中文歌聲合成之線上應用平台架設

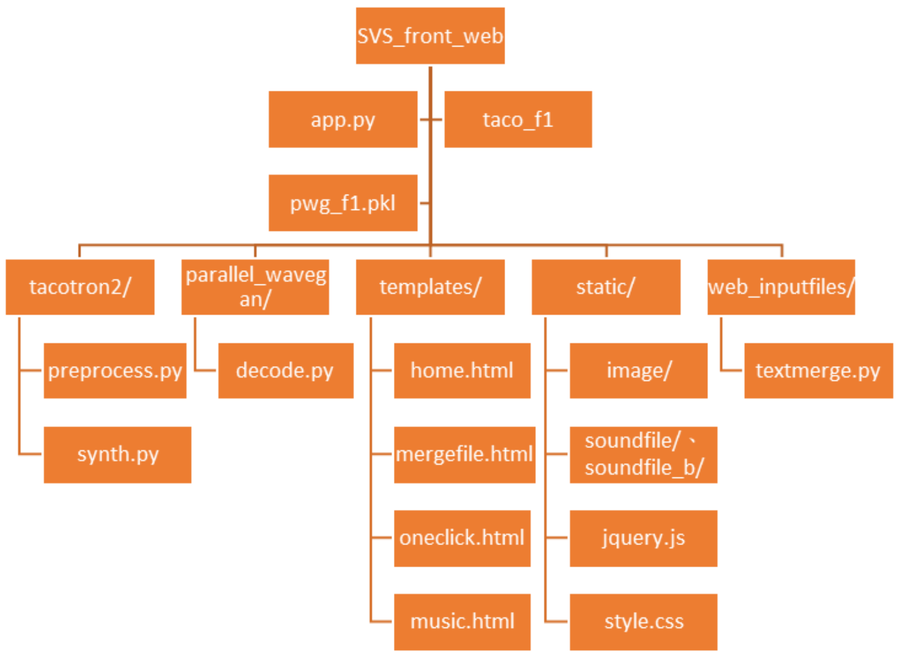
Construction of an Online Application Platform for Chinese Singing Voice Synthesis

組別：A212 指導教授：劉奕汶　組員姓名：王曉涵、王威智

INTRODUCTION

整理簡化實驗室研究現存的中文歌聲合成模型，並透過網路平台連接，只要上傳符合規則的文檔，後端即可在已經建置好的環境中做相關的前處理，並且使用中文歌聲合成模型生成與使用者自己編寫的文檔對應之音檔，不必設定複雜的運作環境以及各式輸入輸出的路徑，讓不熟悉相關技術內容的一般使用者也能輕鬆操作。

MAIN STRUCTURE



TECHNOLOGY

❑ **Tacotron 2**

Tacotron 2為一個長短期記憶模型（long short-term memory，LSTM）架構的的遞歸神經網路（recurrent neural networks，RNN），屬於文字轉語音（Text-to-Speech）的相關應用。考量到歌唱比起說話，若要聽起來自然會有更嚴格的音韻學觀念，經過先前的實驗室學長姐的編寫，將此Tacotron 2轉換成更適合用於歌聲合成（Singing-Voice-Synthesis）的架構，Tacotron2能夠利用訓練好的taco_f1模型，將包含歌詞與旋律的文字文件，轉換成Mel-spectrogram。

❑ **Parallel WaveGAN**

Parallel WaveGAN為神經網路聲碼器（Neural Vocoder），用途為將Mel-spectrogram解碼成時域的波型，也就是人耳所能夠聽到的聲音檔。藉由實驗室畢業學長過往蒐集之Mpop600 Dataset中的女聲進行訓練而成的pwg_f1.pkl模型，Tacotron 2所產生之Mel-spectrogram經過此聲碼器即可轉換成相當擬真且自然的歌聲。

LAYOUT

OPERATING PROCESS

I. 使用說明

一進入網址首頁會看到三行可下拉或隱藏的使用說明，初次使用的使用者需花3-5分鐘閱讀輸入檔案需遵守的格式以及限制，接著按照指示再使用者自己的本地完成欲輸入的文檔，包括歌詞、音高、音長三個檔案(或者合併，後面會有較詳細的敘述)。考慮到使用手機的用戶可能沒有適合的文字編輯器，我們也提供了適合的線上文字編輯器連結，讓使用者能夠直接切換過去完成自己想要輸入的文檔。

II. 一鍵合成

準備好文字檔後可選擇兩種合成方式，第一種方法是使用者自行將歌詞、音高、音長三者用既定格式打在同一文字檔上傳後進行合成，第二種方法是使用三個個別的文字檔上傳，由後端將三個文檔的字串合併做適當的前處理後進行合成。為保持頁面整齊美觀，上傳文檔及合成的畫面會跳轉至與使用說明所在的首頁不同的頁面，並在按下合成鍵後以閃爍的圓形提示提醒使用者合成正在進行中。

III. 儲存播放

音檔合成好後頁面會自動跳轉至有播放器的頁面，此處會顯示使用者剛剛合成好的音檔，以及曾經合成過的使用者的作品，點選檔名即可播放，亦可下載。在確認過當前合成音檔和歷史合成音檔後，當前的使用者可以選擇是否儲存自己的作品到歷史檔案供後續使用者欣賞，或是選擇直接刪除。

IV. 了解更多

完成以上動作後，使用者可以透過長駐在各個頁面的導覽列回到首頁進行其他作品的合成，或是前往此網站所使用中文歌聲合成技術的paper連結，以及指導教授實驗室的簡介。

CONCLUSION

I. Flask（v2.0.2，後端框架）

一般來說，網站架設的後端架構並非使用python，application server（含API）本身可以直接與Web Server進行響應。但若後端要架設一個以python 為正常運行的網站，則需透過WSGI Server（Web Server Gateway Interface Server），這是由於一般Web Server無法判讀python語法，WSGI Server定義了HTTP Request和python相互溝通的規範，使開發人員可以專注於應用程式的開發。

本實驗室的機器學習演算法皆以python程式語言撰寫，網路上也有將機器模型使用Flask開發應用程式發布於網路的先例，故我們選擇使用python之Flask套件架設後端API。使用Flask框架好處是我們可以直接以同樣的語法執行固有模型並且進行網頁開發，尤其它還內建豐富的函式，可與前端html網頁端進行良好的整合，這個套件同時具備簡易的（同一時間只能接收單一指令）WSGI Server，因此在網站的測試階段，搭配開發時所用的電腦（充當Web Server），只要在Flask應用程式運行當中，就可以完成簡易的網站供使用者遠端測試，接收並處理個別使用者的需求。

II. Bootstrap（v5.1.2，前端網頁框架）

Bootstrap為知名網頁框架，利用簡單的引用就可以使網頁有好看的外觀，這個網頁框架的便利性也受到市場上許多網頁開發人員的喜愛，經常被使用於製作響應式網頁。舉例來說，若我們需要製作一個按鈕，可以到其官網瀏覽官方文件 並複製引用至欲設計的HTML網頁中，如此一來就可以很快速地做好一個按鈕。

Bootstrap可以大大的省去學習CSS與JavaScript所需的時間，因此我們可以更專注於開發應用程式之功能。

CONCLUSION

為測試我們建置的網頁是否達到將中文歌聲合成模型更簡單的推廣給更多人了解、使用的目的，我找了6位測試者試用網頁，對於使用者給予的回饋，大致可分為中文歌聲合成模型本身以及網頁使用介面，使用者在操作過程中會發現模型本身的限制（部分字句不在字庫、趕拍、吃音等）以及體驗到模型實際合成出擬真且自然的歌聲，雖然合成限制會讓網頁回報錯誤給使用者，但這也符合我們希望使用者能夠了解此模型的初衷，故對此我們不做過多調整。而網頁使用介面我們發現了一個統一且能夠改善問題，即檔案上傳的網頁介面操作不直覺，容易讓使用者混淆，因時間限制我們將其視為未來改善的目標，希望能在專題競賽後將其解決。即使在做這個專題的一開始，我們甚麼都看不懂，但經過這段時間尋找、規劃並自主學習所有相關知識，以及向實驗室其他人請教，不只學習到如何架設網站，也見習了許多在機器學習上的課題，感謝一年來在這個實驗室所帶給我們的成長。