國立台北大學資訊工程學系專題報告

死而復聲-人工智慧語音聊天

Voice Chat with Artificial Intelligence

專題組員: 朱昕頤 王立宣 林柏翔 石楚城

專題編號: PRJ-NTPUCSIE-112-005

執行時間: 2023 年 9 月 至 2024 年 6 月

1. 摘要

從古至今，若是想念已故的人們，我們只能利用觀落陰等靈媒或者做夢來和已故的親人或好友交流，但隨著 AI 技術日新月異，為了解決人們的思親之愁，我們結合了聲音克隆模型、文本生成以及語音辨識這三個技術，開發了一個 AI 語音聊天程式。主要作用是與一個由我們開發的虛擬人物進行語音聊天，其中虛擬人物的聲音和說話方式可以是世界上任何人。從而達到模仿已故親人或好友，讓還在世的人能透過此網站體驗親朋好友還在世的時光。

2. 簡介

我們開發了一個 AI 語音聊天的應用程式，其中應用了聲音克隆、文本生成以及語音辨識等工具。聲音克隆技術使得使用者能夠透過模擬他人的聲音與對話，提升了聊天的真實感與情感連結。文本生成則能夠根據對話內容生成自然流暢的對話，增強了對話的自然性與深度。而語音辨識技術則有助於提升對話的準確性與效率，使得系統能夠更快速地理解使用者的需求與意圖。本研究將透過對這些工具的應用與效能分析，探討其在 AI 語音聊天中的潛在應用價值與挑戰。

3. 專題進行方式

使用的工具： 我們利用 VITS、Chatgpt 以及 Web Speech Api 的 SpeechRecognition module 這三個工具進行此專題。

- 聲音克隆：

我們使用的是一個開源在 github 上名為 VITS Fast Fine-tuning 的專案。由於此專案是使用微調一個預訓練模型的方式，透過將聲音特徵加入原本預先訓練好的語音模型，不需要完全從頭訓練，所以需要的聲音樣本

數相對也不需要那麼多。過程中我們只需要準備好訓練資料，也就是目標聲音的語音檔，無須高階顯卡或電腦，會使用 Google Colab 進行線上訓練，我們只要將訓練好的模型下載下來使用即可。根據作者所述，二到十秒的短音檔只需要最少十句，最好 20 句以上，或是有一個三分鐘以上的長音檔就可以達到不錯的效果。

- 風格文本生成：

我們先簡單介紹文本生成的原理，它會將依照一段文本，去回應出另一段符合內容與邏輯的文本，其中文本回應的風格與特色是我們自己可以決定的。我們使用的是 OpenAI 所推出之生成式大型語言模型-Chatgpt 作為基礎。要讓語言模型的回應具有風格，我們以人工萃取特徵的方式從收集到的資料中提取出所模仿之人的「說話風格」如口頭禪、語氣、回答長短等，再利用「提示工程」的方式將這些特徵轉換成「系統提示詞」，這個提示詞作為我們對語言模型的預設輸入，在使用者每個輸入的前綴中都會帶有此提示詞，讓模型先看到這個提示詞，轉換自己的輸出樣式，再以此樣式對使用者所拋出來的對話生成帶有說話風格的文字-風格文本，作為回應。

以下是我們目標收集的特徵：

1. 個性特徵
2. 口頭禪
3. 慣用語
4. 興趣
5. 情感表達
6. 回答問題的風格
7. 常見問答
8. 背景（經歷）
9. 個人理念
10. 給人的形象

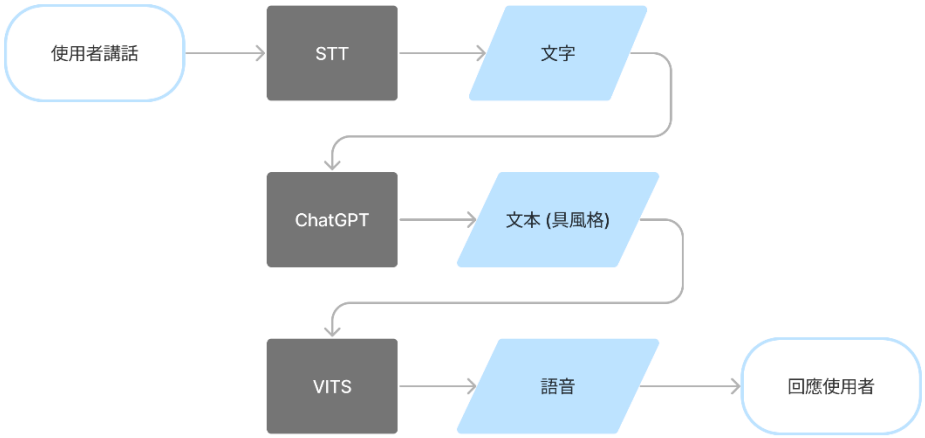
這些資料可以透過被模仿者的家屬描述、他寫過的文章、說過的話進行歸納。

- 語音辨識(STT)：

語音辨識是利用 AI 技術將語音轉換成文字，我們使用的是 SpeechRecognition，一個網頁上就有提供的 API，可以直接調用打開麥克風以收聽使用者的聲音，並且轉換出文字。

➢ 進行方式：

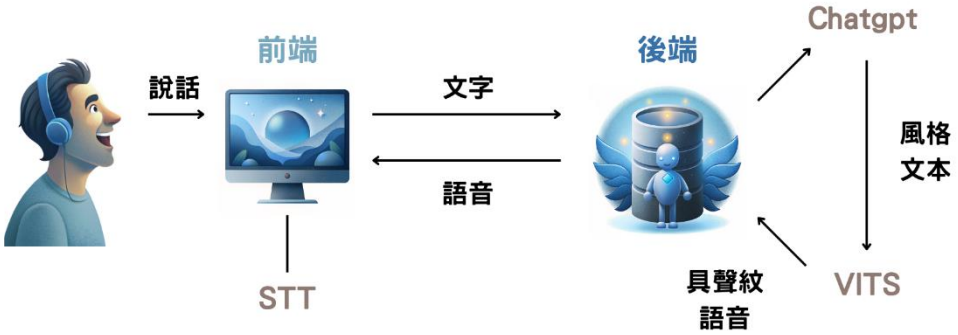
我們先準備三段要模仿的人說話的聲音檔，每一段聲音檔大約五分鐘左右。接著將聲音檔放到 VITS 聲音克隆模型，它會先將聲音檔進行降噪，降噪完之後就會開始對聲音檔進行訓練。之後再結合 ChatGPT 進行風格文本生成，我們把想要模仿的人之說話風格與特色輸入進去，接下來使用者輸入一段想要溝通的對話，ChatGPT 會將所生成之對話丟入 VITS 模型即會得到一段經過聲音克隆的聲音檔。最終得出來的成果就是一個 AI 語音聊天應用程式。



上圖為最終的流程圖，首先我們使用者先說一段話，STT 語音辨識會先幫我們把語音轉成文字，接下來將文字放入訓練好的 ChatGPT 進行文本生成，把生成出來的文本丟入 VITS，即會得到一段語音，此段語音就會是使用者得到的回覆。

整合：

要將我們所準備好的工具整合起來，我們採用了下面的架構：



1、 前端

i. 收聽使用者的聲音，並將使用者所說的話利用 SpeechRecognition Web API 轉換成文字，傳送給後端。

ii. 接收後端回傳的句子和語音，語音會和句子搭配，播放語音的同時將對應的句子打出來。

2、 後端

i. 接收前端所傳過來的文字，並利用 OpenAI 所提供的 API 將它和準備好的「系統提示詞」合併輸入給 Chatgpt。

ii. Chatgpt 生成文字回應，將文字回應依照標點符號分割，每段輸入到已經訓練好並啟動的 VITs 微調預訓練模型做推理生成出帶有「聲紋特徵」的語音，將句子和語音一併傳送給前端。

4. 主要成果與評估

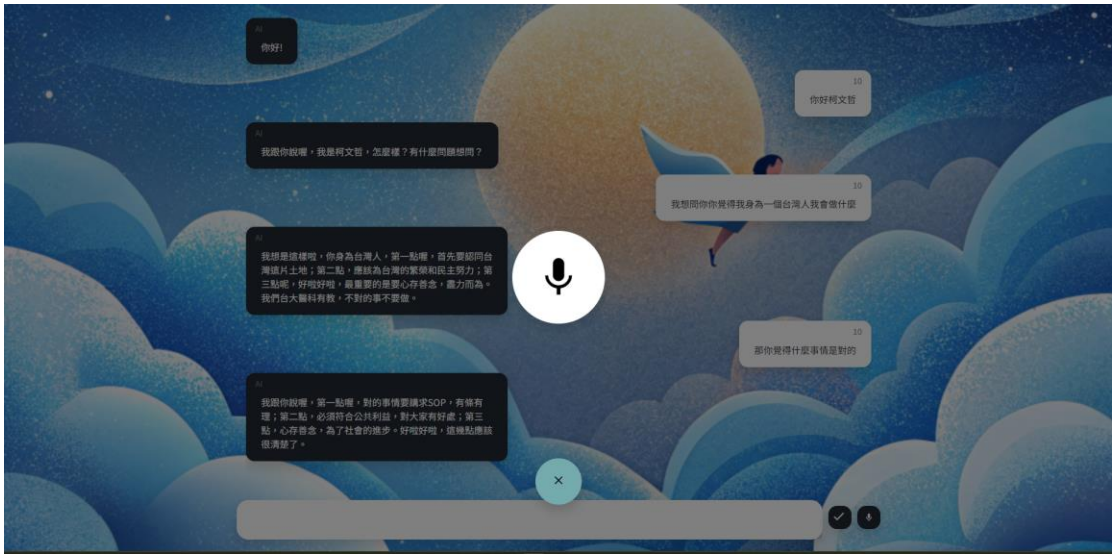
❖ 成果

1. 主頁面

2. 對話頁面

當 AI 開始回答，語音會自動在背景撥放，並且句子如字幕一句一句展示給使用者。

對話中，會有開啟語音的提示畫面（如下圖），告訴使用者現在正在聆聽，並在每次 AI 回答完，都會自動等待使用者的語音輸入，如果使用者不講話，就會關閉語音的提示畫面，需要使用者再次點及右下角麥克風按鈕打開或是使用文字輸入。



❖ 評估

1、 回應速度評估

最一開始我們使用的是 gpt-4 作為我們的語言模型，在最初的系統架構上面，平均的回覆延遲會來到 5 至 10 秒，顯得不像在聊天，我們做了兩項更近：第一是更換模型，由於 OpenAI 進一步的推出了 gpt-

4o 的架構，文本生成速度得到非常大的提升。第二如整合後端所提，我們應用語言模型的「流式」輸出，將已經組合好的句子輸入到 VITs 微調預訓練模型裡面推理出語音，並將這個語音和句子配對送至前端呈現給使用者，而在使用者聆聽回覆的同時，後端繼續將後續生成好的文字和聲音配對送到前端，便可有效降低回覆延遲至 2 至 4 秒。

2、 文本風格評估

文本風格得好壞主要取決於收集特徵的多寡，基本上將上列之特徵收集到，就會有不錯的成果。但是畢竟我們文本生成是使用 ChatGPT，而它主要的用途是回答問題並非聊天，所以有時候文本生成出來的風格還是會有一點像是在問答的感覺，這一點是我們之後必須要努力克服的。

3、 語音聲紋評估

由於我們使用的 VITS 是經過預訓練，後續所需要的聲音資料並不需要太多，以目前所餵進去的量來說，4 段 5 分鐘且比較清晰的音檔，就可以達到很好的效果。然而 VITS 只能生成出相同聲紋的聲音，不能獲得具有明顯情緒的語調聲音。所以在說話的高低起伏這一部分是我們必須要突破的。另一個點是我們為了降低回覆延遲，因此是拆成各個句子生成語音，然而語音和語音的銜接並沒有做潤滑，使得使用者聆聽上面會有較明顯的不自然斷句。

5. 結語與展望

我們這次的專題使用了許多工具，像是聲音克隆模型、文本生成技術、語音辨識技術。從而製作得到一個 AI 語音聊天的應用程式。然而我們最一開始目標是想要做視訊聊天程式而非語音聊天，但是由於時間的不足再加上大部分的人臉合成模型都是屬於付費使用，再加上使用上也有不小的難度，導致我們最終沒有將此相技術加進此專題之中。

未來我們將繼續嘗試加入人臉合成技術，讓我們的 AI 聊天程式的功能更加全面。

6. 銘謝

非常感謝指導教授給予我們的意見以及帶領，讓我們能夠克服一路上所遇到的種種困難。也感謝所有組員的努力，讓我們能順利完成專題。

7. 參考文獻

1. Kristiawan Nugroho, Edy Winarno, “Spoofing Detection of Fake Speech Using Deep Neural Network Algorithm,” International Seminar on Application for Technology of Information and Communication, 2022
2. Jingyi Li, Qin Xu, Michel Kadoch, “A Study Of Voiceprint Recognition Technology Based on Deep Learning,” International Wireless Communications and Mobile Computing, 2022
3. Li Zhao, Feifan Chen, “Research on Voice Cloning with a Few Samples,” International Conference on Computer Network, Electronic and Automation, 2020
4. Jiwon Seong, WooKey Lee, Suan Lee, “Multilingual Speech Synthesis for Voice Cloning,” IEEE International Conference on Big Data and Smart Computing, 2021