mathrm{p}}_{\mathrm{ReconC}} &= \mathrm{RMSE}(\mathbf{p}, \hat{\mathbf{p}}),\\ \mathcal{L}^{\mathrm{p}}_{\mathrm{Recon}} &= \lambda^{\mathrm{p}}_{\mathrm{ReconE}}\mathcal{L}^{\mathrm{p}}_{\mathrm{ReconE}} + \lambda^{\mathrm{p}}_{\mathrm{ReconC}}\mathcal{L}^{\mathrm{p}}_{\mathrm{ReconC}},\end{aligned} \tag{3.3}$$

其中 BCE 代表二元交叉熵損失（binary cross-entropy loss），RMSE 代表均方根誤差損失（root mean square error loss）。$\mathcal{L}^{\mathrm{p}}_{\mathrm{ReconE}}$ 表示音高嵌入的重建損失，$\mathcal{L}^{\mathrm{p}}_{\mathrm{ReconC}}$ 則表示音高曲線的重建損失，$\lambda^{\mathrm{p}}_{\mathrm{ReconE}}$ 和 $\lambda^{\mathrm{p}}_{\mathrm{ReconC}}$ 為這兩個損失函數的權重，根據實驗結果分別設為 1 和 10。而上標 p 代表這些損失函數和權重用於訓練音高轉換模型。

第二部分為顫音損失 $\mathcal{L}^{\mathrm{p}}_{\mathrm{Vib}}$，以幫助模型學到顫音相關的特徵，如式 3.4 所示：

$$\begin{aligned}\mathcal{L}^{\mathrm{p}}_{\mathrm{FT}} &= \mathrm{RMSE}(\mathbf{p}_{\mathrm{FT}}, \hat{\mathbf{p}}_{\mathrm{FT}}),\\ \mathcal{L}^{\mathrm{p}}_{\mathrm{VibExt}} &= \mathrm{RMSE}(\mathbf{p}_{\mathrm{VibExt}}, \hat{\mathbf{p}}_{\mathrm{VibExt}}),\\ \mathcal{L}^{\mathrm{p}}_{\mathrm{Smooth}} &= \mathrm{RMS}(\hat{\mathbf{p}}_{\mathrm{VibDiff}}),\\ \mathcal{L}^{\mathrm{p}}_{\mathrm{Vib}} &= \lambda^{\mathrm{p}}_{\mathrm{FT}}\mathcal{L}^{\mathrm{p}}_{\mathrm{FT}} + \lambda^{\mathrm{p}}_{\mathrm{VibExt}}\mathcal{L}^{\mathrm{p}}_{\mathrm{VibExt}} + \lambda^{\mathrm{p}}_{\mathrm{Smooth}}\mathcal{L}^{\mathrm{p}}_{\mathrm{Smooth}},\end{aligned} \tag{3.4}$$

其中 RMS 代表均方根。而 $\lambda^{\mathrm{p}}_{\mathrm{FT}}$、$\lambda^{\mathrm{p}}_{\mathrm{VibExt}}$ 和 $\lambda^{\mathrm{p}}_{\mathrm{Smooth}}$ 為損失函數的權重，根據實驗結果皆設為 0.1。

最後，用於訓練音高轉換模型的總音高損失函數 $\mathcal{L}^{\mathrm{p}}$ 即為音高重建損失 $\mathcal{L}^{\mathrm{p}}_{\mathrm{Recon}}$ 和顫音損失 $\mathcal{L}^{\mathrm{p}}_{\mathrm{Vib}}$ 的總和。

3.2 能量轉換

首先，我們將介紹能量資料處理的方法，以及提出的能量轉換模型架構；接著，我們同樣對顫音進行建模和平滑處理，這部分的處理方法與音高轉換時相同，因此不再重複說明；最後，我們將介紹訓練能量轉換模型時所使用的損失函數。

doi:10.6342/NTU202402100

3.2.1 資料處理

在資料前處理的部分，我們首先將來源音檔下採樣至 16 kHz；接著，我們以 5 毫秒的音框跳距計算每個音框的均方根數值，再以 10 為底取對數得到能量曲線（energy contour）$\mathbf{e} \in \mathbb{R}^T$；之後，我們先將能量範圍定義為 0.0001（$10^{-4}$）到 1（$10^0$）之間，並將此範圍以對數尺度切分為 128 個區段，再將能量曲線 $\mathbf{e}$ 轉換為能量嵌入（energy embedding）$\mathbf{e}_e \in \mathbb{R}^{T\times 128}$，其中每個能量會以一個 128 維的向量表示，這個向量的 128 維即對應切分後的 128 個區段；最後，我們使用與 3.1.1 小節相同的線性插值方法計算能量嵌入的數值。

資料後處理的部分同樣是要將模型輸出的能量嵌入 $\hat{\mathbf{e}}_e \in \mathbb{R}^{T\times 128}$ 轉換回輸出的能量曲線 $\hat{\mathbf{e}} \in \mathbb{R}^T$，我們採用與 3.1.1 小節相同的加權平均值方法，以此得到代表的能量。

3.2.2 模型架構

圖 3.6 呈現了能量轉換的模型架構，其與音高轉換的模型相似，不同之處在於輸入和輸出改為能量曲線。此外，值得注意的是，該模型需要音高曲線 $\mathbf{p}$ 作為額外的輸入，這有助於確保輸出的能量曲線與其對應的音高曲線能相互配合。這種方法特別適合於提升顫音的轉換效果，因為顫音通常涉及音高曲線和能量曲線一致的波動，當這兩個曲線的波動不一致時，歌唱風格將不容易被人們識別，而這將在 5.3 和 5.4 小節中進行驗證。

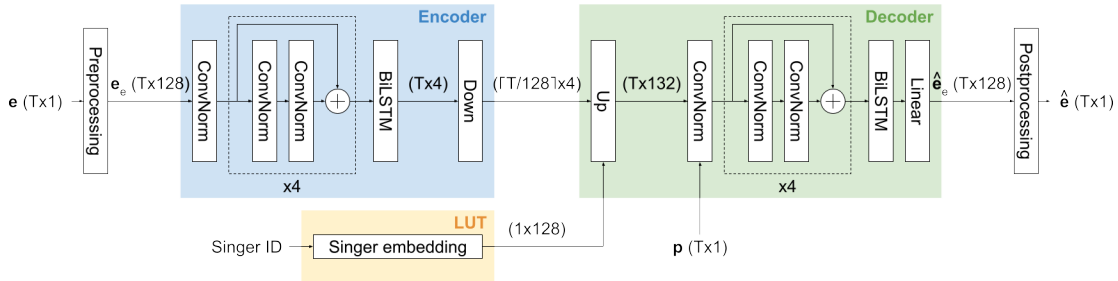


圖 3.6: 能量轉換模型架構圖

在實際操作中，訓練時，我們將真實的音高曲線 $\mathbf{p}$ 作為輔助資訊輸入到能量轉換模型中。而在測試時，我們首先進行音高轉換，得到轉換後的音高曲線 $\hat{\mathbf{p}}$，接著再將 $\hat{\mathbf{p}}$ 輸入到能量轉換模型中。由於 $\hat{\mathbf{p}}$ 的歌唱風格已經轉換，理想情況下，

doi:10.6342/NTU202402100

並不會泄漏與來源歌手歌唱風格相關的資訊，因此能量轉換模型仍然可以輸出具有目標歌手歌唱風格的能量曲線。

編碼器、解碼器以及歌手嵌入查詢表的架構基本上與音高轉換模型相同，唯一的區別在於解碼器部分。在解碼器中，我們會將音高曲線與上採樣後的結果串接在一起，其餘的部分皆與音高轉換模型保持一致。

3.2.3 損失函數

損失函數的部分與訓練音高轉換模型時所使用的相同，換句話說，總能量損失函數 $\mathcal{L}^{e}$ 即為能量重建損失 $\mathcal{L}^{e}_{\text{Recon}}$ 和顫音損失 $\mathcal{L}^{e}_{\text{Vib}}$ 之和，這些損失函數如式 3.5 所示：

$$\begin{aligned}\mathcal{L}^{e}_{\text{Recon}} &= \lambda^{e}_{\text{ReconE}}\mathcal{L}^{e}_{\text{ReconE}} + \lambda^{e}_{\text{ReconC}}\mathcal{L}^{e}_{\text{ReconC}},\\ \mathcal{L}^{e}_{\text{Vib}} &= \lambda^{e}_{\text{FT}}\mathcal{L}^{e}_{\text{FT}} + \lambda^{e}_{\text{VibExt}}\mathcal{L}^{e}_{\text{VibExt}} + \lambda^{e}_{\text{Smooth}}\mathcal{L}^{e}_{\text{Smooth}},\\ \mathcal{L}^{e} &= \mathcal{L}^{e}_{\text{Recon}} + \mathcal{L}^{e}_{\text{Vib}},\end{aligned} \tag{3.5}$$

其中 $\lambda^{e}_{\text{ReconE}}$、$\lambda^{e}_{\text{ReconC}}$、$\lambda^{e}_{\text{FT}}$、$\lambda^{e}_{\text{VibExt}}$ 和 $\lambda^{e}_{\text{Smooth}}$ 皆為損失函數的權重，根據實驗結果分別設為 1、10、0.01、0.01、0.01。而針對能量曲線以判斷音框是否包含顫音的閾值 thres_{e}，我們設為 20，其餘細節請參見 3.1.3 至 3.1.5 小節。

值得注意的是，在能量轉換模型的損失函數中我們也包括了顫音損失 $\mathcal{L}^{e}_{\text{Vib}}$，這是因為顫音通常包含音高曲線和能量曲線同步的波動。圖 3.7 是一個實際的例子，顯示了在第 800 到 1000 個音框之間存在一個顫音，導致兩個曲線同時波動。這樣的特徵可以更好地利用 $\mathcal{L}^{e}_{\text{Vib}}$ 來學習。

3.3 單階段轉換

3.1 和 3.2 小節描述的是一個二階段轉換的模型，首先透過音高轉換模型進行音高曲線轉換，然後將轉換後的音高曲線輸入能量轉換模型，進行能量曲線的轉換。在這個小節，我們也提出了另一個單階段轉換的模型，即可以同時轉換音高曲線和能量曲線。我們將先介紹音高和能量資料處理的方法，以及提出的單階段轉換模型架構，接著介紹訓練單階段轉換轉換模型時所使用的損失函數。

doi:10.6342/NTU202402100

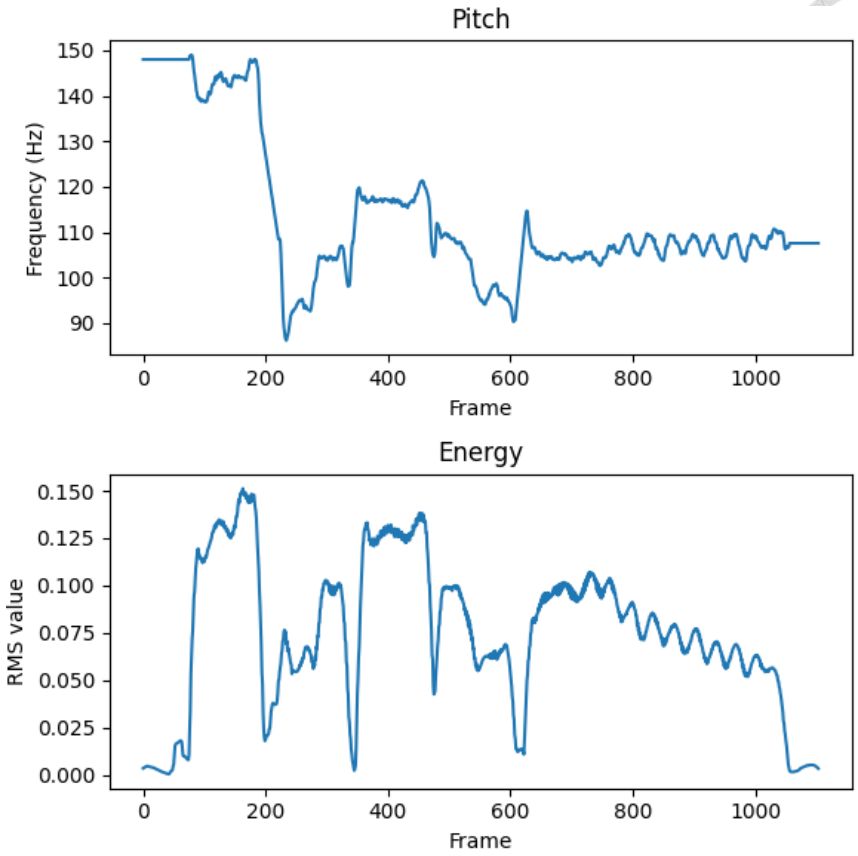


圖 3.7: 音高曲線和能量曲線之顫音範例圖

3.3.1 資料處理

資料前處理的部分，我們採用了 3.1.1 和 3.2.1 小節所描述的方法。首先，我們從來源音檔中提取出音高曲線 $\mathbf{p} \in \mathbb{R}^T$ 和能量曲線 $\mathbf{e} \in \mathbb{R}^T$；接著，利用線性插值的方法將兩者分別轉換為音高嵌入 $\mathbf{p}_e \in \mathbb{R}^{T\times 72}$ 和能量嵌入 $\mathbf{e}_e \in \mathbb{R}^{T\times 128}$；最後，將這兩者串接在一起，即可形成模型的輸入。

至於資料後處理的部分，我們會依據維度將模型的輸出拆解為輸出的音高嵌入 $\hat{\mathbf{p}}_e \in \mathbb{R}^{T\times 72}$ 和輸出的能量嵌入 $\hat{\mathbf{e}}_e \in \mathbb{R}^{T\times 128}$，然後再利用加權平均值的方法，將這兩者轉換為輸出的音高曲線 $\hat{\mathbf{p}} \in \mathbb{R}^T$ 和輸出的能量曲線 $\hat{\mathbf{e}} \in \mathbb{R}^T$。

3.3.2 模型架構

單階段轉換的模型架構如圖 3.8 所示，其與音高轉換模型很相似，唯一的差別在於輸入和輸出被替換為音高嵌入和能量嵌入串接的結果。此外，由於輸入和輸出的維度增加，我們也將 BiLSTM 每個方向的輸出維度調整為 4，因此總共會

doi:10.6342/NTU202402100

輸出 8 維，以允許更多的資訊能通過信息瓶頸。其餘的架構皆與音高轉換模型相同。

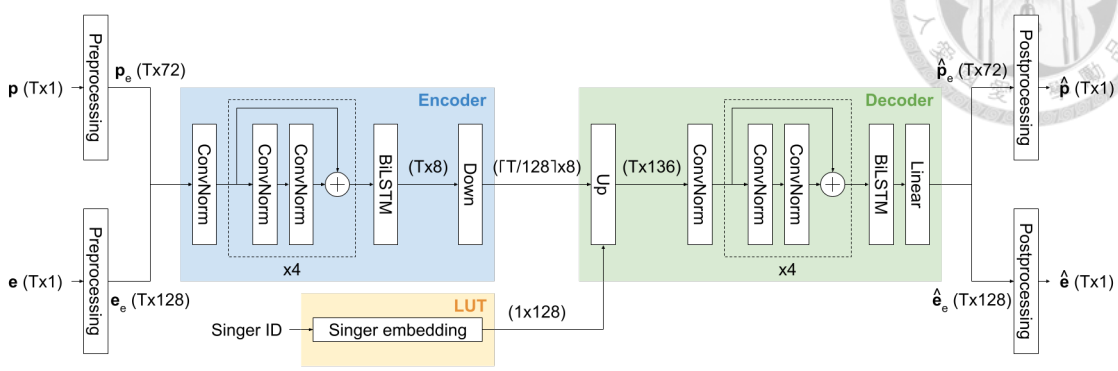


圖 3.8: 單階段轉換模型架構圖

3.3.3 損失函數

訓練單階段模型所使用的總損失函數 $\mathcal{L}^{\text{one}}$ 為總音高損失函數 $\mathcal{L}^{\text{p}}$ 和總能量損失函數 $\mathcal{L}^{\text{e}}$ 之和，其中各損失函數的權重以即是否包含顫音的閾值，皆與前述設定相同。相關細節請參見 3.1.5 和 3.2.3 小節。

doi:10.6342/NTU202402100

第四章　實驗相關設定

本章節將詳細介紹本論文的實驗相關設定，包括資料集、評量指標、實驗環境、實驗參數設定以及實驗項目。首先，我們將介紹所使用的資料集，包括資料的特性和處理方法；接著，將闡述評量指標的選擇，這些指標用於衡量實驗結果的效能；之後，我們將說明實驗環境的配置，以及實驗參數的設定，包含各種超參數和實作細節；最後，我們將概述各項實驗的內容。

4.1　資料集

在本論文中，我們總共使用到了四個資料集，分別為 Opencpop [33]、TONAS [24]、M4Singer [37] 和 OpenSinger [10]。以下將依序介紹這些資料集。

4.1.1　Opencpop

Opencpop [33] 包含了 100 首中文流行歌的無伴奏歌唱，總計約 5.2 小時的音檔，由一位專業女歌手演唱。這位女歌手的歌唱風格為不常使用顫音（vibrato）的技巧，如圖 4.1 所示。

在訓練和測試模型時，我們都使用了這個資料集。而在資料切分的過程中，我們首先依照官方提供的方式，將其中的 5 首歌劃分為測試集；之後，將剩餘的 95 首歌按照 8：1 的比例切割為訓練集和驗證集。我們以歌曲為單位進行切分，以確保同一首歌的所有片段都位於同一個子集中。

doi:10.6342/NTU202402100

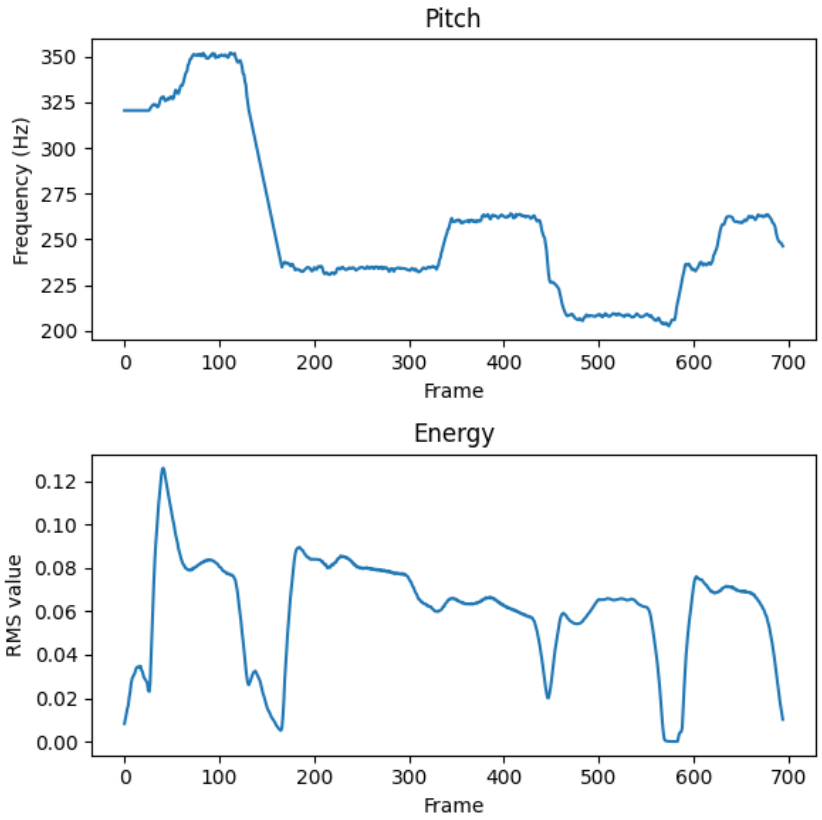


圖 4.1: Opencpop 之音高曲線和能量曲線範例圖

4.1.2 TONAS

TONAS [24] 包含了 72 首佛朗明哥風格的無伴奏歌曲，總計約 0.34 小時的音檔。這些音檔的演唱者並非完全相同，但考慮到本論文的方法是進行音高轉換（pitch conversion）和能量轉換（energy conversion），所以音色上的差異應不會對結果產生太大的影響。因此，我們將整個資料集視為同一位歌手的演唱，此歌手代表佛朗明哥這種歌唱風格，即具有極為強烈的顫音，如圖 4.2 所示。

在訓練和測試模型時，我們都使用了這個資料集。而在資料切分的過程中，我們根據 8：1：1 的比例將所有歌曲切割成訓練集、驗證集和測試集，並以歌曲為單位進行切分。

4.1.3 M4Singer

M4Singer [37] 包含了 700 首中文流行歌的無伴奏歌唱，總計約 29.77 小時的音檔，由 20 位歌手演唱。這 20 位歌手涵蓋了四個聲部，分別是女高音（soprano）、

doi:10.6342/NTU202402100

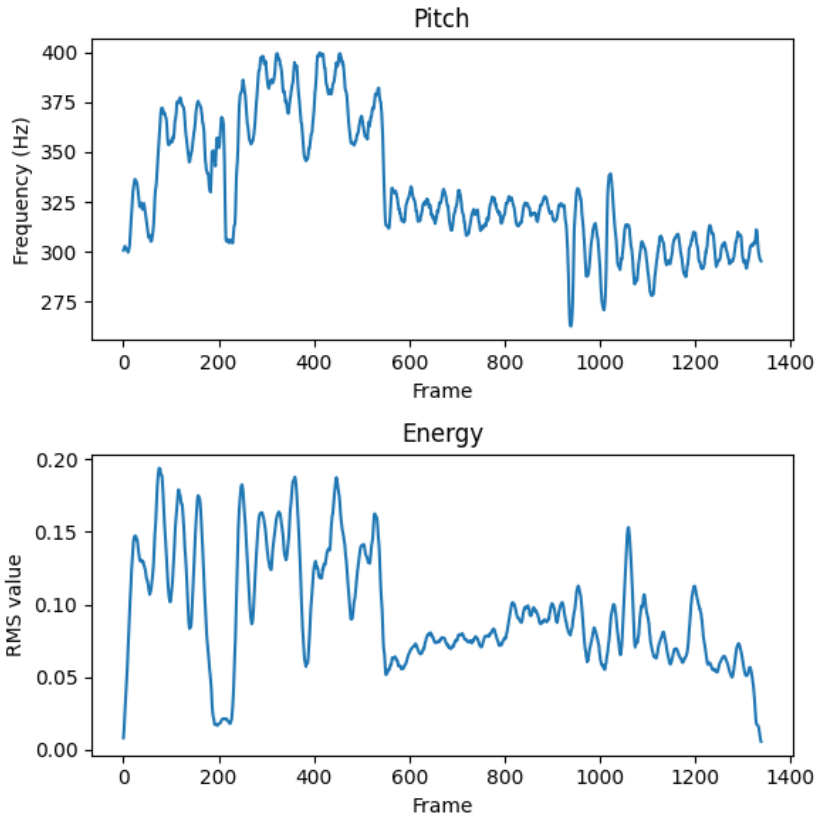


圖 4.2: TONAS 之音高曲線和能量曲線範例圖

女低音（alto）、男高音（tenor）和男低音（bass），詳細資訊如表 4.1 所示，包含每位歌手的平均音高、音高範圍和歌曲總時長。

我們在訓練和測試模型時都有使用這個資料集。而在資料切分的過程中，我們以歌曲為單位，將所有歌曲根據 8：1：1 的比例切割成訓練集、驗證集和測試集。

4.1.4 OpenSinger

OpenSinger [10] 包含了 50 小時的中文流行歌無伴奏歌唱，總共由 76 位歌手演唱。因為每位歌手的平均資料量並不如 M4Singer [37] 那麼多，所以我們僅在測試模型時使用了這個資料集。因此，其代表了訓練過程中未曾見過的歌手歌唱風格。

doi:10.6342/NTU202402100

表 4.1: M4Singer 之歌手資訊表 [37]

Gender	Singer ID	Avg. pitch	Note range	# Hours
Female	Alto-1	325.76	55 - 72 (G3, 196.0Hz - C5, 523.3Hz)	0.97
	Alto-2	317.80	54 - 72 (F#3, 185.0Hz - C5, 523.3Hz)	0.94
	Alto-3	352.19	54 - 73 (F#3, 185.0Hz - C#5, 554.4Hz)	0.91
	Alto-4	292.16	51 - 72 (D#3, 155.6Hz - C5, 523.3Hz)	1.05
	Alto-5	324.15	52 - 73 (E3, 164.8Hz - C#5, 554.4Hz)	2.63
	Alto-6	301.26	50 - 73 (D3, 146.8Hz - C#5, 554.4Hz)	2.53
	Alto-7	289.22	50 - 72 (D3, 146.8Hz - C5, 523.3Hz)	1.68
	Soprano-1	465.89	61 - 77 (C#4, 277.2Hz - F5, 698.5Hz)	0.88
	Soprano-2	475.63	63 - 77 (D#4, 311.1Hz - F5, 698.5Hz)	0.62
	Soprano-3	513.47	63 - 78 (D#4, 311.1Hz - F#5, 740.0Hz)	1.63
Male	Tenor-1	283.47	51 - 69 (D#3, 155.6Hz - A4, 440.0Hz)	1.12
	Tenor-2	222.33	45 - 67 (A2, 110.0Hz - G4, 392.0Hz)	1.15
	Tenor-3	219.07	43 - 67 (G2, 98.0Hz - G4, 392.0Hz)	1.32
	Tenor-4	214.61	45 - 67 (A2, 110.0Hz - G4, 392.0Hz)	1.00
	Tenor-5	177.05	43 - 64 (G2, 98.0Hz - E4, 329.6Hz)	1.76
	Tenor-6	173.86	40 - 63 (E2, 82.4Hz - D#4, 311.1Hz)	1.12
	Tenor-7	178.94	41 - 65 (F2, 87.3Hz - F4, 349.2Hz)	2.41
	Bass-1	109.98	35 - 54 (B1, 61.7Hz - F#3, 185.0Hz)	2.48
	Bass-2	110.93	38 - 55 (D2, 73.4Hz - G3, 196.0Hz)	2.14
	Bass-3	109.72	37 - 51 (C#2, 69.3Hz - D#3, 155.6Hz)	1.43

4.2 評量指標

本論文的評量標準可分為客觀指標和主觀指標兩個部分。客觀指標是利用量化標準來評量轉換結果的優劣，而主觀指標則是透過問卷調查，邀請受測者對轉換結果進行評分。接下來將分別詳細介紹這兩種指標。

4.2.1 客觀指標

我們參考了語者驗證（speaker verification）的方法 [4]，其旨在從音檔中驗證語者的身份。該方法的具體步驟為：首先，訓練一個語者編碼器（speaker encoder），將音檔作為輸入並產生固定維度的語者嵌入（speaker embedding）；接著，將待識別的音檔輸入模型，產生對應的語者嵌入；最後，通過比對這個語者嵌入與目標語者的平均嵌入是否足夠接近，即可確定語者的身份。

由於本論文旨在評量音高曲線（pitch contour）和能量曲線（energy contour）中的歌唱風格，因此我們分別使用音高嵌入（pitch embedding）和能量嵌入（energy embedding）作為輸入，透過語者驗證的方法訓練兩個編碼器。我們採用 Thin

doi:10.6342/NTU202402100

表 4.2: Thin ResNet-34 之模型架構表 [3]

Layer name	Thin ResNet-34
Conv1	7×7, 16, stride 2 3×3, max pool, stride 2
Conv2	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 3$, stride 1
Conv3	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 4$, stride 2
Conv4	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 6$, stride 2
Conv5	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$, stride 2
	Self-attentive pool
FC	9×1, 512, stride 1

ResNet-34 [3] 作為模型，其架構如表 4.2 所示，該模型與一般的 ResNet-34 相似，但通道（channel）數為原本的四分之一，以降低計算成本。此外，其採用了自注意力池化（self-attentive pooling）[1] 來聚合時間上的資訊，即通過自注意力機制計算每個時間點的重要程度。考慮到輸入變更為音高嵌入或能量嵌入，我們將模型中所有的二維卷積（convolution）都替換為一維卷積。

訓練模型所使用的損失函數為 AM-Softmax [32]，具體如式 4.1 所示：

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s(\cos(\theta_{y_i,i})-m)}}{e^{s(\cos(\theta_{y_i,i})-m)} + \sum_{j\neq y_i} e^{s(\cos(\theta_{j,i}))}}\right) \tag{4.1}$$

其中 N 表示批次大小（batch size），$\cos(\theta_{j,i})$ 代表最後一層的第 j 個權重向量和前一層的第 i 個輸出嵌入之內積，這兩個向量皆已進行正規化。m 表示餘弦邊界（cosine margin），用以確保向量之間有足夠的差距，而 s 是比例因子（scale factor），用以避免梯度過小的情況，同時有助於加速模型的收斂。我們將 m 和 s 分別設為 0.3 和 30。

訓練模型所使用的資料集包括 Opencpop [33]、TONAS [24] 和 M4Singer [37] 的訓練集，共 22 位歌手。而經訓練後的音高編碼器和能量編碼器在這三個資料集的測試集上，得到的相等錯誤率（equal error rate, EER）分別為 31.96% 和 24.13%。我們認為這可能是由於訓練資料量不足，以及僅使用音高曲線或能量曲線來辨識歌手相比直接使用歌聲更加困難，因而導致 EER 表現不佳。未來可再改進此部分，但目前我們仍將基於這些模型進行客觀評量。

在評量過程中，我們利用已經訓練好的兩個編碼器，分別從音高曲線和能量

doi:10.6342/NTU202402100

曲線中提取固定維度的歌手嵌入（singer embedding）。我們會先提取轉換結果的歌手嵌入，以及目標歌手的平均歌手嵌入，接著再計算兩者之間的餘弦相似度（cosine similarity），此即為客觀指標。

4.2.2 主觀指標

主觀評量的部分，我們將注意力轉移到提出的歌唱風格轉換模型實際應用於歌聲轉換（singing voice conversion, SVC）模型時的表現上。我們選用了 Diff-SVC[1] 來進行評量，這是一個基於擴散過程（diffusion process）的任意對一（any-to-one）SVC 模型。在每個評量中，我們比較了兩種設置。第一種是直接運行 Diff-SVC，將歌聲轉換為目標歌手的音色。第二種則是將轉換後的音高曲線和能量曲線輸入 Diff-SVC ，然後再進行歌聲轉換。換言之，在第一種設置中，我們僅轉換了音色；而在第二種設置中，除了音色之外，我們還進一步使用提出的歌唱風格轉換模型轉換了音高曲線和能量曲線。

圖 4.3: 主觀評量問卷範例圖

由於 Diff-SVC 是一個任意對一的 SVC 模型，我們首先選擇了四位目標歌手來訓練各自的 Diff-SVC 模型，這四位歌手分別是 Opencpop [33] 中的唯一女歌手，

[1] https://github.com/prophesier/diff-svc

doi:10.6342/NTU202402100

表 4.3: 受測者整體資訊表

Total subjects	10
Male / Female	9 / 1
Age range	20 - 40 years

以及 M4Singer [37] 中的 Alto-1、Tenor-3 和 Tenor-7。接著，我們進行了比較平均意見分數（comparison mean opinion score, CMOS） [20] 評量，此即為主觀指標。圖 4.3 呈現了主觀評量問卷的一組問題範例，在每組問題中，我們要求受測者先聆聽兩段轉換後的音檔片段（一段僅轉換了音色，另一段則音色、音高曲線和能量曲線都轉換了），以及目標歌手的一段音檔片段。然後，受測者需根據兩段音檔片段的自然度和與目標歌手歌唱風格的相似度，評價兩段轉換後音檔之間的相對偏好。CMOS 的評分範圍為 -3 到 3，受測者相關資訊則如表 4.3 所示。

4.3 實驗環境

本論文使用之機器規格如下：

- CPU：Intel(R) Core(TM) i7-6800K CPU @ 3.40GHz
- RAM：94 GB
- GPU：NVIDIA GeForce GTX 1080 Ti

在這個實驗環境下，訓練一個音高轉換模型或能量轉換模型需要大約 20 個小時的時間。

4.4 實驗參數設定

我們將來源音檔重新取樣為 16 kHz，並利用 CREPE [13] 來提取音高曲線，窗口大小（window size）和音框跳距（hop size）分別設為 1024 和 80。我們使用 0.05 的信心分數閾值來區分有聲音與無聲音的音框（frame）。對於無聲音的音框，我們則以兩個最接近的有聲音音框之音高進行線性插值（linear interpolation），以替換其音高數值，這是因為在計算損失函數時，我們會對音高曲線進行短時距傅立

doi:10.6342/NTU202402100

葉變換（short-time Fourier transform, STFT），所以需要避免有不連續曲線的狀況。至於能量曲線的部分，我們使用相同的窗口大小和音框跳距來計算每個音框的均方根（root mean square）數值。

我們將所有隱藏層（hidden layer）的維度設置為 128，每組使用 16 個通道進行組正規化（group normalization）[34]。訓練模型時的超參數設定如下：

- 批次大小：16
- 訓練步數（training step）：400k
- 優化器（optimizer）：AdamW
- 學習率（learning rate）：0.0001
- 權重衰減（weight decay）：0.0001

在訓練音高轉換模型的過程中，我們採用了隨機移調的資料擴增（data augmentation）方法，即將整個音高曲線在 C1 到 B6 這個範圍內，以半音為單位隨機進行平移。這有助於模型學習與絕對音高無關的歌唱風格。

4.5 實驗項目

- 實驗一：單階段轉換與二階段轉換模型之比較
  我們將 3.1 和 3.2 小節中提出的二階段轉換模型，以及 3.3 小節中提出的單階段轉換模型進行比較，評量內容包含客觀指標和聽感評量。
- 實驗二：顫音建模與顫音幅度平滑之消融實驗
  我們將對 3.1.3 和 3.1.4 小節中提出的方法進行消融實驗，並與原先提出的模型進行比較，評量內容包含客觀指標和聽感評量。
- 實驗三：有無提供音高曲線對能量轉換模型之影響
  我們在未提供音高曲線的情況下訓練了能量轉換模型，並與原先提出的模型進行比較，評量內容包含客觀指標和聽感評量。

doi:10.6342/NTU202402100

- 實驗四：主觀評量指標之結果分析
  我們透過問卷的方式邀請受測者進行主觀評量，並分析主觀指標的整體結果以及目標歌手為不同性別時的結果。

- 實驗五：任意對多情境下之案例分析
  針對任意對多（any-to-many）的情境，我們進行了實際案例分析，試圖了解提出的模型對音高曲線、能量曲線和聽感的實際影響，並進一步分析模型的優缺點。

- 實驗六：歌唱風格轉換與顫音轉換之比較
  我們將提出的歌唱風格轉換模型與僅轉換顫音的方法進行比較，評量內容包含客觀指標和聽感評量。

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

第五章　實驗結果與探討

本章節將呈現各項實驗的結果，並對其進行分析與討論，以探討本論文的貢獻和價值，並為未來的相關研究提供參考和建議。

5.1　實驗一：單階段轉換與二階段轉換模型之比較

在 3.1 和 3.2 小節中，我們提出了一個二階段轉換的模型（以下簡稱為 Proposed），即先將音高曲線（pitch contour）轉換到目標歌手的歌唱風格，然後根據轉換後的音高曲線，再將能量曲線（energy contour）也轉換到目標歌手的歌唱風格。而在 3.3 小節中，我們將兩個模型合併成一個，提出了單階段轉換的模型（以下簡稱為 Proposed (one-stage)），即同時將音高曲線和能量曲線進行轉換。在這個實驗中，我們將比較這些模型在客觀指標和聽感上的差異。此外，我們將來源音檔的音高曲線和能量曲線作為下界（以下簡稱為 Source），因為其代表了轉換前的結果；而目標歌手的音高曲線和能量曲線則作為上界（以下簡稱為 Target），因為其代表了同一歌手之間的相似度。

在客觀評量的部分，我們分別評量了兩種情境：第一種是使用訓練過程中見過的歌手作為來源音檔的歌手和目標歌手（以下簡稱為 seen-to-seen），即多對多（many-to-many）的情境；第二種是使用訓練過程中未見過的歌手作為來源音檔的歌手，而目標歌手則保持為訓練過程中見過的歌手（以下簡稱為 unseen-to-seen），即任意對多（any-to-many）的情境。在實際操作中，我們對每個來源歌手和目標歌手的組合都隨機選擇了 2 段音檔，因此在 seen-to-seen 的情境下，由於訓練時共有 22 位歌手（Opencpop [33] 1 位歌手，TONAS [24] 1 位歌手，M4Singer [37] 20 位歌手），且我們排除了來源歌手和目標歌手相同的組合，因此總共有 $22 \times 21 \times 2 = 924$ 個音檔；而在 unseen-to-seen 的情境下，由於 OpenSinger [10] 中有 76 位歌手，因此總共有 $76 \times 22 \times 2 = 3344$ 個音檔。

doi:10.6342/NTU202402100

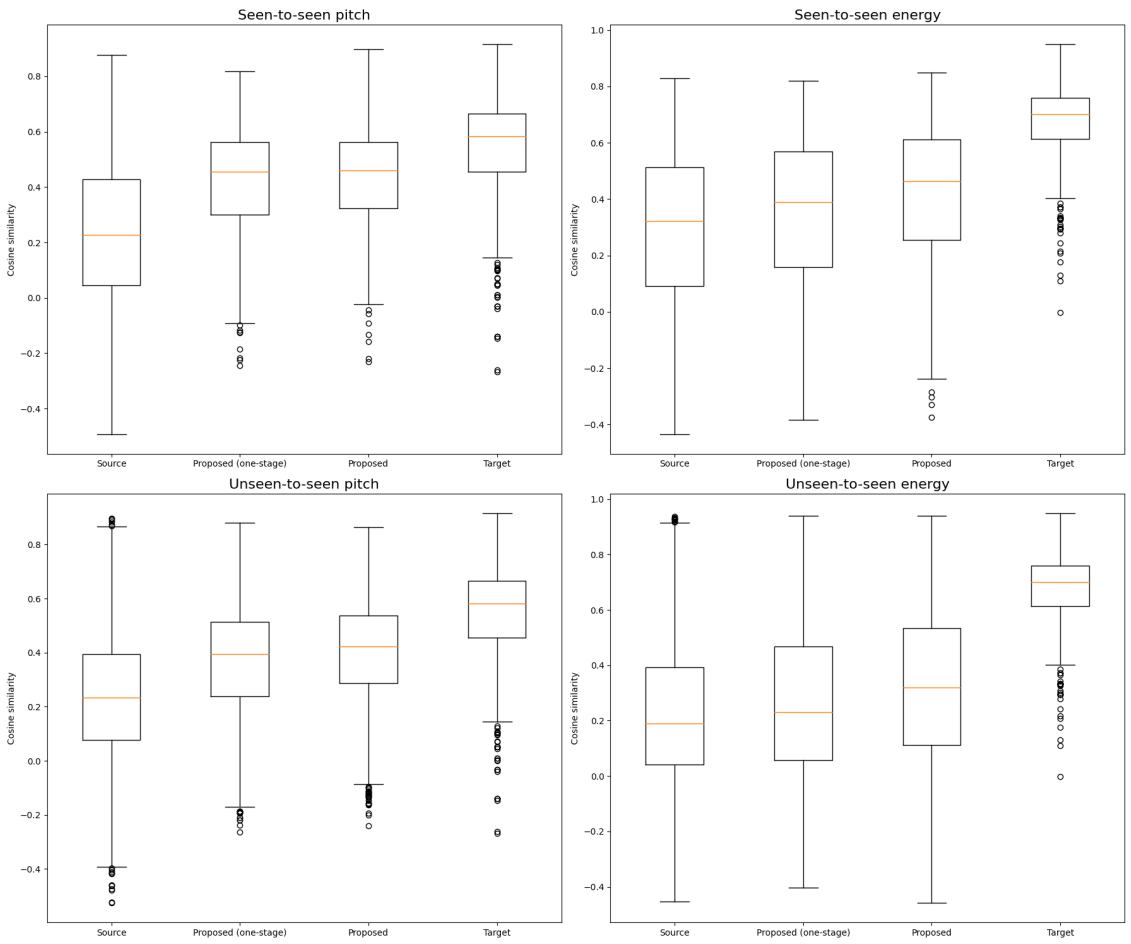


圖 5.1: 單階段轉換與二階段轉換模型之客觀指標箱形圖

圖 5.1 顯示了將各音檔轉換後的音高曲線和能量曲線對應的歌手嵌入（singer embedding），經過與目標歌手的平均歌手嵌入計算餘弦相似度（cosine similarity）後的箱形圖，而平均餘弦相似度則列於表 5.1。在所有情況下，與 Source 相比，Proposed (one-stage) 和 Proposed 皆提升了歌手嵌入的相似度，顯示提出的模型能在多對多和任意對多的情境下，將音高曲線和能量曲線轉換成更接近目標歌手的歌唱風格。然而，與 Target 相比，Proposed (one-stage) 和 Proposed 的相似度仍有明顯差異，顯示提出的模型在歌唱風格轉換方面仍有改進的空間，尤其在能量曲線轉換的部分。我們認為使用更先進的模型架構或更大量的資料訓練可能會帶來進一步的改善，這些部分可作為未來改進的方向。而在 Proposed (one-stage) 和 Proposed 的比較中，Proposed 在兩種情境下的相似度明顯高於 Proposed (one-stage)，顯示使用二階段轉換模型能夠實現更好的歌唱風格轉換。我們認為這是因為音高曲線和能量曲線具有不同的特性，例如音高曲線通常按照樂譜的音符趨勢，因此整體上呈現較明顯的階梯狀，而能量曲線則沒有此限制。因此，同時轉

doi:10.6342/NTU202402100

表 5.1: 單階段轉換與二階段轉換模型之客觀指標比較表

Model	Seen-to-seen		Unseen-to-seen	
	Pitch	Energy	Pitch	Energy
Source	0.226	0.295	0.230	0.224
Proposed (one-stage)	0.424	0.354	0.367	0.267
Proposed	0.435	0.414	0.403	0.321
Target	**0.555**	**0.681**	**0.555**	**0.681**

換兩者可能導致互相干擾，進而降低轉換後的相似度。

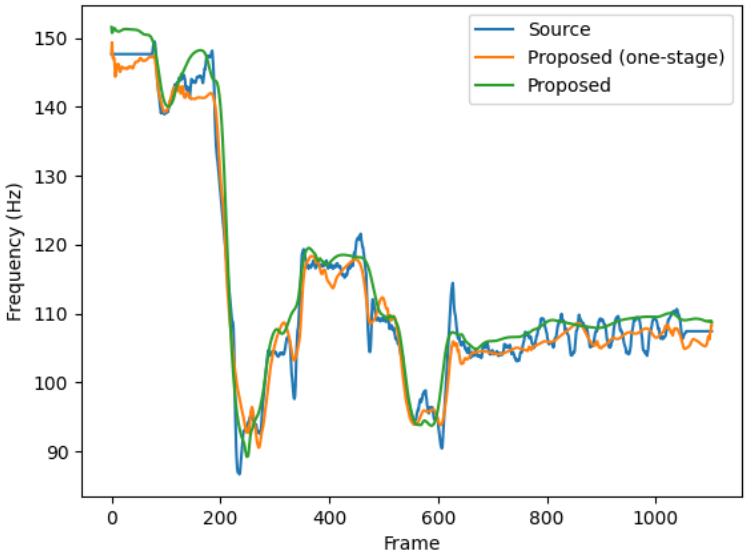


圖 5.2: 單階段轉換與二階段轉換模型之音高曲線結果範例圖

在聽感評量方面，我們首先隨機挑選了幾組轉換結果，並使用 Diff-SVC 生成了相應的歌聲，然後以人工方式進行聽測。我們發現 Proposed (one-stage) 轉換結果中的歌聲容易出現走音的問題，尤其是在長音部分。圖 5.2 展示了一個實際的例子，轉換的目標歌手是 Opencpop [33]，這位歌手的歌唱風格為不常使用顫音（vibrato）的技巧。從圖中可以看出，在第 800 到 1000 個音框（frame）之間，原始音檔存在一個顫音，而 Proposed (one-stage) 和 Proposed 都學到了這種歌唱風格的差異，將顫音轉換為長音。然而，Proposed (one-stage) 的長音存在不自然的抖動，因此在聽感上會有走音的問題。綜上所述，我們認為使用二階段轉換模型可以獲得更好的表現，因此後續實驗中的模型都將採用二階段轉換的方法。

doi:10.6342/NTU202402100

5.2 實驗二：顫音建模與顫音幅度平滑之消融實驗

在 3.1.3 小節中，我們參考了 [17] 的方法，透過對音高曲線進行短時距傅立葉變換（short-time Fourier transform, STFT），以凸顯顫音的特徵。我們進一步修改了原方法中的一階差分（first-order difference），改為使用高通濾波器（high-pass filter）來強調曲線中相對高頻的成分，確保顫音幅度曲線（vibrato extent contour）僅反映顫音振幅。而在 3.1.4 小節中，我們對輸出的顫音幅度曲線進行了平滑處理，以避免音高曲線中顫音振幅的不規則波動。在這個實驗中，我們將進行四個消融實驗（ablation study），分別是未加入顫音幅度曲線的重建損失（以下簡稱 Proposed (w/o VibExt)）、將高通濾波器改回原方法中的一階差分（以下簡稱 Proposed (HPF $\rightarrow$ diff)）、未加入功率時頻譜（power spectrogram）的重建損失（以下簡稱 Proposed (w/o FT)），以及不對顫音幅度曲線進行平滑處理（以下簡稱 Proposed (w/o smooth)），並比較這些模型在客觀指標和聽感上的差異。同時，和 5.1 小節一樣，我們以 Source 作為下界，以 Target 作為上界，並且對 seen-to-seen 和 unseen-to-seen 這兩種情境進行評量，相關細節請參見 5.1 小節。

表 5.2: 顫音建模與顫音幅度平滑消融實驗之客觀指標比較表

Model	Seen-to-seen		Unseen-to-seen	
	Pitch	Energy	Pitch	Energy
Source	0.226	0.295	0.230	0.224
Proposed (w/o VibExt)	0.309	0.427	0.293	0.340
Proposed (HPF $\rightarrow$ diff)	0.365	0.417	0.362	0.353
Proposed (w/o FT)	0.407	0.452	0.376	0.370
Proposed (w/o smooth)	0.421	0.435	0.396	0.331
Proposed	0.435	0.414	0.403	0.321
Target	**0.555**	**0.681**	**0.555**	**0.681**

圖 5.3 展示了將各音檔轉換後的音高曲線和能量曲線對應的歌手嵌入，經過與目標歌手的平均歌手嵌入計算餘弦相似度後的箱形圖，而相應的平均餘弦相似度則列於表 5.2。相較於 Source，這四個消融實驗的模型在兩種情境下的歌手嵌入相似度均有所提升，顯示它們在歌唱風格轉換上皆有不錯的表現。而對於音高曲線而言，Proposed 在兩種情境下的相似度均優於這四個消融實驗的模型，顯示了我們提出的顫音建模改進方法以及對顫音幅度曲線進行平滑處理，確實能達到更好的歌唱風格轉換效果。至於能量曲線的部分，在兩種情境下則是 Proposed (w/o FT) 有較佳的表現。我們認為這也與音高曲線和能量曲線具有不同特性有關，所以在音高曲線轉換表現最好的方法，未必在能量曲線轉換方面也有最佳的表現。

doi:10.6342/NTU202402100

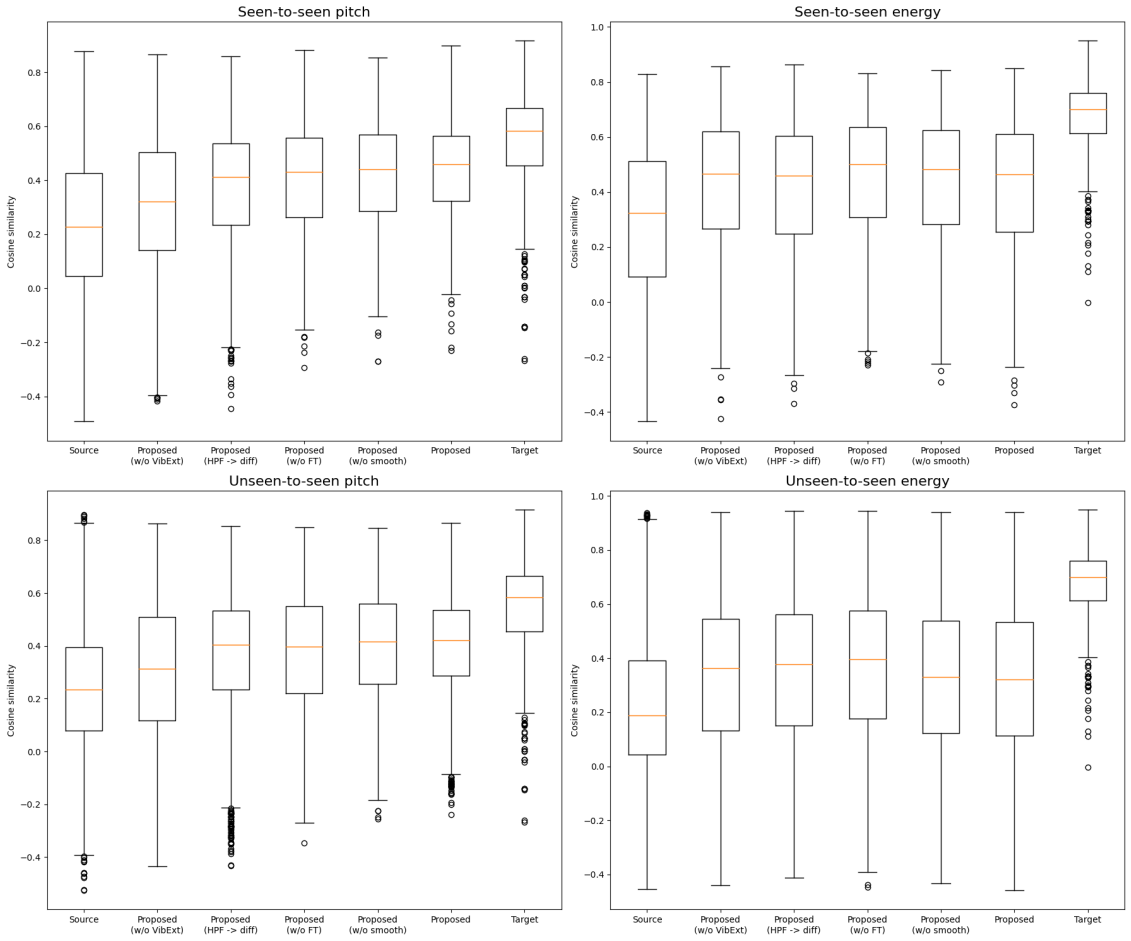


圖 5.3: 顫音建模與顫音幅度平滑消融實驗之客觀指標箱形圖

因此，針對能量曲線的特性提出更客製化的方法，將會是未來可以再改進的方向。

在聽感評量方面，我們同樣隨機挑選了多組轉換結果，並利用 Diff-SVC 生成相應的歌聲，接著進行人工聽測。我們觀察到這四個消融實驗的模型轉換出的歌聲，在顫音的部分顯得較不自然，容易出現顫音頻率或顫音幅度不穩定的問題。因此，在後續的實驗中，我們將持續使用 Proposed 作為主要架構。

5.3 實驗三：有無提供音高曲線對能量轉換模型之影響

在 3.2.2 小節中，我們將音高曲線作為能量轉換（energy conversion）模型的額外輸入，旨在使輸出的能量曲線與其對應的音高曲線相互配合。在這個實驗中，我們將比較未提供音高曲線對能量轉換模型在客觀指標和聽感上的影響（以下簡稱為 Proposed (w/o pitch)）。此外，我們一樣以 Source 來作為下界，以及以 Target

doi:10.6342/NTU202402100

來作為上界，並且評量了 seen-to-seen 和 unseen-to-seen 這兩種情境，相關細節請參見 5.1 小節。

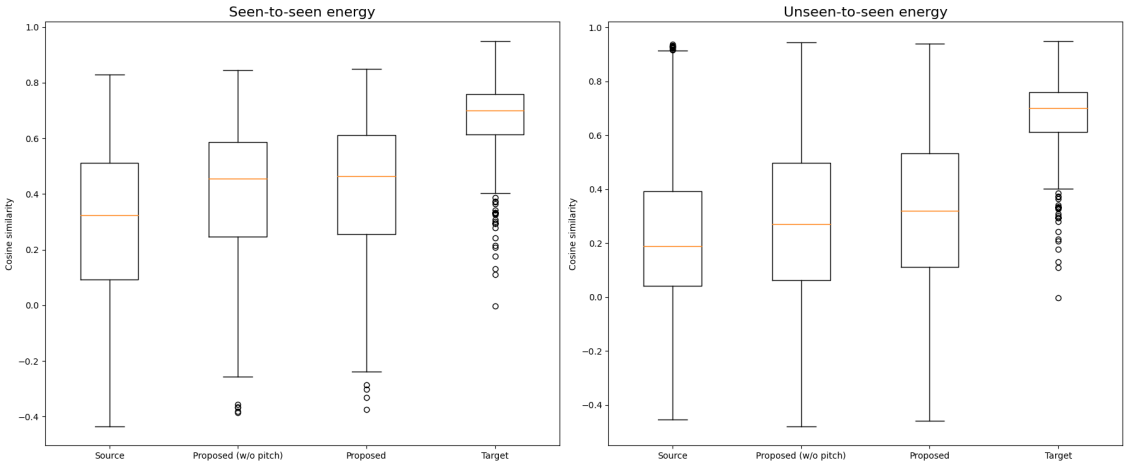


圖 5.4: 有無提供音高曲線對能量轉換模型影響之客觀指標箱形圖

表 5.3: 有無提供音高曲線對能量轉換模型影響之客觀指標比較表

Model	Seen-to-seen energy	Unseen-to-seen energy
Source	0.295	0.224
Proposed (w/o pitch)	0.402	0.287
Proposed	0.414	0.321
Target	**0.681**	**0.681**

圖 5.4 顯示了將各音檔轉換後的能量曲線對應的歌手嵌入，經過與目標歌手的平均歌手嵌入計算餘弦相似度後的箱形圖，而相應的平均餘弦相似度則在表 5.3 中列出。由於此實驗僅修改了能量轉換模型，音高轉換（pitch conversion）模型的結果保持不變，因此這裡僅呈現了能量曲線的結果。與 Source 相比，Proposed (w/o pitch) 在兩種情境下的歌手嵌入相似度均有所提升，顯示 Proposed (w/o pitch) 在歌唱風格轉換上也表現出不錯的效果。然而，就 Proposed (w/o pitch) 和 Proposed 的比較而言，Proposed 在兩種情境下的相似度皆稍微優於 Proposed (w/o pitch)，進一步顯示了將音高曲線輸入能量轉換模型的有效性。

在聽感評量方面，我們同樣隨機挑選了多組轉換結果，利用 Diff-SVC 生成相應的歌聲，隨後進行人工聽測。我們觀察到 Proposed (w/o pitch) 轉換出的歌聲在聽感上顯得較不自然，容易出現音量不穩定的問題，特別是在顫音的部分。圖 5.2 展示了一個實際的例子，左右兩側分別是 Proposed (w/o pitch) 和 Proposed 轉換後的音高曲線和能量曲線。從圖中可以看出，在第 400 到 600 個音框之間，轉換後的結果存在一個顫音，而 Proposed 轉換後的能量曲線明顯更加貼合音高曲線，

doi:10.6342/NTU202402100

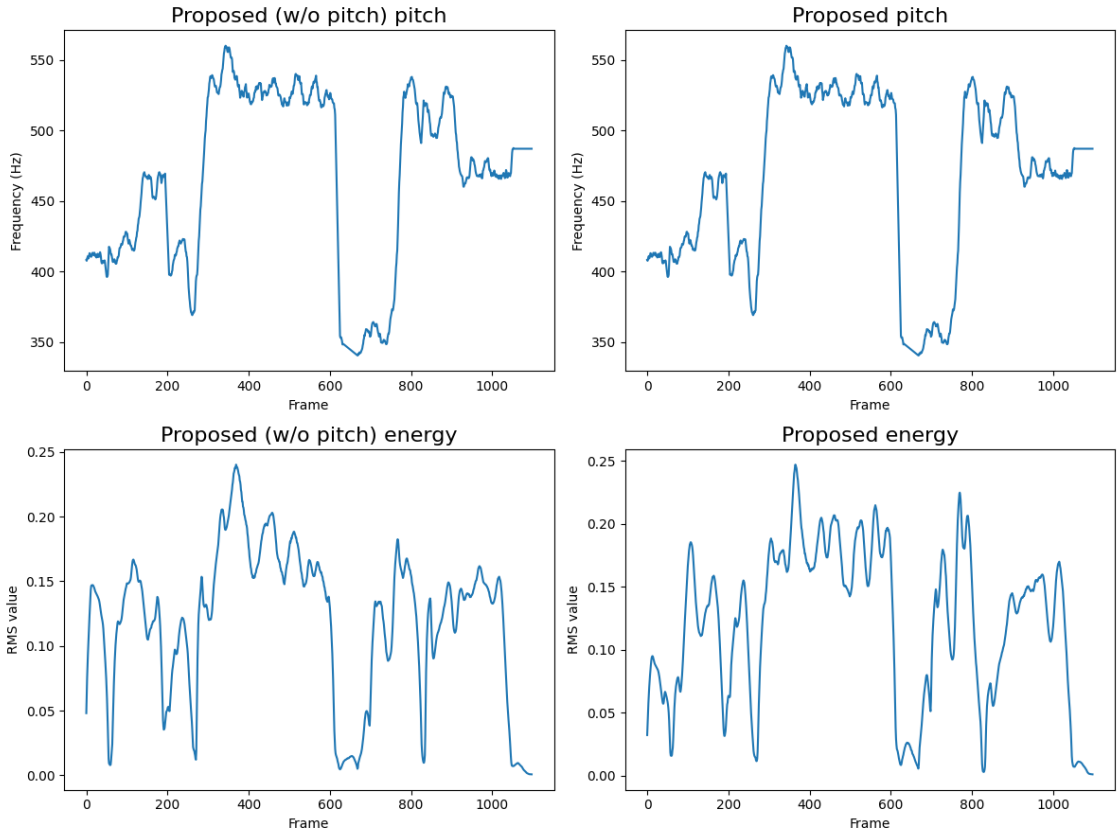


圖 5.5: 有無提供音高曲線對能量轉換模型之影響範例圖

在這段範圍內以相同的頻率和相位一同抖動，但是 Proposed (w/o pitch) 轉換後的能量曲線則沒有這種現象。這種差異也影響了聽感的表現，我們認為在這個例子中，與 Proposed 相比，Proposed (w/o pitch) 的顫音聽起來相對較不明顯也不太自然，這可能是由於音高曲線和能量曲線之間缺乏良好的配合所導致的。

雖然從客觀指標來看，Proposed (w/o pitch) 和 Proposed 之間的表現沒有太大的差異，但我們認為在聽感上兩者之間的區別相當明顯。因此，後續的主觀評量將對這兩者進行更深入的比較。

5.4 實驗四：主觀評量指標之結果分析

在這個實驗中，我們將提出的模型實際應用於歌聲轉換（singing voice conversion）模型，並比較其在主觀指標上的表現。為了減輕受測者的負擔，我們僅評量了 seen-to-seen 的情境，並進行了兩組比較平均意見分數（comparison mean opinion score, CMOS）[20] 評量：第一組比較了僅轉換音色（以下簡稱為 Source

doi:10.6342/NTU202402100

(SVC)）和 Proposed (w/o pitch)，第二組則是比較了 Source (SVC) 和 Proposed。我們共邀請了 10 位受測者，每位受測者將被要求對 12 組音檔片段配對進行評分。

5.4.1 整體主觀評量結果

表 5.4 展示了相似度和自然度的整體主觀評量結果，其長條圖呈現於圖 5.6 中。在相似度方面，有趣的是，雖然在客觀評量中，Proposed (w/o pitch) 的表現優於 Source (SVC)，但在主觀評量中，Proposed (w/o pitch) 的相似度卻顯著劣於 Source (SVC)（p 值約為 0.040）。這說明了 Proposed (w/o pitch) 中音高曲線和能量曲線的不一致可能會嚴重影響人們對歌唱風格的感知。因此，受測者可能會認為 Proposed (w/o pitch) 的歌唱風格與目標歌手不相似。而 Proposed 則顯著優於 Source (SVC)（p 值約為 0.017），再次證明了提出的模型確實能夠改善歌手的歌唱風格相似度，而這是人們可以識別的。

表 5.4: 多對多情境下整體主觀指標比較表

Model	Seen-to-seen	
	Similarity	Naturalness
Source (SVC)	0.000	**0.000**
Proposed (w/o pitch)	-0.250 ± 0.279	-1.308 ± 0.280
Proposed	**+0.342 ± 0.315**	-1.233 ± 0.290

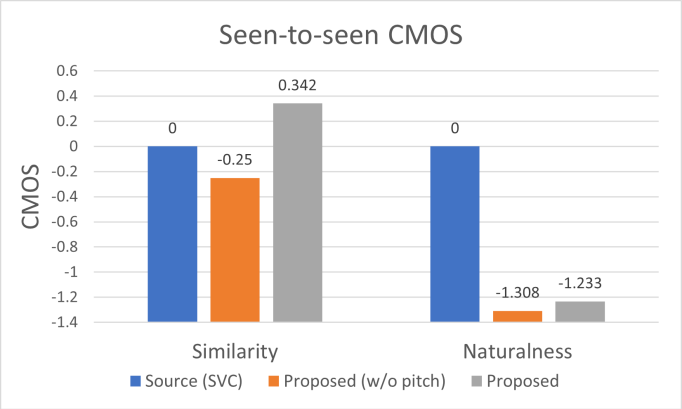


圖 5.6: 多對多情境下整體主觀指標長條圖

在自然度方面，Source (SVC) 顯著優於 Proposed (w/o pitch) 和 Proposed（p 值分別約為 6×10^{-16} 和 5×10^{-14}）。這反映了轉換音高曲線和能量曲線的副作用，進而導致自然度下降。而這也指出了未來可以再改進的另一個方向：專注於提升轉換後音高曲線和能量曲線的自然度，同時保留提出的模型所產生的歌唱風格轉換效果。

doi:10.6342/NTU202402100

5.4.2 不同性別目標歌手之主觀評量結果

為了進一步了解 Proposed 相對於 Proposed (w/o pitch) 的改進之處，我們根據目標歌手的性別將主觀評量結果分為兩個部分，該分析結果如表 5.5 所示，其對應的長條圖請參見圖 5.7。我們發現，在不同性別的目標歌手下，Proposed 相對於 Proposed (w/o pitch) 在相似度和自然度方面都有所提升，尤其是當目標歌手為女性時，相似度的提升最為明顯。我們認為這可能與人耳對中高音的敏感度較高有關。當目標歌手為女性時，音高曲線和能量曲線的不一致更容易被人們所察覺，因此受測者可能會給出較低的相似度評分；而當音高曲線和能量曲線能夠相互配合時，女聲的歌唱風格由於音高較高，相較於男聲更容易被辨識出來，因此目標歌手為女性的相似度評分也相對較高。

表 5.5: 多對多情境下不同性別目標歌手之主觀指標比較表

Target singer	Model	Seen-to-seen	
		Similarity	Naturalness
Male	Source (SVC)	0.000	**0.000**
	Proposed (w/o pitch)	-0.133 ± 0.364	-1.267 ± 0.392
	Proposed	+0.133 ± 0.424	-1.167 ± 0.356
Female	Source (SVC)	0.000	**0.000**
	Proposed (w/o pitch)	-0.367 ± 0.433	-1.350 ± 0.411
	Proposed	**+0.550 ± 0.472**	-1.300 ± 0.467

5.5 實驗五：任意對多情境下之案例分析

在這個實驗中，我們針對 unseen-to-seen 的情境進行了案例分析，旨在了解提出的歌唱風格轉換模型對音高曲線、能量曲線和聽感的實際影響，特別是對於未見過的來源歌手。

5.5.1 成功轉換案例

圖 5.8 展示了一個轉換成功的實際案例，呈現了來源音檔和經由 Proposed 轉換後的音高曲線和能量曲線，轉換的目標歌手為 M4Singer [37] 中的 Tenor-7，這位歌手的歌唱風格為較少使用顫音的技巧。我們發現，即使來源歌手是在訓練過程中未見過的，提出的模型仍然能成功捕捉到他們之間歌唱風格的差異，並進一步修改了音高曲線和能量曲線以去除顫音，使得轉換後的歌唱風格更加符合

doi:10.6342/NTU202402100

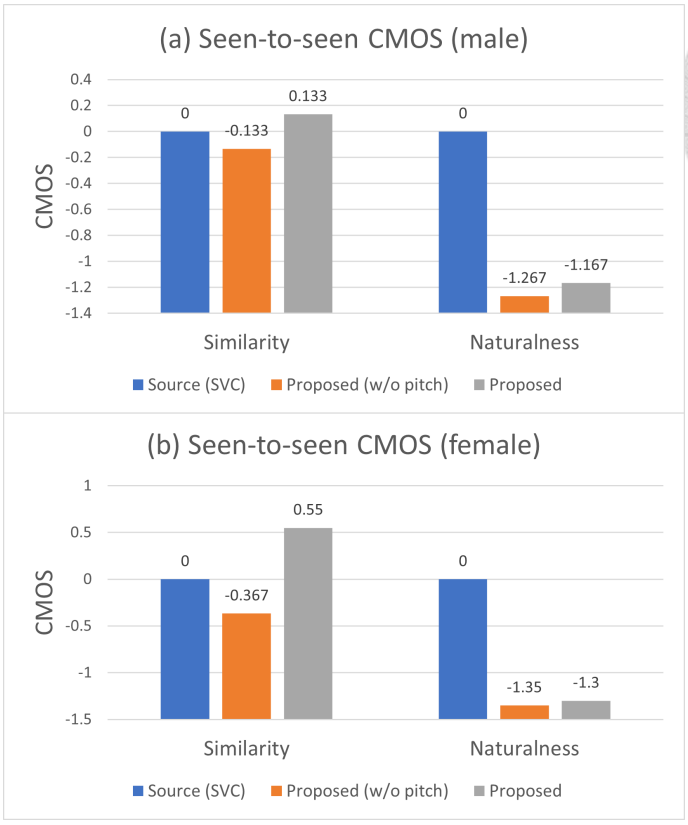


圖 5.7: 多對多情境下不同性別目標歌手之主觀指標長條圖

Tenor-7。這也反映在客觀指標上，與 Tenor-7 平均歌手嵌入的音高曲線相似度從 -0.245 提升到 0.124。

除此之外，在第 1800 個音框附近，來源音檔的音高曲線存在一些轉音，而轉換後的曲線也對這些特徵進行了相應的修改。這顯示了我們提出的模型在顫音之外的風格差異方面，仍然具有一定的轉換能力。

5.5.2 失敗轉換案例

圖 5.9 展示的則是一個轉換失敗的實際案例，呈現了來源音檔和經由 Proposed 轉換後的音高曲線和能量曲線，轉換的目標歌手為 M4Singer [37] 中的 Alto-1，這位歌手的歌唱風格具有較為明顯的顫音。在這個例子中，提出的模型未能成功捕捉到他們之間歌唱風格的差異，導致沒能將來源音檔中的顫音幅度放大。客觀指標數據顯示，與 Alto-1 平均歌手嵌入的音高曲線相似度從 0.550 下降到 0.448。

除此之外，在聽感評量方面，我們發現轉換後的歌聲存在些微走音的問題。例如，在第 1000 個音框附近，轉換後的音高曲線與原始曲線存在較大的誤差，這

doi:10.6342/NTU202402100

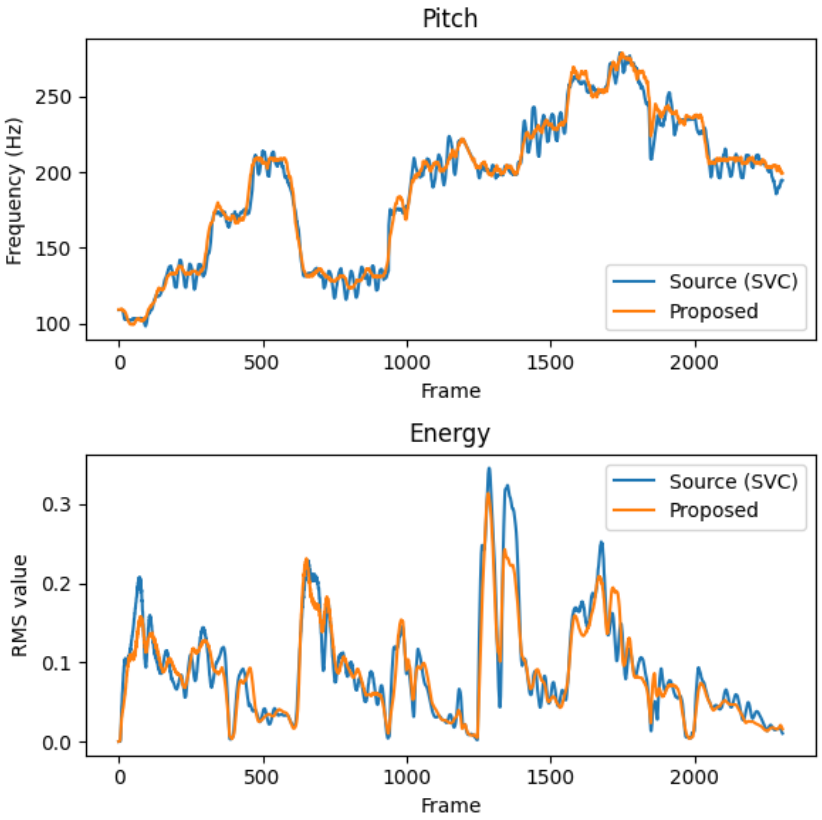


圖 5.8: 任意對多情境下之轉換成功結果範例圖

會使聆聽者感受到音準有偏差。我們認為這部分的問題源於模型對細節的處理不夠精確，導致在曲線變化較快的區域，重建時容易有較大的誤差。而這也可能是為什麼在 5.4 小節中，自然度的結果會表現不佳的原因，因為這種細節的缺失，會使受測者感覺到轉換後的音檔聽起來相對不自然。因此，調整模型架構以更好地學習細節特徵，將是未來改進的方向之一。

5.6 實驗六：歌唱風格轉換與顫音轉換之比較

在這個實驗中，我們將比較提出的歌唱風格轉換模型和僅轉換顫音的方法（以下簡稱為 Vib-scaling）在客觀指標和聽感上的差異，目的是證明我們提出的模型能夠轉換顫音以外的歌唱風格。我們一樣以 Source 作為下界，以 Target 作為上界，並評量了 seen-to-seen 的情境，相關細節請參見 5.1 小節。

圖 5.10 是顫音轉換的方法說明圖。其中，圖 5.10（a）展示了一段範例音高曲線 $\mathbf{p}$，其對應的功率時頻譜 $\mathbf{p}_{\mathrm{FT}}$ 和顫音幅度曲線 $\mathbf{p}_{\mathrm{VibExt}}$ 分別如圖 5.10（b）和圖 5.10（c）所示。首先，我們從訓練資料中統計每位歌手的顫音幅度平均值和標準

doi:10.6342/NTU202402100

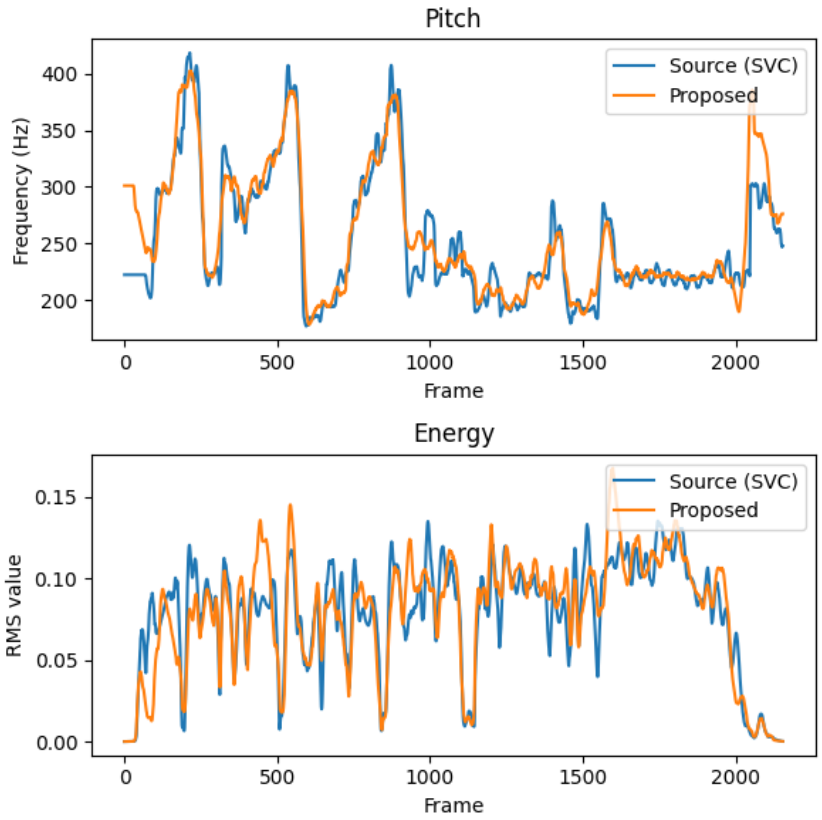


圖 5.9: 任意對多情境下之轉換失敗結果範例圖

差；接著，找出顫音幅度曲線大於顫音閾值的部分，我們將閾值設置為 0.25；之後，將功率時頻譜中對應位置的數值經由來源歌手和目標歌手的統計量進行標準化（standardization）和反標準化（inverse standardization）運算，如圖 5.10（d）所示；最後，再將其轉回音高曲線即為顫音轉換的結果，如圖 5.10（e）所示。

表 5.6: 歌唱風格轉換與顫音轉換之客觀指標比較表

Model	Seen-to-seen pitch	Seen-to-seen energy
Source	0.226	0.295
Vib-scaling	0.283	0.292
Proposed	0.435	0.414
Target	**0.555**	**0.681**

圖 5.11 展示了將各音檔轉換後的音高曲線和能量曲線對應的歌手嵌入，經過與目標歌手的平均歌手嵌入計算餘弦相似度後的箱形圖，而相應的平均餘弦相似度則列於表 5.6。由於此實驗需要歌手資料來計算顫音相關統計量，因此僅呈現 seen-to-seen 的結果。相較於 Source，Vib-scaling 在音高曲線方面的相似度有所提升，但在能量曲線方面並未顯示出更佳的表現，說明僅轉換顫音可能對能量曲線的轉換效果比較有限。而就 Vib-scaling 和 Proposed 的比較而言，Proposed 明顯優

doi:10.6342/NTU202402100

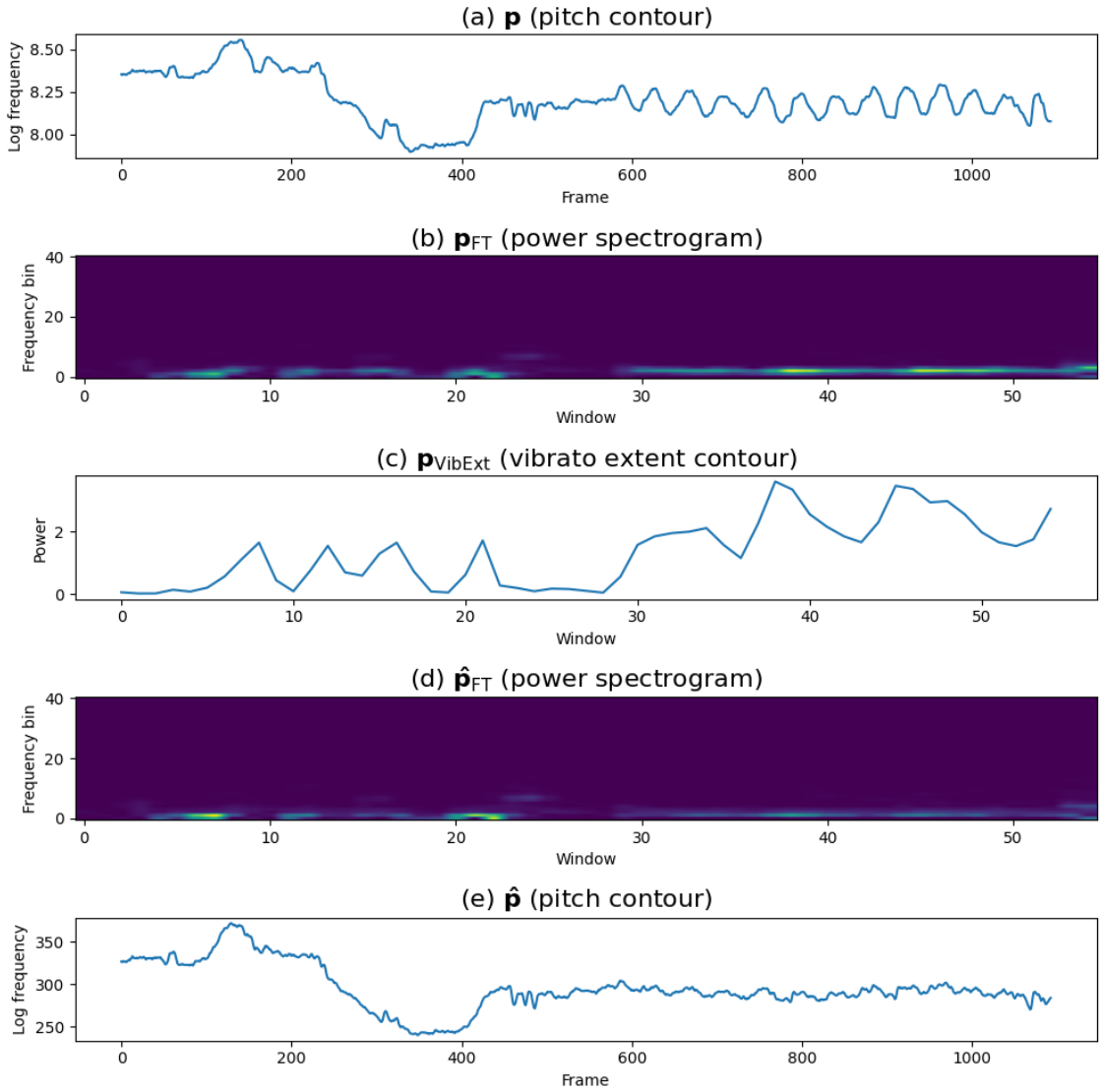


圖 5.10: 顫音轉換方法說明圖

於 Vib-scaling，這表明即使增加了強調顫音的損失函數，我們提出的模型也沒有退化到僅修改顫音的情況，反而是通過考慮多種歌唱風格，取得了更好的表現。

在聽感評量方面，我們同樣隨機挑選了多組轉換結果，並利用 Diff-SVC 生成相應的歌聲，之後進行人工聽測。我們觀察到 Vib-scaling 轉換出的歌聲在自然度上優於 Proposed，但在相似度上則是 Proposed 更勝一籌。例如，當目標歌手的歌唱風格為不常使用顫音技巧時，Vib-scaling 轉換後的結果仍會有些許顫音，而 Proposed 則可以完全轉換掉顫音。

doi:10.6342/NTU202402100

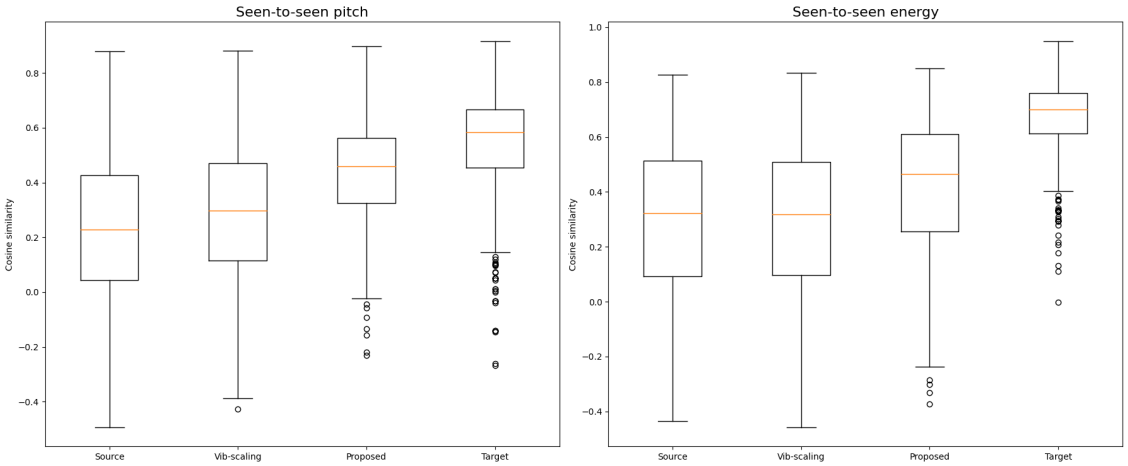


圖 5.11: 歌唱風格轉換與顫音轉換之客觀指標箱形圖

doi:10.6342/NTU202402100

第六章　結論與未來展望

本章節總結本論文的主要發現和貢獻，概述研究成果的實際意涵和學術價值。同時，我們將討論研究中遇到的挑戰和限制，並提出未來研究的方向和改進建議，以促進相關領域的進一步發展。

6.1　結論

本論文提出了首個任意對多（any-to-many）歌唱風格轉換模型，主要專注於將音高曲線（pitch contour）和能量曲線（energy contour）轉換成更接近目標歌手的歌唱風格。本論文所得出的重要結論如下：

1. 受第二章所介紹的各種研究方法啟發，我們提出了一個類似的新任務，旨在轉換歌聲中的歌唱風格。我們將歌唱風格轉換的任務分解為音高轉換（pitch conversion）和能量轉換（energy conversion），並採用了類似於 AutoVC [28] 的模型架構來處理這兩個子任務。

2. 根據 5.1 小節的實驗一結果顯示，在多對多（many-to-many）和任意對多的情境下，我們提出的模型在音高曲線和能量曲線等方面皆能提升與目標歌手的歌唱風格相似度。其中，使用了二階段轉換的模型呈現出比單階段轉換模型更優異的表現。因此，我們採用此模型進行後續實驗。

3. 根據 5.2 小節的實驗二結果顯示，在多對多和任意對多的情境下，我們提出的顫音建模改良方法以及對顫音幅度曲線（vibrato extent contour）進行平滑處理，能夠在音高曲線方面提升與目標歌手的歌唱風格相似度，但在能量曲線方面未必能夠有效提升相似度。而在聽感評量方面，我們的改良方法仍然能使轉換後的歌聲聽起來比較自然。

doi:10.6342/NTU202402100

4. 根據 5.3 小節的實驗三結果顯示，在多對多和任意對多的情境下，將音高曲線作為能量轉換模型的輔助資訊，能夠有效提升與目標歌手的歌唱風格相似度。同時，這也促使轉換後的音高曲線和能量曲線更好地配合，進而使得轉換後的歌聲更加自然，且歌唱風格更為明顯。

5. 根據 5.4 小節的實驗四結果顯示，在多對多的情境下，我們提出的模型在相似度方面獲得了較高的比較平均意見分數（comparison mean opinion score, CMOS），顯示轉換後的歌唱風格在聽感上確實更接近目標歌手。然而，在自然度方面卻下降了，顯示轉換所帶來的副作用。此外，音高曲線和能量曲線的不一致會對聽感產生嚴重的影響，足以使相似度的 CMOS 分數比未轉換的情況更低，尤其是在目標歌手為女性時，影響最為明顯。

6. 根據 5.5 小節的實驗五結果顯示，在任意對多的情境下，我們提出的模型仍然能夠捕捉到來源歌手和目標歌手的歌唱風格差異，將音高曲線和能量曲線轉換以符合目標歌手的歌唱風格。然而，模型在細節方面仍有許多可以改進的空間，例如在曲線變化較快區域的轉換，以及走音的問題等。

7. 根據 5.6 小節的實驗六結果顯示，在多對多的情境下，我們提出的模型在相似度表現上優於僅轉換顫音的方法，這表明該模型不僅能夠有效地轉換顫音，還能處理和轉換其他不同的歌唱風格，進一步提升整體的歌手身份相似度。

6.2 未來展望

儘管本論文在歌唱風格轉換任務上已取得一定的研究成果，然而，由於時間和研究資源的限制，仍有許多值得進一步探討和延伸的研究方向，以下列舉一些例子：

1. 在本論文中，音高嵌入（pitch embedding）和能量嵌入（energy embedding）的表示方法雖然可以輕易轉換回音高曲線和能量曲線，但僅有其中兩個維度不為零，這樣所包含的資訊可能相對有限。未來可以探索其他表示方法，例如獨熱編碼（one-hot encoding），或是讓模型自動學習適當的表示方法，以提升嵌入表示的豐富性和精確性。

doi:10.6342/NTU202402100

2. 本論文提出的模型採用了類似 AutoVC [28] 的自編碼器（autoencoder）架構，主要透過信息瓶頸（information bottleneck）的設計實現風格轉換效果。然而，在這類生成任務中，仍有多種模型架構值得嘗試，例如生成對抗網路（generative adversarial network, GAN）[6] 和擴散模型（diffusion model）[9]。期待通過改進模型架構來更好地學習歌唱風格的細部特徵。

3. 在本論文中，訓練能量轉換模型的方法基本上與訓練音高轉換模型的方法相同，並未專門針對能量曲線的特性進行調整。因此，能量曲線的相似度表現未必能夠像音高曲線一樣好。未來的研究可以嘗試根據兩種曲線的特性進行改進。例如，音高曲線需按照樂譜中的音符趨勢，而能量曲線則不受此限制。

4. 本論文所使用的訓練資料僅包含了 22 位歌手，然而，有些歌手的歌唱風格可能並不明顯。因此，若能使用更龐大的資料集進行訓練，尤其是包含了各種不同歌唱風格的歌手資料，或許能夠使模型更有效地學習到歌手間歌唱風格的差異，進一步提升整體表現。透過增加訓練資料的多樣性，模型將有更多的機會學習到歌唱風格的各種特徵，進而提高其在歌唱風格轉換任務中的泛化能力。

5. 本論文所使用的客觀指標和主觀指標，大多參考了歌聲轉換（singing voice conversion）的評量方法，未來也可嘗試加入針對音高曲線或能量曲線中歌唱風格特徵的評量指標，例如顫音相關的統計量等。在主觀評量方面，有些受測者反映音檔的長度會影響其對歌唱風格的判斷，還有部分音檔的歌唱風格並不容易辨識。因此，如何調整問卷，使受測者能更有效地評量歌唱風格，也是未來可再研究的方向。

6. 本論文所探討的轉換情境為任意對多，然而在客觀評量過程中，我們參考了語者驗證（speaker verification）的方法，訓練了音高編碼器和能量編碼器。未來或許可以嘗試將這兩個編碼器改良後分別取代音高轉換模型和能量轉換模型中的歌手嵌入（singer embedding）查詢表，從而進一步拓展至任意對任意（any-to-any）的轉換情境。

doi:10.6342/NTU202402100

doi:10.6342/NTU202402100

參考文獻

[1] W. Cai, J. Chen, and M. Li. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. arXiv preprint arXiv:1804.05160, 2018.

[2] H.-Y. Choi, S.-H. Lee, and S.-W. Lee. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. International Speech Communication Association, pages 2283–2287, 2023.

[3] J. S. Chung, J. Huh, and S. Mun. Delving into voxceleb: environment invariant speaker recognition. arXiv preprint arXiv:1910.11238, 2019.

[4] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han. In defence of metric learning for speaker recognition. arXiv preprint arXiv:2003.11982, 2020.

[5] C. Deng, C. Yu, H. Lu, C. Weng, and D. Yu. Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7749–7753. IEEE, 2020.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

[7] H. Guo, Z. Zhou, F. Meng, and K. Liu. Improving adversarial waveform generation based singing voice conversion with harmonic signals. In ICASSP 2022-2022 IEEE

doi:10.6342/NTU202402100

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6657–6661. IEEE, 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[9] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[10] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In Proceedings of the 29th ACM International Conference on Multimedia, pages 3945–3954, 2021.

[11] W. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda. The singing voice conversion challenge 2023. In IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023, pages 1–8. IEEE, 2023.

[12] T. Jayashankar, J. Wu, L. Sari, D. Kant, V. Manohar, and Q. He. Self-supervised representations for singing voice conversion. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[13] J. W. Kim, J. Salamon, P. Li, and J. P. Bello. Crepe: A convolutional representation for pitch estimation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 161–165. IEEE, 2018.

[14] C. Le Moine, N. Obin, and A. Roebel. Towards end-to-end f0 voice conversion based on dual-gan with convolutional wavelet kernels. In 2021 29th European Signal Processing Conference (EUSIPCO), pages 36–40. IEEE, 2021.

[15] J. Lee, H. Choi, J. Koo, and K. Lee. Disentangling timbre and singing style with multi-singer singing synthesis system. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, pages 7224–7228. IEEE, 2020.

doi:10.6342/NTU202402100

[16] X. Li, S. Liu, and Y. Shan. A hierarchical speaker representation framework for one-shot singing voice conversion. arXiv preprint arXiv:2206.13762, 2022.

[17] R. Liu, X. Wen, C. Lu, L. Song, and J. S. Sung. Vibrato learning in multi-singer singing voice synthesis. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 773–779. IEEE, 2021.

[18] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng. Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation. In 2021 ieee international conference on multimedia and expo (icme), pages 1–6. IEEE, 2021.

[19] S. Liu, Y. Cao, D. Su, and H. Meng. Diffsvc: A diffusion probabilistic model for singing voice conversion. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 741–748. IEEE, 2021.

[20] P. C. Loizou. Speech quality assessment. In Multimedia analysis, processing and communications, pages 623–654. Springer, 2011.

[21] J. Lu, K. Zhou, B. Sisman, and H. Li. Vaw-gan for singing voice conversion with non-parallel training data. In 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 514–519. IEEE, 2020.

[22] Y. Lu, Z. Ye, W. Xue, X. Tan, Q. Liu, and Y. Guo. Comosvc: Consistency model-based singing voice conversion. arXiv preprint arXiv:2401.01792, 2024.

[23] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans. Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3277–3281. IEEE, 2020.

[24] J. Mora, F. Gómez, E. Gómez, F. Escobar, and J. M. Díaz-Báñez. Melodic characterization and similarity in a cappella flamenco cantes. In International Society for Music Information Retrieval Conference (ISMIR), 2010.

[25] S. Nercessian. Zero-shot singing voice conversion. In ISMIR, pages 70–76, 2020.

doi:10.6342/NTU202402100

[26] Z. Ning, Y. Jiang, Z. Wang, B. Zhang, and L. Xie. Vits-based singing voice conversion leveraging whisper and multi-scale f0 modeling. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 1–8. IEEE, 2023.

[27] B. D. O'Connor, S. Dixon, and G. Fazekas. Zero-shot singing technique conversion. CoRR, abs/2111.08839, 2021.

[28] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In International Conference on Machine Learning, pages 5210–5219. PMLR, 2019.

[29] T. Saitou, M. Unoki, and M. Akagi. Extraction of f0 dynamic characteristics and development of f0 control model in singing voice. In Proc. ICAD, pages 275–278, 2002.

[30] Y. Song, W. Song, W. Zhang, Z. Zhang, D. Zeng, Z. Liu, and Y. Yu. Singing voice synthesis with vibrato modeling and latent energy representation. In 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), pages 1–6. IEEE, 2022.

[31] C. Wang, Z. Li, B. Tang, X. Yin, Y. Wan, Y. Yu, and Z. Ma. Towards high-fidelity singing voice conversion with acoustic reference and contrastive predictive coding. arXiv preprint arXiv:2110.04754, 2021.

[32] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. IEEE Signal Processing Letters, 25(7):926–930, 2018.

[33] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. arXiv preprint arXiv:2201.07429, 2022.

[34] Y. Wu and K. He. Group normalization. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[35] Z. Wu, T. Kinnunen, E. Chng, and H. Li. Text-independent f0 transformation with non-parallel data for voice conversion. In INTERSPEECH, pages 1732–1735, 2010.

doi:10.6342/NTU202402100

[36] F.-L. Xie, Y. Qian, F. K. Soong, and H. Li. Pitch transformation in neural network based voice conversion. In The 9th International Symposium on Chinese Spoken Language Processing, pages 197–200. IEEE, 2014.

[37] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen, et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. Advances in Neural Information Processing Systems, 35:6914–6926, 2022.

[38] Y. Zhou, M. Chen, Y. Lei, J. Zhu, and W. Zhao. Vits-based singing voice conversion system with dspgan post-processing for svcc2023. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 1–8. IEEE, 2023.

doi:10.6342/NTU202402100