臺灣口音中英雙語之多語者影音合成系統
Taiwanese-Accented Mandarin and English Multi-Speaker Talking-Face Synthesis System

Chia-Hsuan Lin, Jian-Peng Liao, Cho-Chun Hsieh
Kai-Chun Liao and Chun-Hsin Wu
Department of Computer Science and Information Engineering
National University of Kaohsiung, Taiwan
{a1085512, a1085508, a1085540, a1085520}@mail.nuk.edu.tw
wuch@nuk.edu.tw

摘要

本論文提出一個多語者影音合成系統，結合語音複製與嘴型同步技術，透過取得任意語者短暫的說話語音及影像片段，以零樣本之遷移學習，來實現可即時翻譯的文字轉人物說話影像。除此之外，我們利用開源語料集訓練了多個臺灣口音的模型，同時也提出使用注音作爲合成器之文字嵌入的方式，來提升系統合成中英交雜語句的能力。透過此系統，使用者便可創造出豐富的應用，且此技術之研究與應用，在影音合成領域具有相當的新穎性。

Abstract

This paper proposes a multi-speaker talking-face synthesis system. The system incorporates voice cloning and lip-syncing technology to achieve text-to-talking-face generation by acquiring audio and video clips of any speaker and using zero-shot transfer learning. In addition, we used open-source corpora to train several Taiwanese-accented models and proposed using Mandarin Phonetic Symbols (Bopomofo) as the character embedding of the synthesizer to improve the system's ability to synthesize Chinese-English code-switched sentences. Through our system, users can create rich applications. Also, the research on this technology is novel in the audiovisual speech synthesis field.

關鍵字：多語者語音合成、語者驗證、語音複製、語碼轉換、嘴型同步、人物說話影像

Keywords: Multi-Speaker TTS, Speaker Verification, Voice Cloning, Code-Switching, Lip-Syncing, Talking-Face Generation

1 緒論

隨著人工智慧技術的蓬勃發展，大量人機互動相關的應用如雨後春筍般出現。而文字轉語音技術便扮演了一個不可或缺的角色，相關的研究也成了熱門的議題。現今，普通的單語者語音合成技術已經非常成熟了，如 Google、Microsoft 與 AWS 等皆有提供相關的應用程式介面（Application Programming Interface，API）服務可供使用。有了這樣的基礎後，眾多研究便希望可以將其延伸，發展出多語者語音合成系統。

爲了要實現多語者語音合成，並能應用在沒有見過的目標語者上，亦即達到語音複製（Voice Cloning）的效果，可透過引入語者驗證（Speaker Verification）機制到文字轉語音系統中的方式，也就是取一小段目標語者的聲音，再利用遷移學習來合成與目標語者相似的語音。該方法同時解決了訓練單語者語音合成模型，需要同一名語者大量語音資料的困難。然而目前大多數的研究都僅限在英文等外語，缺乏有效且實用的中文內容，同時臺灣口音的中文語音研究更是稀少。

有鑑於此，我們透過訓練影響口音最主要的合成器模型，並實驗了多組語料集，其中包含一個全臺灣口音中文語音的公開語料集 Common Voice Corpus[1]，藉此提出了一個臺灣口音的零樣本多語者語音合成系統（見圖 1），只需要參考短暫的目標人聲，便可複製目標語者之聲音，以產生任意內容的語音。我們也嘗試以注音來建立合成器中的字元嵌入（Character Embedding），並進行中英混雜語句合成的開發，有了語音複製的技術後，我們還可以將電腦生成的語音，利用嘴型同步（Lip-Syncing）的方式來合成出人物說話影像。

整合語音複製文字轉語音（Text-to-Speech，TTS）與嘴型同步，我們便可以建立一個文字轉人物說話影像合成系統，透過輸入目標文字、目標語者之語音與人臉（圖片或影片），即可產生目標語者在說目標文字的影像。我們亦加入了語音辨識和語言翻譯模組，可以辨識語者說話的原始內容，再進行中英雙向翻譯，使目標語者說出不同語言。透過該系統，能創

[1] https://commonvoice.mozilla.org

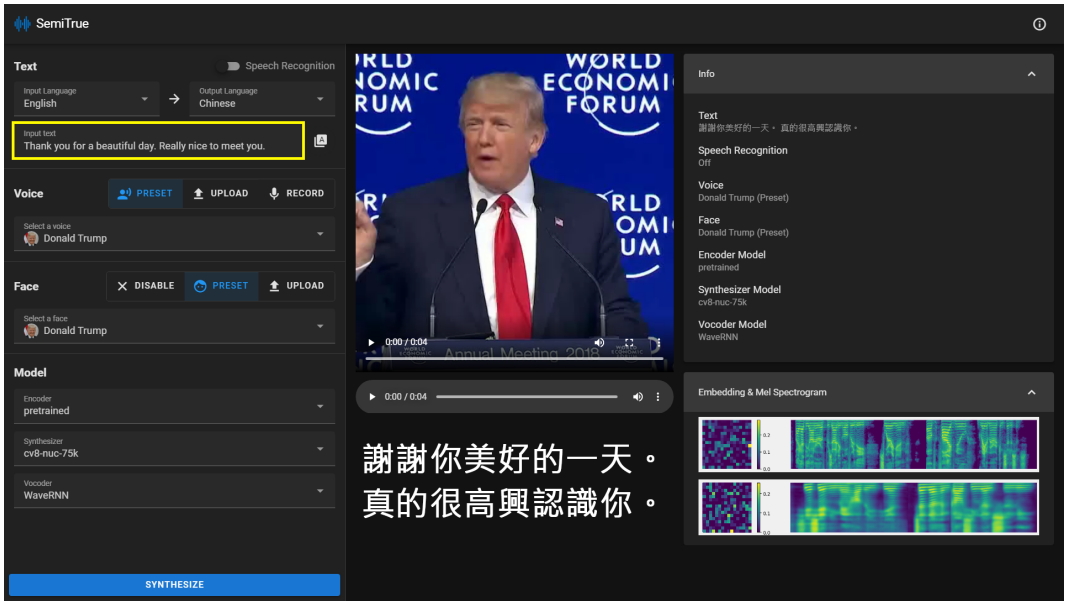

圖 1. SemiTrue 系統介面

造出豐富的應用，如多國語言電影的配音工作、演講錄音生成講者跨語影像等，而此技術的應用在臺灣可說是前所未見的。

本文之架構分成第 1 章簡述研究之內容，第 2 章介紹該領域的相關工作，第 3 章說明系統設計，第 4 章設置實驗來驗證成果，最後第 5 章爲本研究之結論。

2 相關工作

語音合成可以從文字來產生自然且可以理解的語音內容，而若要測量語音合成品質，主要有兩個重要的依據：可理解性及自然度。在語音合成技術發展的歷史中，經歷了許多的改進與演變，目前最爲熱門的技術，是基於神經網路的端到端語音合成 (Wang et al., 2017; Shen et al., 2018)，能在有限資料的背景下，合成出逼眞的語音 (Ning et al., 2019; Tan et al., 2021)。在此基礎上，許多關於多語者語音合成的想法被提出，包括建立語者編碼，並訓練多個語者、利用語者驗證 (Jia et al., 2018; Lőrincz et al., 2021; Neekhara et al., 2021)，來實現語音複製，或者透過語音轉換 (Zhang et al., 2019; Casanova et al., 2022) 來改變語者的聲音。而隨者全球化發展，很多人會同時使用多種語言，如中英夾雜的語句，這類可支援語碼轉換（Code-Switching）的語音合成系統 (Hung et al., 2019; Zhou et al., 2020)，也成了熱門的研究重點。

在中文語音合成領域，也有不少的研究出現 (Cheng and Chen, 2019; Hung et al., 2019; Wang and Chen, 2020; Wang and Huang, 2021)，其中 Wang and Huang, 2021 在改進 Tacotron 架構的同時，提供了有關語者驗證和語音轉換，來實現語音複製的實作。 Wang and Chen, 2020 針對臺灣口音的中文語音合成進行了研究，但未支援中英混雜語句的合成。

雖然中文語音合成的研究日漸增多，但卻鮮少有研究將文字轉語音合成與人物說話影像生成的領域結合，即便是外國有類似之研究 (Song et al., 2022)，也沒有包含語音複製的功能，我們的研究涉及了在語音領域的多項研究重點，包含臺灣口音、中文語音合成、語音複製以及語碼轉換等，而配合人物說話影像生成領域中嘴形同步技術，而提出的文字轉人物說話影像合成系統，實屬相當創新的概念，值得進一步研究發展。

3 系統設計

本研究提出了一個文字轉人物說話影像合成系統，並將其命名爲 SemiTrue。在本系統中，使用者可以選擇設定輸入語言、輸出語言、目標文字、3∼8 秒的目標語音以及目標人臉，其中目標人臉允許是圖片，或是長度大於合成音訊的影片。若是選擇輸入圖片，在最終所呈現的人物說話影像，則大部分會是靜態的，僅有嘴部區域會隨音訊內容而產生變化。另外，語音、圖片與影像皆有支援多種常見的格式，但在音訊進入系統後會被轉成 WAV 格式，而圖片或影像則會被自動縮放爲適當的大小，以利後續的合成工作。除了透過使用者上傳自定義的目標語音與目標人臉外，系統中已有事先載入好預設的語音與人臉，使用者僅需要輸入任意文字，就能合成出人物說話影像。

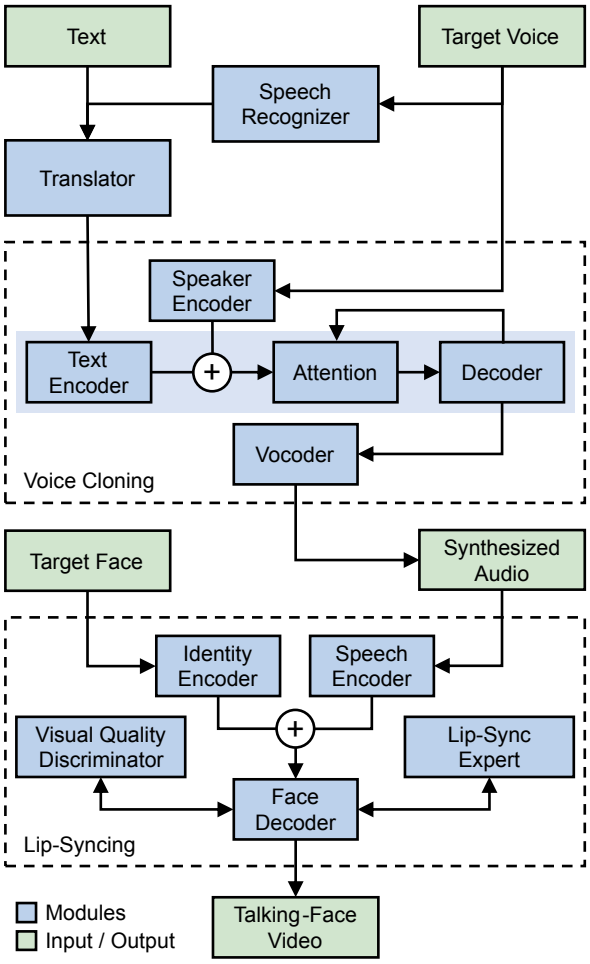

圖 2. SemiTrue 系統架構

SemiTrue 的整體架構，大致可以區分成以下四個模組：語音辨識、翻譯、語音複製以及嘴型同步。SemiTrue 整合了多個開源專案，並利用 FastAPI[2] 實作成網頁 API 服務，供使用者以網路請求的方式，來進行語音和影像的合成。同時，我們也建置了網頁前端，以圖形使用者介面來展現系統的功能。

SemiTrue 的運作流程如圖 2 所示，先將目標文字與目標音訊，依照輸出語言的不同，輸入語音複製模組中對應語言的複製器，來產生合成音訊，再將合成音訊與目標人臉之圖像或影像輸入嘴型同步模組（見圖 3），最終得到一段人物說話影像。在系統處理的過程中，根據不同使用者的需求，輸入之語音可以進行語音辨識，轉成文字內容，再進行文字翻譯後，才放入語音複製模組。如此一來就可以使目標人物，以其聲音說出不同語言的翻譯內容。

在後面的幾個小節中，我們將針對 SemiTrue 的四個模組，進行更詳細的說明。

[2]https://fastapi.tiangolo.com/

3.1 語音辨識模組

核心的文字轉人物說話影像功能，主要是透過使用者手動輸入目標文字，來作爲合成語音的文字來源。但在部分情境中，如僅有會議錄音，而無會議逐字稿的情況下，要合成出講者的語音及影像，甚至是進行語言的轉換，就會顯得較爲麻煩。有鑑於此，SemiTrue 另外加入了語音辨識模組，並利用了 Google 雲端平台的 Speech-to-Text API[3]，故在語音辨識穩定度與準確度上，能有一定程度的品質。

3.2 翻譯模組

SemiTrue 還可對使用者輸入的內容進行翻譯。舉例來說，使用者可將一段英文演講者的原始音訊輸入系統中，並經由語音辨識模組辨識其說話內容成英語原文，再透過翻譯模組將英文語句翻譯成中文語句，最後才進入到語音複製模組進行後續的複製工作。該模組的實作則使用了 Google Translate API[4] 進行翻譯。

3.3 語音複製模組

SemiTrue 語音複製模組的實作，大部分是基於 Jemine, 2019 的 Real-Time Voice Cloning[5]（RTVC）與其在中文的版本 MockingBird[6] 修改而來的。RTVC 的設計是源自於一個零樣本語音複製框架 (Jia et al., 2018)，其架構大致由一個三階段管線構成，分別爲語者編碼器（Speaker Encoder）、合成器（Synthesizer）與聲碼器（Vocoder），下文將分別介紹這三個階段的設計與我們改進的地方。

3.3.1 語者編碼器

爲了要實現語音複製，系統需要分離出不同語者的聲音特徵，合成器才可依此特徵產生出不同音色的人聲，這樣的方式稱作語者驗證。而要區分不同語者聲音的特徵，可以透過一個向量來表示，因此語者編碼器便負責從目標語音中抽取其特徵，包括音色、聲調、字詞發音，來形成語者嵌入向量，供後面的合成器使用。

RTVC 使用 GE2E (Wan et al., 2018) 方法來實現語者驗證，該模型將聲音轉換後的對數梅爾頻譜資訊，映射到一個固定的嵌入向量，以此來區分語者。一開始，目標語音被轉換成 40-Channel 的對數梅爾頻譜資訊，透過特徵萃取得到特徵向量，再經過 3 層 LSTM 組成之 768 個單元網路後，輸出 256 維的向量，再套用 L2 正規化，才能取得語者嵌入向量。

[3]https://cloud.google.com/speech-to-text
[4]https://cloud.google.com/translate
[5]https://github.com/CorentinJ/Real-Time-Voice-Cloning
[6]https://github.com/babysor/MockingBird

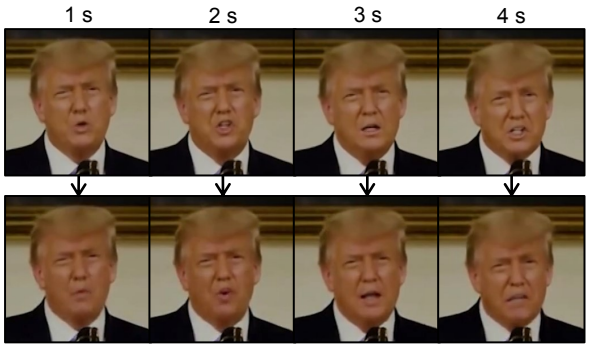

圖 3. 原始影像與嘴型同步後影像之比較

Method	Character Embedding
Pinyin	mei3 li4 de5 tai2 wan1
Bopomofo	ㄇㄟˇ ㄌㄧˋ ㄉㄜ˙ ㄊㄞˊ ㄨㄢ

表 1. 「美麗的臺灣」的字元嵌入

Method	Character Embedding
Pinyin	Hello，shi4 jie4 ！
Bopomofo	Hello，ㄕˋ ㄐㄧㄝˋ ！

表 2. 「Hello，世界 !」的字元嵌入

3.3.2 合成器

合成器部分採用了 Tacotron 2 (Shen et al., 2018) 的架構，其基本上是由文字編碼器、注意力層與解碼器所組成。文字編碼器輸入部分，可以選擇使用語音的文字符號（Grapheme），或者聲音符號（Phoneme）的串列，並將結果與目標語者的語者嵌入向量串接，使嵌入的內容得以傳給注意力層與解碼器，最後產生合成語音的梅爾頻譜資訊。該架構可以使模型將輸入文字映射到每個字音上，也可加快模型收斂的速度，並從訓練資料中學習生僻字詞與專有名詞的發音。

SemiTrue 在合成英文語句時，使用 RTVC 的設計，以文字符號序列輸入文字編碼器，該序列會映射到 26 個英文字母及常用標點符號。合成中文語句時，原本 MockingBird 採用了漢語拼音的聲音符號輸入方式，拼音由英文字母組合成發音，聲調則以數字表示，字與字之間以空白分隔，但在本研究的 SemiTrue 系統中，提出改以注音作爲聲音符號輸入，其中包含了 37 個注音、4 種聲調（一聲沒有標記）與標點符號。主要是因爲原 MockingBird 中的字元嵌入採用拼音，而拼音的聲音符號會與英文的文字符號重疊，使其不能應用於中英混雜的語句。基於此問題，我們設計了一個中英混雜的版本，輸入的映射序列含有 37 個注音符號、26 個英文字母、阿拉伯數字 0～9 及常用標點符號。在表 1 與表 2 中顯示了輸入的文字被轉爲字元嵌入符號的範例。

爲了抑制背景雜訊與強化複製人聲的韻律及音色，又加入了全域風格標籤（Global Style Token）(Wang et al., 2018)。原本在 TTS 系統中，若要學習語者的語調及風格，就必須提供上下文的風格關係，即便如此也難以評估是否爲正確的聲音風格，而全域風格標籤在訓練時，不需要任何實際的標籤，而是讓內部架構自行決定風格的樣式。也正是因爲沒有固定的標籤，可以讓系統實現多樣化的風格，且更可適應於長句子的合成上。

3.3.3 聲碼器

從前面的合成器，只能得到語音的梅爾頻譜資訊，並不是一般可以直接聆聽的語音文件，因此聲碼器主要的工作便是將梅爾頻譜資訊轉換成聲音波型，也就可以產生普通 WAV 格式的語音檔案。現今聲碼器的發展相當多元，有很多不同的實作，而每個方法都有其優缺點，因此 SemiTrue 加入了三種基於神經網路的聲碼器：WaveRNN (Kalchbrenner et al., 2018)、HiFi-GAN (Kong et al., 2020)、Fre-GAN (Kim et al., 2021) 以及一個基於演算法的聲碼器：Griffin-Lim，下文將對這四種聲碼器進行簡略的介紹。

WaveRNN：利用 Tacotron 時一般會配合 WaveNet 作爲聲碼器，WaveNet 使用自回歸（Auto-Regression）的方式生成聲音，根據前一刻的輸出來預測下一刻的數值，因採用多層架構，讓 WaveNet 可以生成高品質的自然人聲，但也因爲其高複雜度，進而導致生成速度緩慢，不適合在實際應用場景中使用。於是 WaveRNN 便針對生成速度做優化，透過將模型簡化與序列調整，使其可僅依靠 CPU 運算，來實現語音合成。

HiFi-GAN：使用生成對抗網路（Generative Adversarial Network）作爲模型基礎，與 MelGAN (Kumar et al., 2019) 較爲相似，其中包含了一個生成器與兩個鑑別器：多週期鑑別器（Multi-Period Discriminator）和多尺度鑑別器（Multi-Scale Discriminator），以此來強化 GAN 評斷眞實語音的能力。

Fre-GAN：在 HiFi-GAN 的基礎上，改良了生成器與鑑別器，採用 RCG（Resolution-Connected Generator）與 RWD（Resolution-Wise Discriminator）。而 Fre-GAN 最核心的改進是：傳統中在採樣時，使用了如平均池化等方法，會損失高頻區域，而 Fre-GAN 則選擇使用離散小波變換，可以完整保留聲音的所有資訊。

Synthesizer Model	Phonetic Method	Corpus
MockingBird	Pinyin	aidatatang_200zh + AISHELL-3
CV8-75K	Pinyin	Common Voice 8.0
Chinese-Hybrid	Pinyin	Common Voice 8.0 + aidatatang_200zh
Bopomofo	Bopomofo	Common Voice 8.0
Code-Switching	Bopomofo	Common Voice 8.0 + VCTK

表 3. 模型的字元嵌入方式與使用之語料集

Griffin-Lim：基於演算法的經典聲碼器，利用短時距傅立葉變換的大小，來重建聲音訊號。該演算法簡單且高效，但因爲 Griffin-Lim 在生成時沒有使用神經網路，所以在語音的合成品質較遜於上面所提及的聲碼器。

3.4 嘴型同步模組

對於 SemiTrue 嘴型同步模組的建構，本研究參考 Wav2Lip (Prajwal et al., 2020) 的方法與實作。Wav2Lip 主要針對一段人物說話影像，依據輸入語音的資訊，來產生與該語音同步的人物嘴型，並置換掉原始影片中人物的嘴部區域，最終產生出自然且準確的嘴型同步影像。

Wav2Lip 改進了由 Prajwal et al., 2019 所提出的方案，藉由引入了一個預訓練的嘴型同步鑑別器（Pretrained Lip-Sync Discriminator），其專注於調整嘴型同步的效果，並回饋到下一次生成，以此來提高嘴型同步的精準度，雖然透過嘴型同步鑑別器能夠產生準確的嘴形，卻可能會在連續影像中導致嘴部區域的模糊以及瑕疵。因此，Wav2Lip 又加入了一個影像品質鑑別器（Visual Quality Discriminator）作爲輔助，透過考慮連續多個影格，修正嘴型同步鑑別器所導致的模糊問題，進一步提升了影像的品質。同時，又因爲 Wav2Lip 是透過讀入語音頻譜資訊來生成影像的，使其可以運用於任何目標人臉與任何語言的人聲，即便是透過 TTS 合成出來的語音也沒問題，這便相當適合應用於我們的系統，來產生指定目標人物的說話影像。

Wav2Lip 的架構區分爲生成器與鑑別器兩部份，而生成器又可以再細分爲三個區塊，分別是人物編碼器（Identity Encoder）、語音編碼器（Speech Encoder）與臉部解碼器（Face Decoder）。在實際運作時，嘴型同步模組需要輸入目標人臉，與經由語音複製模組所產生的合成語音，之後系統會隨機挑選目標人臉影像中的一部分參照片段，與另一段移除嘴部區域後的片段，並將兩個片段連接在一起，作爲人物編碼器的輸入，產生出臉部嵌入向量。而合成音訊的梅爾頻譜資訊，則會進入語音編碼器來產生語音嵌入向量。將臉部嵌入向量與語音嵌入向量合成後，再送入臉部解碼器中，依目標人物的特徵來產生相應的嘴型同步影片，生成結果經過嘴型同步鑑別器與影像品質鑑別器來評斷生成品質，最終合成人物說話影像。

4 實驗

4.1 研究方法

在本章節中，我們將探討語音複製與嘴型同步的品質及效能，並針對中文語音合成部分，比較 MockingBird 與我們提出之改進方案的差異，包括使用拼音字元嵌入、使用注音字元嵌入、混合語料集、中文混雜語句之四個模型（見表 3），又以其中一個模型去探討：搭配 WaveRNN、HiFi-GAN、Fre-GAN 與 Griffin-Lim 聲碼器對語音合成效果的影響。爲了評斷 SemiTrue 所合成之人物說話影像的整體效果，我們也對比了同一名語者之原始影片、英文語句合成與中文語句合成的語音品質、相似程度與嘴型同步品質。

鑒於 Wang et al., 2017 等多篇論文都是以平均意見分數（Mean Opinion Score，MOS）(P.800.2, 2016) 來評斷影音之品質，因此，我們將不具標籤之語音及影像，整理成問卷的形式，每段語音或影像以最差 1 分至最佳 5 分進行匿名評分，來取得其平均意見分數。

另外，我們參考了 Jia et al., 2018 與 Wang and Huang, 2021 的實驗，透過語者相似度分析圖來檢視，不同語者間的離異性與語音複製的相似程度，及使用梅爾頻譜圖來觀察，中英混雜版本與純中文版本在合成效果上之差別，以提供除了合成語音經人類觀測的主觀評分外，較爲客觀的驗證方法。

4.2 模型訓練

SemiTrue 語音複製模組的架構中，包含語者編碼器、合成器與聲碼器，其中影響合成品質及口音最大的是合成器模型。若期望合成語音可以更貼近臺灣人的口音特色，必須使用含有臺灣口音的語料集來進行合成器模型的訓練。下文即列出了 SemiTrue 與 MockingBird 有使用到的語料集。

Common Voice Corpus：由 Mozilla 發行的開源多語言語音資料集，任何人都可以提供錄音，固定每 3 個月發布新版本。SemiTrue 大部分的模型是採用華語（臺灣）第 8 版進行訓練的，該版本有 1,695 名語者提供上萬筆共 89 小時的錄音。

NER-Trs-Vol1[7]：臺灣口音中文語料集，由北科大教育電台廣播節目錄製音檔組成，經過人工校正、切割並整理。因爲每段錄音時間較長，不宜用作訓練，故我們抽取其中一部分用作實驗中驗證的語音。

VCTK[8]：全英語之語料集，包含 110 名不同口音的語者，每位語者約提供 400 句朗讀報紙、演講段落的錄音，我們用於訓練中英混雜語句的英語部分。

Aidatatang_200zh[9]：中文北京口音普通話語料集，錄音主要爲多名語者在安靜室內透過 Android 與 iOS 手機錄製而成的，長達 200 小時，爲 MockingBird 作者訓練模型的語料集。

AISHELL-3[10]：中文北京口音普通話語料集，採用高保眞度麥克風錄製，經過嚴格的品質檢驗，包含 218 名語者，共 85 小時的錄音，爲 MockingBird 作者訓練模型的語料集。

爲了實驗拼音字元嵌入符號、注音字元嵌入符號、混和口音語料集、中英混雜語句合成的效果，我們分別訓練了四個模型，並命名爲 CV8-75K、Bopomofo、Chinese-Hybrid 與 Code-Switching，加上 MockingBird 作者提供的模型作爲對照組。特別要注意的是，在訓練 Code-Switching 模型時，因爲沒有大量乾淨的中英混雜語料集，於是採用了中文與英文語料集交錯訓練的方式。

4.3 實驗結果

原本 256 維的語者嵌入向量經過 UMAP (McInnes et al., 2018) 方法降維後，形成了可視化的圖表。透過這個分群，可從客觀的角度得知語料集中的語者聲音，經過 SemiTrue 的語者編碼器可以被有效的區分開來，其中圖 4 顯示出語者編碼器可以清晰的分離男女語者的聲音，而圖 5 則進一步顯示出同性別語者間，也能達成有效的分群，且經過合成後的語音也被分在同一群中，即語音複製的效果可以更貼近目標人聲。

在語音與影像合成品質部分，最終我們收集到了 120 份問卷回應，經過統計得到了測試資料的平均意見分數。表 4 中比較了四種合成器模型的語音品質，其中我們所訓練的三個

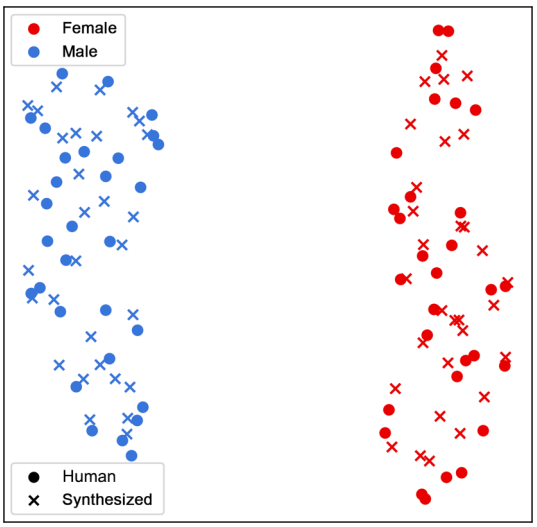

圖 4. 男女語者聲音分群的語者相似度分析

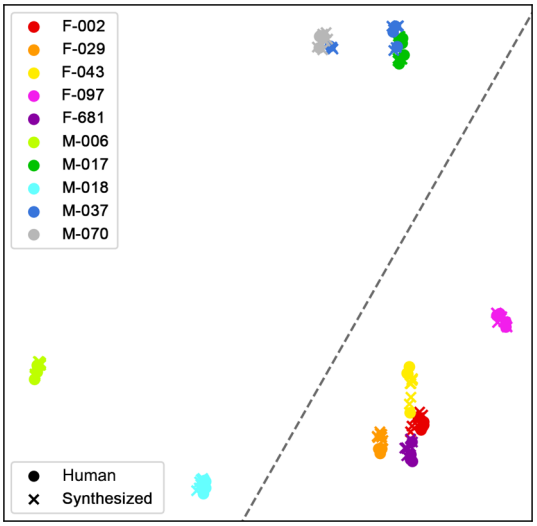

圖 5. 不同語者聲音分群的語者相似度分析（左上爲男性語者，右下爲女性語者）

合成器模型，皆取得了比 MockingBird 更高的分數。而如表 5 中所示，在不同聲碼器之間，又以 WaveRNN 取得 3.8 分較高。

從表 4 的結果可發現，用注音作爲文字嵌入符號的想法是可行的，故我們進一步以梅爾頻譜圖來觀察中英混雜模型 Code-Switching 的合成效果是否也有所提升。我們使用 MockingBird、CV8-75K、Bopomofo 與 Code-Switching 四個模型生成一段中英混雜語句「我明天要報五篇 Paper」後得到了語音的梅爾頻譜圖（見圖 6）。仔細觀察其中「Paper」部分的頻譜，因爲 MockingBird 與 CV8-75K 使了用拼音，故合成英文字時容易造成映射與中文拼音衝突，因此不論是用什麼語料集訓練，採用拼音方案，都較容易導致頻譜顯示模糊。反觀 Code-Switching 分離了

[7] http://www.aclclp.org.tw/use_mat_c.php#ner_edu
[8] https://datashare.ed.ac.uk/handle/10283/3443
[9] https://openslr.org/62
[10] https://www.openslr.org/93

Synthesizer Model	MOS
MockingBird + WaveRNN	2.43±0.19
CV8-75K + WaveRNN	3.80±0.16
Chinese-Hybrid + WaveRNN	2.82±0.17
Bopomofo + WaveRNN	4.13±0.16

表 4. 以不同模型合成之語音的平均意見分數

Vocoder Model	MOS
CV8-75K + WaveRNN	3.80±0.16
CV8-75K + HiFi-GAN	3.77±0.16
CV8-75K + Fre-GAN	3.32±0.16
CV8-75K + Griffin-Lim	2.92±0.17

表 5. 以不同聲碼器合成之語音的平均意見分數

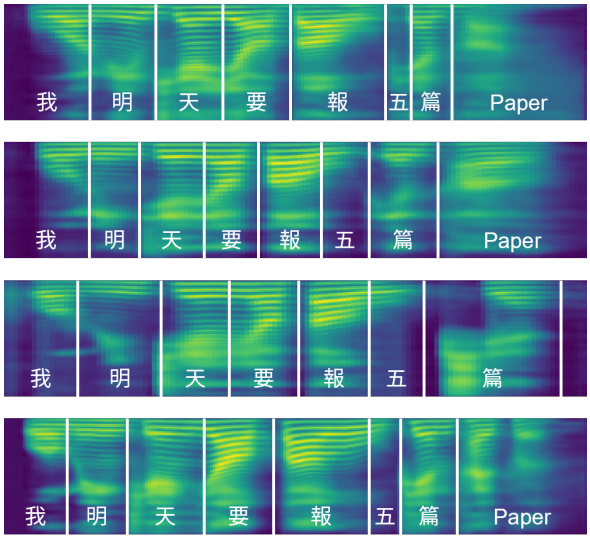

圖 6. 以不同模型合成之語音的梅爾頻譜圖（由上而下分別爲 MockingBird、CV8-75K、Bopomofo、Code-Switching 模型）

中文與英文的字元嵌入，使其可以讀出當中夾雜的英文字，這也反映出了較爲清晰的頻譜。另外，因爲 Bopomofo 的字元嵌入不含英文字母，故合成時系統會直接剪掉英文的部分。

在 SemiTrue 整體合成品質部分，我們使用前美國總統唐納·川普（Donald Trump）的演講影片作爲輸入，合成一段英文語句及一段中文語句，並比較其語音品質、相似程度與嘴型同步品質，結果呈現於表 6 中。透過比較原始影片、英文語句及中文語句合成影片之平均意見分數差異，可以發現中英語之說話影像的語音與嘴型同步品質，皆達到了一定水準，受測者對這些影像之嘴型同步品質所給出的分數與原始影片相當接近，惟中文語音之相似度較爲差強人意，因與原始影片的語言不同，僅取得受測者認爲「普通」的評價。

Case	Criteria	MOS
Original Video	Audio Quality	4.39±0.15
	Voice Similarity	4.35±0.16
	Lip-Sync Quality	4.23±0.17
English	Audio Quality	3.97±0.16
	Voice Similarity	3.90±0.16
	Lip-Sync Quality	4.24±0.15
Mandarin	Audio Quality	3.47±0.16
	Voice Similarity	2.91±0.19
	Lip-Sync Quality	3.79±0.17

表 6. 系統整體品質的平均意見分數

Case	Inference Time
Audio only	8.31 s
Audio + Image	13.30 s
Audio + Video	22.10 s

表 7. 不同合成方案的推論時間

另外在評估系統運作效能上，我們針對僅進行語音複製、語音複製配合以圖片合成影像、語音複製配合以影片合成影像，這三種情形在搭載 RTX2060 圖形處理器的設備上，以合成器模型 CV8-75K，並搭配合成品質最佳，但生成時間最長的聲碼器 WaveRNN，生成語句「任何人有三百萬美金都能參加慈善撲克大賽」，所消耗的時間進行統計，結果顯示於表 7。從表中可見，上述三種情況的生成時間呈現遞增，語音複製配合以影片合成影像，共用了 22.1 秒進行推論，但若僅進行語音複製則可在 8.31 秒完成，從以上的數據可看出我們的系統，在推論時間上有良好的表現。

5 結論

在本研究中提出了一個基於零樣本學習的 SemiTrue 多語者影音合成系統，其中包含了兩個主要改進。首先，我們利用臺灣口音語料集來訓練語音合成模型，且經匿名問卷調查後，得到了較高之平均意見分數，故使用我們的模型合成出的語音，確實更能令受測者感覺到自然與流暢。除此之外，我們以注音符號來表示合成器中的字元嵌入，從而在中英語句的英文部分，獲得了更清晰的梅爾頻譜圖，使系統具備初步合成中英混雜語句的能力。本研究仍有許多值得改進的地方，未來我們會朝向提升混雜語句能力的方向研究，最終期望在單一模型中，合成出自然且逼真的雙語語音。

Acknowledgments

特別感謝國家實驗研究院資安卓越中心規劃建置計畫、台灣人工智慧學校，提供運算資源，協助本研究順利進行。

References

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

An-Chieh Cheng and Chia-Ping Chen. 2019. 即時中文語音合成系統 (Real-time Mandarin speech synthesis system). In *Proceedings of the 31st Conference on Computational Linguistics and Speech Processing (ROCLING 2019)*, pages 256–265, New Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yi-Hsiang Hung, Yi-Chin Huang, and Guang-Feng Deng. 2019. 應用文脈分析於中英夾雜語音合成系統 (Linguistic analysis for English/Mandarin speech synthesis system). In *Proceedings of the 31st Conference on Computational Linguistics and Speech Processing (ROCLING 2019)*, pages 368–377, New Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Corentin Jemine. 2019. Real-time voice cloning. Master's thesis, University of Liége, Liége, Belgium.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee. 2021. Fre-GAN: Adversarial frequency-consistent audio synthesis. In *Proc. Interspeech 2021*, pages 2197–2201.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Beáta Lőrincz, Adriana Stan, and Mircea Giurgiu. 2021. Speaker verification-derived loss and data augmentation for DNN-based multispeaker speech synthesis. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 26–30. IEEE.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Paarth Neekhara, Jason Li, and Boris Ginsburg. 2021. Adapting TTS models for new speakers using transfer learning.

Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19).

Recommendation P.800.2. 2016. Mean opinion score interpretation and reporting. Technical Report P.800.2 E 40885, ITU-T, Geneva Switzerland.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C V Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 484–492, New York, NY, USA. Association for Computing Machinery.

K R Prajwal, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. ACM.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-wook Kim. 2022. Talking face generation with multilingual TTS. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21425–21430.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Sheng-Yao Wang and Yi-Chin Huang. 2021. Incorporating speaker embedding and post-filter network for improving speaker similarity of personalized speech synthesis system. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 326–332, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yih-Wen Wang and Chia-Ping Chen. 2020. Real-time single-speaker Taiwanese-accented Mandarin speech synthesis system. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 87–101, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech 2017*, pages 4006–4010.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5180–5189. PMLR.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning.

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. 2020. End-to-end code-switching TTS with cross-lingual language model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7614–7618. IEEE.