基於時域上基週同步疊加法之歌聲合成系統

Singing Voice Synthesis System Based on Time Domain-Pitch Synchronized Overlap and Add

吳銘冠 Ming-Kuan Wu　　陳嘉平 Chia-Ping Chen
國立中山大學資訊工程系
Department of Computer Science and Engineering
National Sun Yat-Sen University
m003040056@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

摘要

在本研究中，我們提出並實作一個串接式的歌聲合成系統，用來產生具有配樂的合成歌聲。語料庫的錄製是根據注音符號檢字表來錄製，並錄製三種不同的音高。我們將主旋律中的力度、音符編號、起始時間和結束時間來當作合成資訊，並加入了轉音的資訊。在合成單元的處理上，採用時域上基週同步疊加法來對合成單元做時域上的修改。我們提供一個歌曲的選擇介面供使用者來進行歌曲的合成，並加入了一些對於合成歌曲的調整。包括了整體上音符編號的調整、歌詞的修改等等。此外，也做了一些聽測實驗，來進行合成歌曲的品質、清晰度和相似度的評估。品質評估方面，合成歌曲加上配樂有改善的效果。清晰度和相似度評估方面，簡單的歌曲有較好的表現。

關鍵詞： 串接合成方法、歌聲合成、時域上基週同步疊加法

Abstract

In this study, we propose and implement a concatenation-based singing synthesis system to synthesize the singing voice with background music. We record three different pitches to build our corpus for all syllables. The synthesis informations, including velocity, note number, start time and end time are extracted from the main melody. Runs and riffs information was added into consideration afterward. We use TD-PSOLA to modify the synthesis units in time domain. At last, we add back the background music extracted from MIDI to our synthesis song. We implemented a user interface for users to synthesize songs. This interface can be used to adjust the synthesis songs, for example, adjust the overall pitches in the song, modify syllables, etc. Finally, we evaluate the quality, clarity and similarity of the synthesis songs. The results show that the proposed method achieve better results with simple songs than with fast songs.

keywords: concatenation synthesis, singing synthesis, TD-PSOLA

一、緒論

(一)、研究背景、目的

近年來，電腦的普及和效能大大的提升，也有愈來愈多的訊號處理能藉由電腦得到更好的成效。如何利用電腦來完成歌聲合成系統，需要考慮到兩個部分。其一為處理輸入的音樂訊號，其二為根據這些音樂訊號去處理所錄製的合成單元，以便完成歌聲合成。歌聲合成的目的，是能夠像人類一樣演唱出樂譜上的旋律和歌詞。因此，本研究的重點是完成一個歌聲合成系統。

(二)、相關研究

1、聲音合成

基週同步疊加法 (Pitch Synchronous Overlap and Add, PSOLA) [1] 主要是為了解決文字轉語音 (Text to Speech, TTS) 上合成品質的問題，他同時也提供了音高升降的處理方式。其演算法能盡量不改變波形的輪廓來變換語音訊號的週期。 PSOLA 演算法可分為時域上基週同步疊加法 (Time Domain-Pitch Synchronous Overlap and Add, TD-PSOLA) [2] [3] [4] 和頻域上基週同步疊加法 (Frequency Domain-Pitch Synchronous Overlap and Add, FD-PSOLA) [5] 兩種。 FD-PSOLA 是將語音訊號轉到頻域上做處理，處理完之後再轉回時域上。 TD-PSOLA 則是運用窗函數直接在時域上對波形做處理。而 TD-PSOLA 的優點則能改善在頻域所耗費太多時間，以及在時域上接合效果太差等問題。

2、歌聲合成

語料庫為主的合成系統，是將事先錄製好的歌手聲音，分析並儲存在資料庫裡。接著利用串接的方式，來產生合成歌曲 [6] [7]。例如由 YAMAHA 公司所開發的商業化軟體 (VOCALOID) [8]。圖 1 為概略的系統架構圖。由圖中可知，使用者只需要歌詞和音符就可以合成出歌曲。

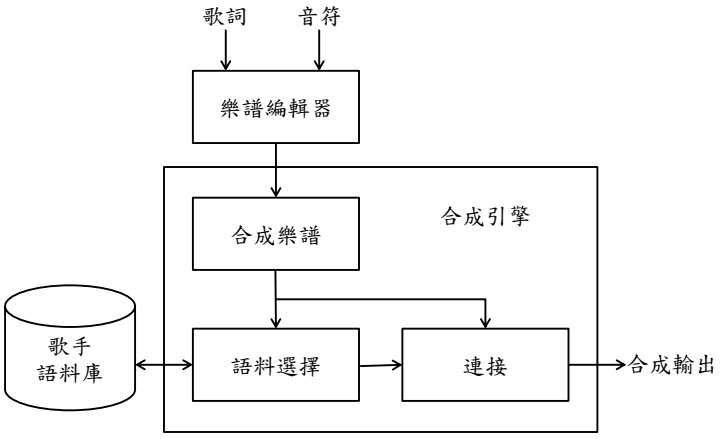


圖 1、 Vocaloid System Diagram

台灣科技大學古鴻炎教授，在語音合成，歌手聲音的合成，以及電腦音樂方面也有相關的研究 [9] [10]。主要採用諧波加噪音模型 (Harmonic plus Noise Model, HNM) 的方式來合成中文歌聲。此模型是由語音訊號的諧波部分 $h(t)$ 和噪音部分 $n(t)$ 所組成。 HNM 會先依訊號的頻譜計算出最大的有聲頻率 (Maximum Voiced Frequency, MVF) $F_m(t)$ 。頻率值小於 MVF 的頻譜部份，在HNM中視為諧波部分，而頻率值大於 MVF 的頻譜部份，在 HNM 中視為噪音部份。諧波部分為語音週期訊號的分量，而噪音部分解釋了非週期分量。這兩個分量在頻域上是分開的，其中諧波部分 $h(t)$ 如式子 1 所示。

$$h(t) = \sum_{k=1}^{K(t)} a_k(t)\cos(\phi_k(t)), \tag{1}$$

其中 $a_k(t)$ 和 $\phi_k(t)$ 表示在時間為 t 時，第 k 個弦波的振幅和相位， $K(t)$ 則表示此時諧波部份所包含的諧波個數。最後，合成的訊號 $s(t)$ 就是由諧波 $h(t)$ 和噪音 $n(t)$ 兩部份的訊號值相加而得到，如式子 2 所示。

$$s(t) = h(t) + n(t). \tag{2}$$

(三)、系統概述及研究方向

本研究是以時域上基週同步疊加法為基礎，來對錄製的合成單元做時域上的修改。接著實作出一個能夠合成出自然歌聲的合成系統。其中，要注意轉音的處理、音量的處理、音節連結上的處理和音節時間的對應，且音色要盡量保持不變。另外，還要能夠使合成出來的歌聲，與配樂同步播放。大致上的流程圖如圖 2 所示。

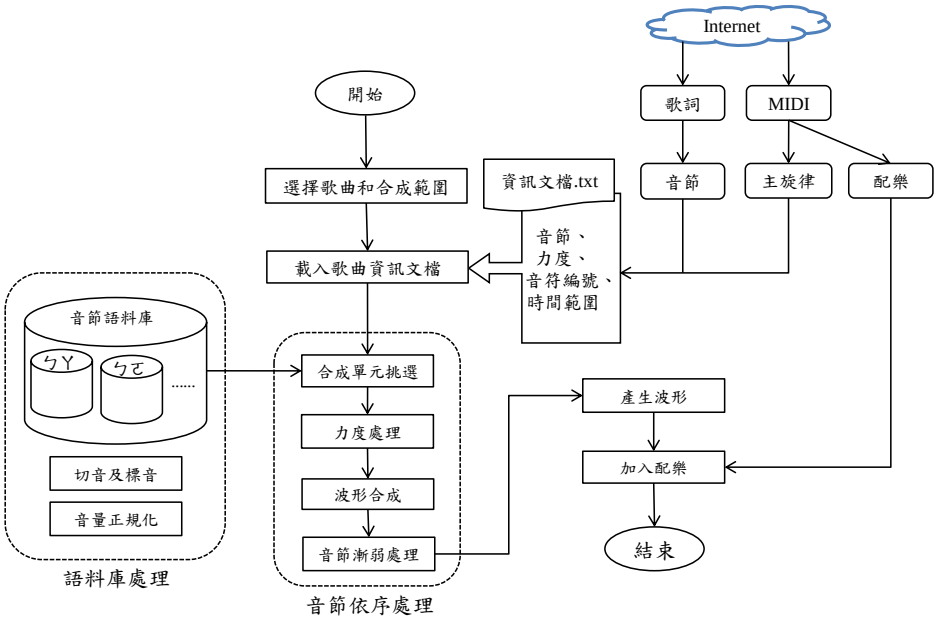


圖 2、 中文歌聲合成系統流程圖

二、資料處理

(一)、訊息處理

本研究中歌詞是根據魔鏡歌詞網來蒐集。音節的部分，是根據網際智慧股份有限公司的中文拼音查詢系統來進行轉換。轉換的步驟為將原本整首歌詞的標點符號去除，然後將整首歌詞放入此系統中，便可以輸出整首歌詞的音節。將收集的 MIDI 利用 Guitar Pro 軟體來分離主旋律和配樂。其中，主旋律中的資訊包括音高 (這裡指的是 MIDI 音符編號) 、音符的起始時間、音符的結束時間和力度。接著將音節歌詞加入，並寫成一個合成資訊的文檔。如表 1 所示。

表 1、 合成資訊文檔片段

音節	力度	音符編號	時間範圍
ㄅㄨ	95	69	11.5384-11.7692
ㄓ	79	70	11.7692-12
ㄗㄣ	79	72	12-12.2308
ㄇㄜ	79	70	12.2308-12.4615
ㄕㄜ	79	69-67	12.4615-12.6923-12.9231
ㄑㄩ	95	69	12.9231-13.7692

根據表 1 來解釋。由左到右分別為歌詞轉換後的音節、力度、 MIDI 音符編號和時間範圍。力度為 MIDI 資訊裡的參數，其解釋為音符所按下去的速度。數值愈高表示按下去的力道愈大，所演奏出來的聲音也就愈大聲。在音符編號的欄位裡，大於兩項代表此音節有轉音的現象。時間範圍為此音節在歌曲裡所在的時間位置，以秒為單位。

(二)、合成單元建立

為了能錄製所有的歌唱音節，錄音是根據注音符號檢字表來錄製。錄製的對象為一個喜歡唱歌，且音準不算太差的男性語者。錄製的地點為安靜的研究室或者安靜的宿舍空間。錄音軟體為 Goldwave ，錄音設備為 Audio-Technica 的麥克風，型號為 AT9942 。規格採用頻率 16000Hz 、 16bit 的 wave 檔格式。為了使合成出來的音色更像語者的聲音，分別錄製在 midi 音符編號 44、50、56 左右 3 組不同的音高來當作合成單元。。錄製時與麥克風的距離為 15 公分，音節的發音盡量持平。之後，我們將錄製好的 3 組音檔分別切除前後非語音的部分，並根據自相關函數 (Auto-Correlation Function, ACF) [11] 來進行音高追蹤。自相關函數是一個基於時域的方法，主要是使用自相關來計算一個音框和本身音框的相似度。根據式子 3 來說明，其中 $s(i)$ 為語音訊號， τ 是時間延遲量。接著，我們可以找出一個合理的 τ 值，就可以算出此音框的音高。

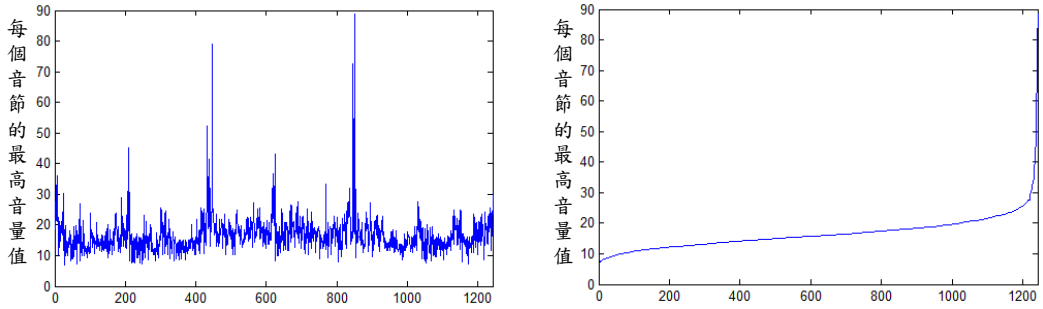
$$acf(\tau) = \sum_{i=0}^{n-1-\tau} s(i)s(i+\tau), \tag{3}$$

換句話說，自相關函數的作法為，將音框每次向右平移一點，和原本音框重疊的部分做內積。重複 n 次後會得到 n 個內積值，根據此方式可以找到音高週期 (pitch period) 。接著算出音節平均的音高週期。最後將平均的音高週期取倒數，就是此音節

的音高頻率 (pitch frequency)。音高頻率轉換成音符編號的式子如 4 所示。為了能更準確的追蹤音高，我們將音符編號取到小數第四位來標記每個錄製音節的名稱。

$$\text{音符編號} = 69 + 12 * \log_2(\text{音高頻率}/440). \quad (4)$$

因為原本錄製的音節，聲音大小不一致，故在這裡將音量做正規化。我們將音框設為 128 個取樣點，並對音節每個音框內的值，取絕對值後累加起來，來當作音量值。接著在這個音節內取最高的音量值來當作音量正規化的標準。根據全部共 1242 個音節，其最大音量值如圖 3 (A) 所示。將其排序後可以發現到，最大音量值集中在 20 左右，故我們將其定為正規音量的標準。



(A) 1242 個音節　　(B) 根據最高音量值排序

圖 3、 正規音量值分析圖

根據以上分析的結果，其音量正規的流程圖如圖 4 所示。 (A) 為原本錄製的音檔，之後將每個音框內的值，取絕對值後累加起來得到 (B) 。接著根據最大音量值 20 來縮放整個音節的音量值 (C) ，最後調整原本的波形使之正規化 (D) 。

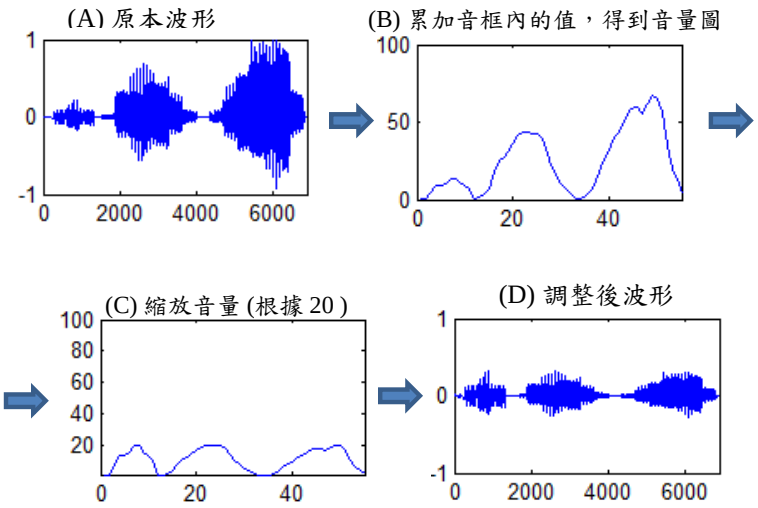


圖 4、 正規音量流程圖

(三)、合成單元挑選

根據之前每個音節，所錄製 3 組不同音高的音符編號 ，此小節說明如何來挑選合成單元。在語料庫裡，有 414 個音節，每個音節有 3 個不同音符編號的 wave 音檔。每個音檔的檔名是根據音符編號來命名。因此，根據欲合成的音符編號，來選取差值最小的合成單元。如圖 5 所示。

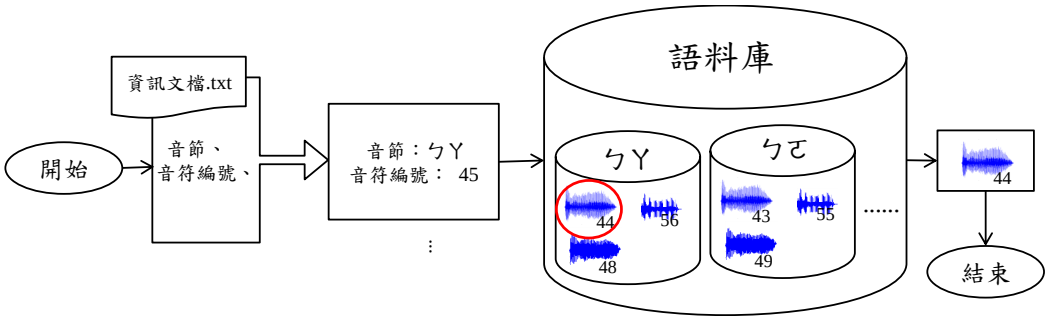


圖 5、 合成單元選擇流程圖

三、合成方法

(一)、音量調整

在做完音量正規化之後，我們就可以根據合成資訊文檔中的力度，來調整每個音節的音量大小。這裡採用的是較為簡單的式子如 5 所示。其中， $y[n]$ 為輸入訊號 $x[n]$ 音量調整後的訊號， V_{max} 為在合成資訊文檔中出現最多次的力度， V_i 為欲調整音量音節的力度。經由這樣的方式，能讓合成的歌曲音量更有變化。

$$y[n] = x[n] * \frac{V_i}{V_{max}} \tag{5}$$

(二)、時域上基週同步疊加法簡介

同步疊加法 (Synchronized Overlap-Add, SOLA) [12] 是由疊加 (overlap-add) 技術經過改良之後的一種方法。疊加的演算法較為簡單，只藉著重疊的方式來處理訊號。但所得到結果不佳，造成的失真也相當大。而同步疊加法藉著重新放置資料來控制聲音的播放速度，因此在時域上以同步疊加法來做修改，可以得到較好的聲音品質。

時域上基週同步疊加法 (Time Domain-Pitch Synchronous Overlap and Add, TD-PSOLA) 是將時域的波形訊號以漢明窗切割，然後分別對切割出來的波形做處理，再重疊相加。這個方法可以盡量不改變波形，將語音訊號的週期拉長或縮短。所以合成出來聲音，音色與原本的聲音不會相差太大。在品質與自然度上也有不錯的表現。因為是在時域上做計算，因此計算度也比在頻域上做處理的合成方式低。

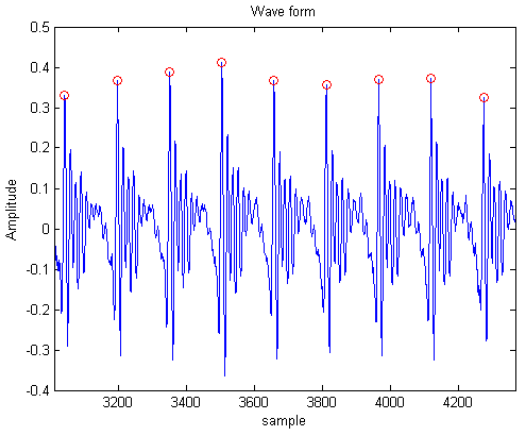


圖 6、 語音訊號中求取基週標位示意圖

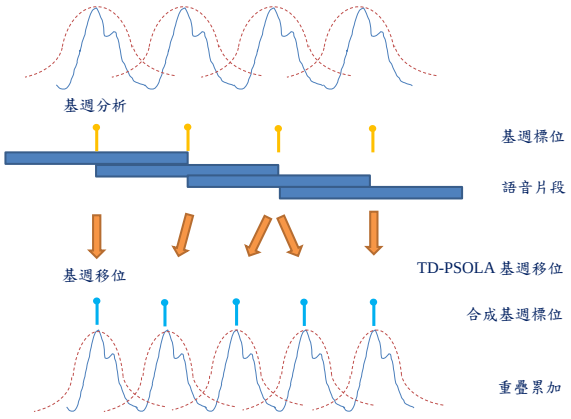


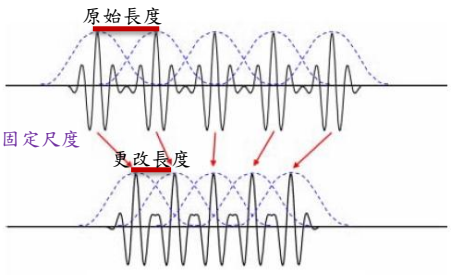
圖 7、 TD-PSOLA 示意圖

在進行 TD-PSOLA 之前，要先做語音的基週標位 (speech pitch mark) 。語音基週標位提供合成階段韻律調整之資訊，其理論的基礎是從一組聲音訊號下找出全域最大值之位置，接著再從左右兩邊來尋找區域最大值位置，圖 6 為基週標位示意圖。有了基週標位的資訊之後，就可以來進行 TD-PSOLA 演算法，如圖 7 所示。將語音訊號根據其基本週期分割成重疊且較小的訊號，接下來根據欲合成的音高來調整其基週標位。在兩兩基週標位上乘上一個漢明窗來重新加成，最後將剩下的語音訊號經過疊加的方式重新結合。使用這樣的方式，可以保持基週標位上主要的特徵。雖然訊號的長短與基本頻率改變了，但對於原本頻譜的破壞相對減少了許多，音色也不太容易變質。在音長調整方面，在此應用線性投射 (linear mapping) 的原則，做指標位置的對應。進行音長延長時，將部份分析音框重複。若要縮短音框，則刪除部分分析音框即可。

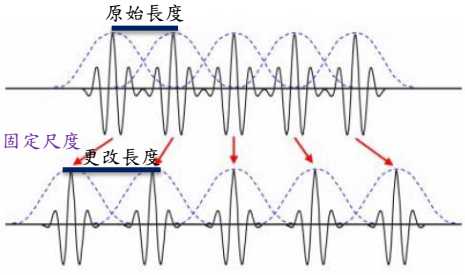
(三)、後續處理

1、轉音處理

根據圖 8 的 TD-PSOLA 變調的示意圖，發現到是根據固定的尺度來修改整體的音高。



TD-PSOLA音高調高示意圖



TD-PSOLA音高調低示意圖

圖 8、 TD-PSOLA 變調示意圖

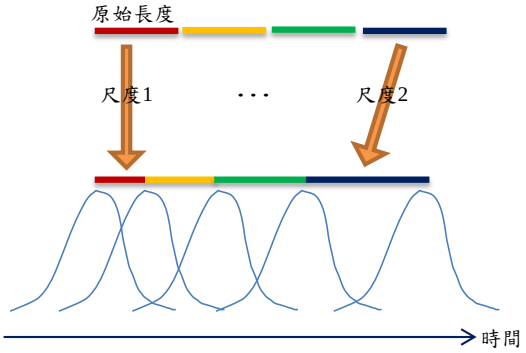


圖 9、 動態更改尺度示意圖

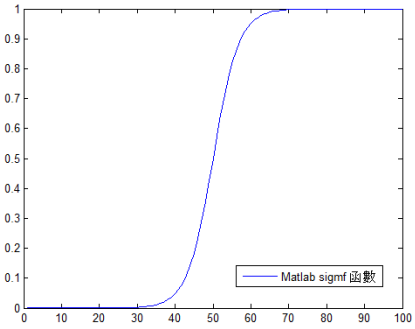


圖 10、 Sigmoid 函數

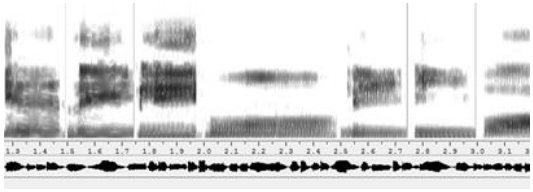


圖 11、 音節尚未做漸弱之頻譜圖

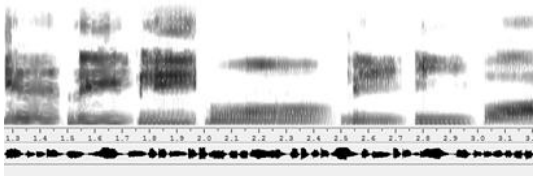


圖 12、 音節做過漸弱處理的頻譜圖

因此我們如果在疊加的過程中動態更改尺度的話，就可以達到轉音的效果，如圖 9 所示。為了更接近真實的轉音，這裡採用 Sigmoid 函數來對尺度的變化作處理。 Sigmoid 函數為一個 S 型函數，其式子如 6 所示。其中 a 表示 Sigmoid 函數開口的方向， c 為偏移值。基本上就是根據音節的長度和轉音音高的差值來做調整。圖 10 為在 Matlab 中的 Sigmoid 函數，在這裡將參數 a 設定為 0.3 來做轉音的處理。

$$f(x) = \frac{1}{1 + e^{-a(x-c)}}. \tag{6}$$

2、音節漸弱處理

根據之前的步驟，將每個音節合成出來，接著根據合成資訊文檔中的時間資訊，將音節對應到相對應的時間位置上。發現到音節彼此相接有雜訊的產生，將其轉到頻域上如圖 11 所示。因此，在每個合成完的音節之後做漸弱處理，如式子 7 所示。

$$y[n] = \begin{cases} x[n], & \text{if } n = 1, ..., (N - L - 1) \\ x[n] * \frac{N-n}{L}, & \text{if } n = (N - L), ..., N. \end{cases} \tag{7}$$

其中 N 為音節的長度， L 為漸弱的長度，範圍為 $1, ..., (N - 2)$ 。經由漸弱的方式處理完之後，音節串聯之後的雜訊有所改善，如圖 12 所示。

四、系統實作

(一)、系統流程

首先會載入歌曲清單，之後使用者可以選擇欲合成的歌曲和欲合成的範圍。經由使用者所選擇的歌曲，系統會去資料庫找出合成的資訊文檔、歌詞和配樂音檔。接著根據其資訊顯示歌詞、音節、力度、音符編號和時間範圍。

合成階段的流程圖如圖 13 所示。一開始會根據所選的歌曲，找出其歌曲資料夾底下的合成資訊文檔，裡面的訊息包括歌曲中每個音節的語言、力度、音符編號、開始時間和結束時間。一開始會將第一個音符編號定位在 48 ~ 60 之間，其餘音節的音符編號根據其差值來做調整。這裡提供了可以修改合成資訊的功能，包括了音節的修改和整體音符的修改。確定完合成資訊之後，系統會根據音節的順序，依序去音節語料庫尋找合適的合成單元進行合成。接著調整音量大小，然後將訊號經由 TD-PSOLA 來變換音高和長度，若有轉音的情形發生則做轉音的處理。接著對每個音節做漸弱處裡，然後根據文檔裡的時間資訊將合成的音檔串接起來。最後這裡提供了加入配樂的功能，配樂音檔是由 midi 中抽取出來並將合成歌曲整批對應到配樂上，並且可以調整配樂的音量大小。

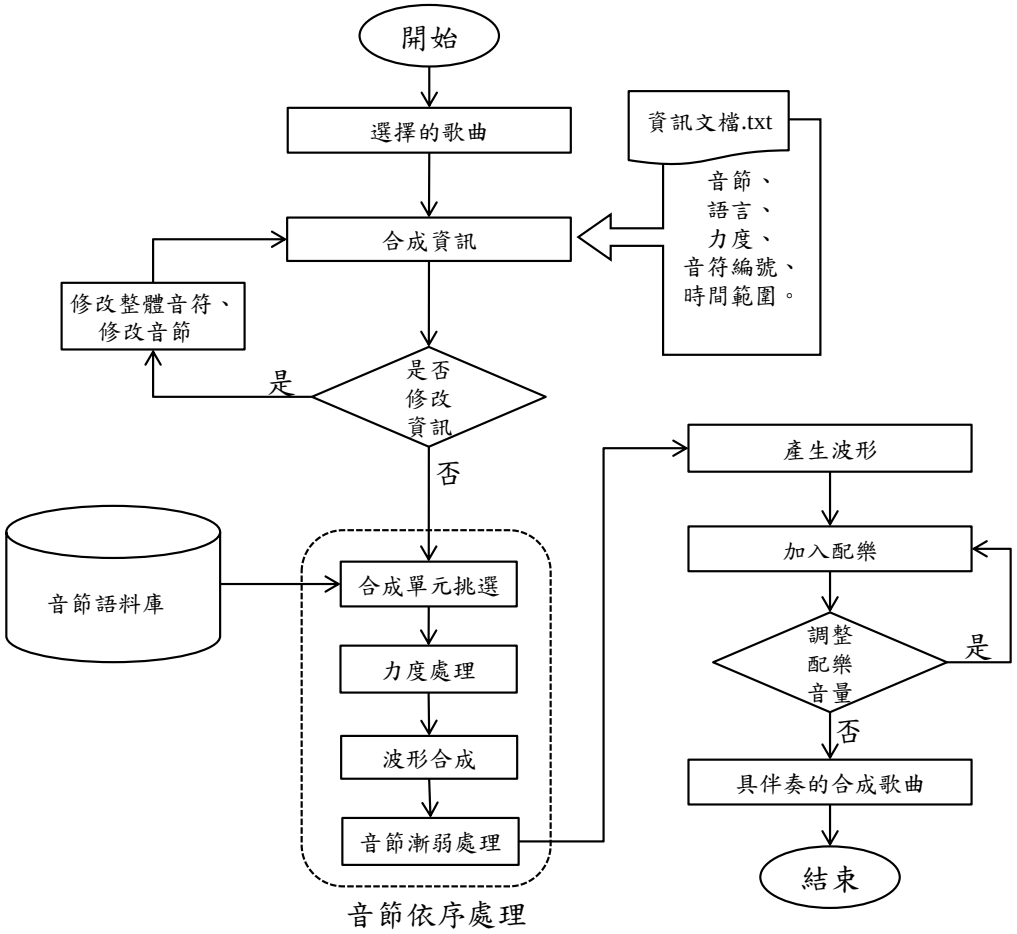


圖 13、 合成階段流程圖

(二)、介面實作

本小節實作一個具有配樂之歌唱合成系統，圖 14 為歌唱合成系統介面圖。 A 部分為載入歌曲清單； B 部分為選擇歌曲和合成範圍； C 部分顯示歌詞、音節、力度、音符編號和時間範圍； D 部分為合成按鈕組； E 部分為合成歌曲的波形圖。

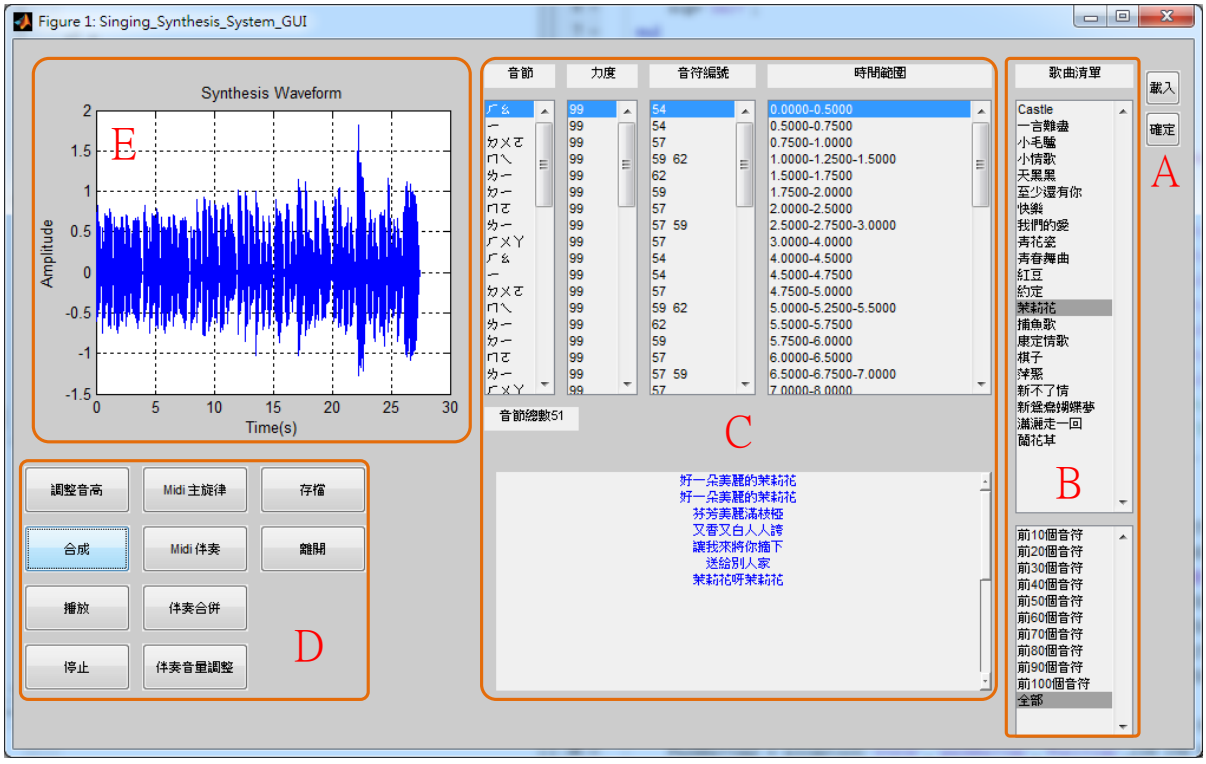


圖 14、 歌唱合成系統介面圖

五、實驗

歌曲的合成評估，主要分為品質、清晰度和相似度三個部份。品質代表合成出來歌聲品質的好壞；清晰度主要是評估合成出來的歌聲訊號聽起來是否清楚無雜訊，以及咬字的清楚程度；相似度則是評估合成歌聲與原唱的接近程度。評測人數為 10 人。為了要盡量分析多種曲風，故先將歌曲做分類。由於抒情類歌曲較多，因此選了兩首來當作評估的歌曲，其它分類各選一首歌曲。最後，我們將選擇出來的歌曲用在 TD-PSOLA 系統加入配樂的品質評估、清晰度和相似度的評估上，分類的種類如下所示。

- ◊ 童謠：適合孩童念誦的歌謠。
- ◊ 民謠：坊間流傳的歌謠。
- ◊ 抒情：節奏慢，且曲風溫柔的歌曲。
- ◊ 快節奏：拍子較短，速度較快的歌曲。
- ◊ 悲壯：歌曲帶有一點悲傷，且副歌通常給人激昂的感覺。
- ◊ 中國風：特性為穿插一些艱深的字詞和文言文。
- ◊ 節奏藍調：特性為歌曲帶有一點憂鬱，且會有重複的現象。

(一)、品質評測

在品質測試中，是根據 平均評定得分 (Mean Opinion Score, MOS) 的 5 分評分制度，如表 2 的評分方式來打分數。因為歌唱總是伴隨著配樂，因此我們將原本歌曲裡的配樂加到合成歌曲，並且比較結果。

首先，我們先做沒有配樂的比較實驗。一開始我們讓受測者聽原唱，並將其定為 5 分。接著，將合成步驟中的轉音處理、音節的串接處理、音量正規化處理和力度處理拿掉，並且只保留一組音高 (音符編號 50 左右) 來當作 1 分 (baseline) 的評測標準。然後讓受測者聽合成出來的歌聲 (TD-PSOLA) 並給予分數。接著，我們進行歌曲加上配樂的實驗。首先讓受測者聽原唱加上配樂，也將其定為 5 分，然後將之前 baseline 的歌曲也加上配樂並評為 1 分，最後讓受測者聽取合成加配樂的歌聲，並給予分數。

表 2、 品質評分表

類別	優秀	很好	普通	不好	糟糕
分數	5	4	3	2	1

TD-PSOLA 的品質評測結果如表 3 所示。整體上來說，加入配樂後的平均分數比未加入配樂要來的高 0.4 分。其解釋為在人類的聽覺上，加入配樂可以縮短和原唱加上配樂的差距。

表 3、 品質評測表

編號	曲風	歌曲片段	未加入配樂			有加入配樂		
			原唱	baseline	合成	原唱	baseline	合成
1	童謠	捕魚歌	5	1	3.2	5	1	3.6
2	民謠	茉莉花	5	1	2.7	5	1	3.8
3	抒情	天黑黑1	5	1	1.9	5	1	2.5
4	抒情	天黑黑2	5	1	1.8	5	1	1.9
5	抒情	紅豆1	5	1	2.4	5	1	2.7
6	抒情	紅豆2	5	1	2.5	5	1	2.6
7	快節奏	新鴛鴦蝴蝶夢1	5	1	2.1	5	1	2.3
8	快節奏	新鴛鴦蝴蝶夢2	5	1	2.6	5	1	2.7
9	悲壯	我們的愛1	5	1	2.2	5	1	2.5
10	悲壯	我們的愛2	5	1	2.9	5	1	3.3
11	悲壯	我們的愛3	5	1	2.7	5	1	3.1
12	中國風	青花瓷1	5	1	2.5	5	1	3.3
13	中國風	青花瓷2	5	1	2	5	1	2.7
14	節奏藍調	龍捲風	5	1	1.9	5	1	2.2
平均			5	1	2.4	5	1	2.8

(二)、清晰度評測

在清晰度測試中，也是根據 MOS 評分制。首先讓受測者先聽一段乾淨無雜訊，且咬字清楚的原唱當作參考，並將其分數評為 5 分， 1 分的評測標準同上。最後，每位受測者在聽完合成的歌聲之後，隨即在聲音清晰程度的表現給予 1 到 5 分的分數。

合成歌曲清晰度的評測結果如表 4 所示；清晰度的評測重點為歌聲訊號是否清楚無雜訊，和咬字的清楚程度。從表中我們可以看到【捕魚歌】和【茉莉花】的分數是較高的。但是在其它歌曲方面，分數明顯的較為低落。根據【新鴛鴦蝴蝶夢1】這首歌曲來分析其原因，本系統沒有做音節上子音和母音的延長處理，使得某些音節在經由TD-PSOLA的過程中，少了重要的發音資訊。因此，若是合成的音節較原本的合成單元短，會讓合成的聲音聽起來不清楚。一方面可能的原因為，音節的起始保留的空白訊號太短，造成串接合成上有雜訊的產生。

表 4、 清晰度和相似度評測表

編號	曲風	歌曲片段	清晰度評分	相似度評分
1	童謠	捕魚歌	3.8	3.6
2	民謠	茉莉花	3.2	3
3	抒情	天黑黑1	2.5	2.1
4	抒情	天黑黑2	2.1	1.9
5	抒情	紅豆1	2.1	2.3
6	抒情	紅豆2	2.8	2.7
7	快節奏	新鴛鴦蝴蝶夢1	1.9	2
8	快節奏	新鴛鴦蝴蝶夢2	2.8	3.2
9	悲壯	我們的愛1	2.3	2.2
10	悲壯	我們的愛2	3.1	3.3
11	悲壯	我們的愛3	2.9	2.6
12	中國風	青花瓷1	2.1	2.5
13	中國風	青花瓷2	2	2
14	節奏藍調	龍捲風	2.3	1.8
平均			2.6	2.5

(三)、相似度評測

在相似度測試中，也是採用 MOS 評分制。首先讓受測著先聽完原本真人的歌聲，分數為滿分 5 分。 1 分的歌曲標準和之前的一樣。接著，讓受測者聽完合成的歌聲之後，隨即給予 1 到 5 分的分數來評測與真人歌聲的相似程度。

最後，相似度的評測結果如表 4 所示。相似度評測的重點為，評估合成歌聲與原唱接近的程度。根據表中我們可以發現到【捕魚歌】和【茉莉花】因為有轉音上的處理，所以具有較好的表現。其它的歌曲分數就普普通通。值得我們注意的是【龍捲風】這首歌曲，在相似度的表現上差強人意。分析其原因為，這首歌曲沒有轉音上的表現，且在真人的歌聲中音節的連接較為連續。因此，之前的合成步驟中，對音節間的連結做漸弱處理，這會導致音節間的不相連。

六、結論與未來方向

在本研究中，我們採用時域上基週同步疊加法來處理合成單元。之後實作一個串接式的歌唱合成系統，用來產生具有配樂的合成歌聲。我們分別錄製 3 組不同音高的語料庫，來當作合成單元。根據 100 首左右的 MIDI 歌曲的資訊來分析訊息，並蒐集歌曲的歌詞和所對應的音節。根據本系統，使用者可以選擇想要合成的歌曲，並且修改歌曲整體上的音符編號和音節。歌曲的合成中，包括了合成單元的挑選、音量正規、力度處理、音節連接時漸弱的處理和合成時間的對應等等，最後將產生出來的合成歌聲與配樂同步結合。接著，利用主觀評測方式來分析此系統的優、缺點並做改進。品質評估方面，串接式的合成還是有一定程度的表現，尤其是在加了配樂之後，分數大部分呈現上升的情形。清晰度評估方面，由於沒有做音節上子音和母音的處理，除了某些歌曲表現較好之外，其它歌曲就表現的普普通通。相似度評估方面，有些歌曲沒有轉音上的表現會造成分數較為低落。另外，本系統沒做連音的處理，會使得合成出來的歌聲不像真人所演唱的歌聲。在轉音、振音或顫音等唱腔的部分，因為著手的語音資料並不是那麼的多，將來如果能蒐集到更多資料來進行分析，應該能使系統更加完善。最後，本研究提出的方式，可以推廣到其他語言的歌聲合成，也可以應用在哼唱的歌唱合成。

參考文獻

[1] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2-3, pp. 175–187, June 1992.

[2] C. Hamon, E. Moulines, and F. Charpentier, "Diphone synthesis system based on timedomain prosodic modifications of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1989, pp. 238–241.

[3] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, December 1990.

[4] V. Colotte and Y. Laprie, "Higher precision pitch marking for TD-PSOLA," in *XI European Signal Processing Conference- EUSIPCO 2002*, Toulouse, France.

[5] F. J. Charpentier and M. Stella, "Diphone synthesis using an overlapadd technique for speech waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1986, pp. 2015–2018.

[6] J. Bonada and A. Loscos, "Sample-based singing voice synthesizer by spectral concatenation," *Proceedings of the Stockholm Music Acoustics Conference*, August 2003. [Online]. Available: files/publications/SMAC2003-aloscos.pdf

[7] X. Rodet, "Synthesis and processing of the singing voice," in *IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, November 2002.

[8] H. Kenmochi and H. Ohshita, "VOCALOID - Commercial singing synthesizer based on sample concatenation," in *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium*. ISCA, August 2007, pp. 4009–4010.

[9] H.-Y. Gu and H.-L. Liau, "Mandarin singing voice synthesis using an HNM based scheme," in *Proceedings of the 2008 Congress on Image and Signal Processing, Vol. 5 - Volume 05*, ser. CISP '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 347–351.

[10] J.-C. Wang, H.-Y. Gu, and H.-M. Wang, "Mandarin singing voice synthesis based on harmonic plus noise model and singing expression analysis," in *Technical Report, Spoken Language Group, Institute of Information Science, Academia Sinica, Taipei*, March 2008, pp. 1–8.

[11] Y. Tabata and T. Shimamura, "Noise robust pitch extraction based on auto-correlation analysis in the frequency somain," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, video and Speech Processing*, May 2001.

[12] S. Roucos and A. Wilgus, "High-quality time scale modification of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1985, pp. 236–239.