

國立台北大學資訊工程學系專題報告

智慧虛擬分身

Intelligent Virtual Avatar

專題組員:林家緯、李岱芸、曾泯薰、張毓旻

專題編號:PRJ-NTPUCSIE-112-008

執行期間:112 年 9 月 至 113 年 6 月

1. 摘要

如今社會科技已然十分卓越，然而人們的心理狀況卻少見關懷。即使是醫療發達的現在，也常見親友的逝去；忙碌的日常生活，使人們難以擁有閒心談天說地。許多話語只能藏在心中，或是難以啟齒，或是再無機會，遺憾並因此而生。科技日新月異，人心卻日漸疏離。為解決此類問題，使人們得以無所顧忌地傾訴，創造一個可以進行正常語音對話與動作的虛擬人物，藉此代替真實的人來與使用者對話。此計畫使用 OpenAI 的 ChatGPT 模型生成對話，結合 Ready Player Me 的 3D 人物模型，使人物能夠自然地與使用者談論各種話題，判斷對話的情緒從而做出相應的肢體反應，增加其真實度與親密感，豐富人們的生活體驗並提升使用者的情感參與感。

2. 簡介

本計畫旨在利用 OpenAI Whisper 平台的語音辨識功能作為輸入，同時利用 ChatGPT 的自然語言處理能力，實現智慧對話系統。透過這個系統，使用者能以自然的語音進行對話，而 ChatGPT 將根據輸入的語音內容分辨其情緒再給予適切的回應，最終以

Coqui 所提供的 Text To Speech 功能來將 ChatGPT 所給予的回應以自訂聲紋的形式撥放。進一步地，我們將此智慧對話系統與 Unity 遊戲開發平台相結合，利用 Ready Player Me 所提供的 3D Avatar 功能，為使用者提供身臨其境的互動體驗。根據 ChatGPT 的回應內容，我們將設計不同的肢體動畫和臉部表情，使得 Avatar 能夠栩栩如生地模擬出對話情境，提高對話的情感連貫性和真實感。

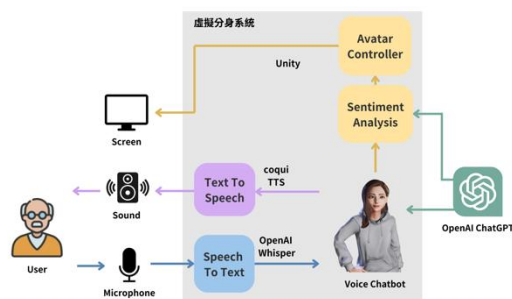
在現今社會中，AI 已成為各個領域的焦點，驅使人們蜂擁而至投入開發。令人惋惜的是，因為使用者並非專家，對於生成式 AI 的了解與應用略顯不足，往往只作查詢資料使用，而忽略其廣大的潛力所在。這一計畫不僅推進人機互動技術的發展，更為虛擬現實、遊戲開發等領域帶來全新的應用可能性。透過融合語音輸入、自然語言處理和虛擬仿真技術，打造出更加人性化、更具互動性的智慧系統，進而豐富人們的生活體驗並提升使用者的情感參與感。本計畫可分為兩大主題進行研究：

- (一) 人機語音對話：如何讓使用者與 ChatGPT 即時語音溝通？
- (二) 虛擬分身：如何讓使用者所創建之人物模型做出應有情緒之

臉部及肢體表現？

3. 專題進行方式

在此計畫中，我們利用 OpenAI 所提供的 Whisper 功能實現 Speech to Text，並使用 OpenAI 的 Unity API 與 ChatGPT 進行溝通。ChatGPT 回覆後，再藉由 Coqui TTS 完成 Text to Speech，並且根據所檢測到之情緒反應，經由 Unity 渲染出相應的肢體與臉部動畫。圖一為本計畫之系統架構圖，大致展現了系統實現的流程與步驟。以下將解釋如何決定使用之工具，並介紹各項流程的實作方法。



圖一. 系統架構圖

(一) ChatGPT 技術

ChatGPT 技術是由 OpenAI 開發的一種基於深度學習的自然語言處理技術。它通過學習大量的文本數據，生成具有上下文連貫性的自然語言文本。

過去的研究已經證實了 ChatGPT 在對話生成、語言理解等方面的有效性和性能。對於情感偵測表現出了大型語言模型如 GPT 比舊有的機器學習方法有更好的結果^[1]，但過於模稜兩可的 Prompt 會使 ChatGPT 產生不理想的判定結果^[2]，所以如何正確決定各種情緒及對應的 Prompt 將成為一項

重要的挑戰。

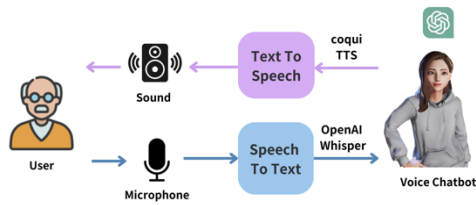
(二) Coqui TTS

人機互動技術在現代科技應用中扮演著關鍵角色。過去的研究探討了語音輸入、自然語言處理以及虛擬仿真技術的應用，而這些技術的發展同時潛移默化人們對於科技產品的接受度和使用體驗。在語音辨識與合成中，人們更傾向與具有真實人聲的系統互動^[3]，並認為此類系統更加自然且易於理解。

有鑑於此，我們採用 Coqui 所提供的 XTTS 功能，透過 10 秒左右的人聲樣本，結合預訓練模型，推理出近似原型的聲紋。Coqui TTS 為開源的語音合成技術，利用高性能深度學習模型將文本快速轉換成高品質語音，並支持多達 16 種語言，同時擁有完整的 API，方便我們有效率地操作，完美符合計畫中所需。相較於傳統 TTS 系統，Coqui TTS 擁有更自然的語言連貫度與穩定的品質，使我們得以運用此工具達成目標。

(三) Chatbot

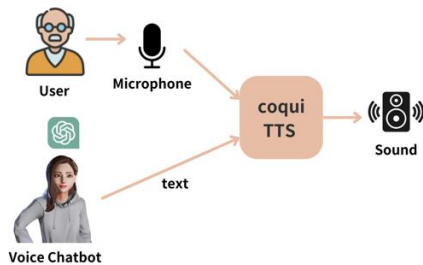
我們選擇 OpenAI 的 ChatGPT 作為本系統的聊天機器人，利用 Prompt 事先設定 ChatGPT 所扮演的角色以及情緒的判定^[2]，讓 ChatGPT 得以相對應角色回覆，並將情緒判定的結果傳至 Unity，讓模型得展現對應情緒之肢體及表情回應。下圖為 Chatbot 在系統架構中所影響之部分：



圖二. Chatbot 與 ChatGPT

(四) 音訊處理

為使虛擬分身更加逼真，我們透過 Coqui TTS 所提供之 XTTS 功能，經由分身原型聲音的樣本來合成聲紋。而後，虛擬分身根據此聲紋，將 ChatGPT 回應之文字轉為語音輸出。



圖三. Coqui TTS

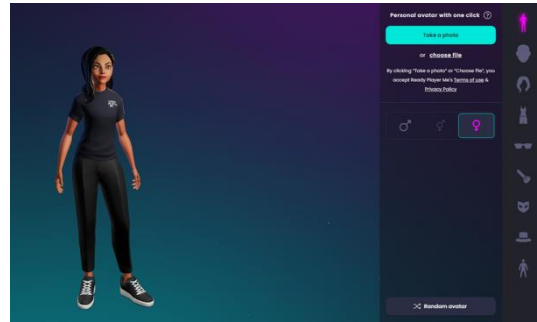
(五) 分身模型

在模型的選擇上，比起相對逼真的人物，象徵性的人物更能讓人取得信任^[4]，因此我們選擇使用 Ready Player Me 作為建立模型的系統。圖二為系統默認之四位分身模型，均為專題小組成員為原型：



圖四. 系統默認分身模型

除了默認角色，我們也讓他們使用這可以自訂虛擬分身，透過自訂分身的過程，增強與分身之間的親近感。若是難以見到親友的人們，亦可使用自定義功能，經由 webcam 拍照，讓虛擬分身產生近似於親友的面貌。圖五為創建虛擬分身之實際頁面：



圖五.

Ready Player Me 的 Avatar Creator

(六) 情緒反應

現實中的溝通除去言語，多數都是倚靠肢體動作及面部表情來區分情緒。為保證虛擬分身足夠真實，情緒反應必然是重要的一環。

我們利用 Ready Player Me 內建的肢體動畫庫及 Adobe 的 Mixamo 所提供的肢體動作，再經由調整模型臉部上的 BlendShape 值來控制面部表情展現，用以建立完整的情緒反應，協助使用者更好的帶入情境。下圖分別為 Ready Player Me 內建動畫庫及 Mixamo 提供之動作，圖八則為 BlendShape 之展示畫面：



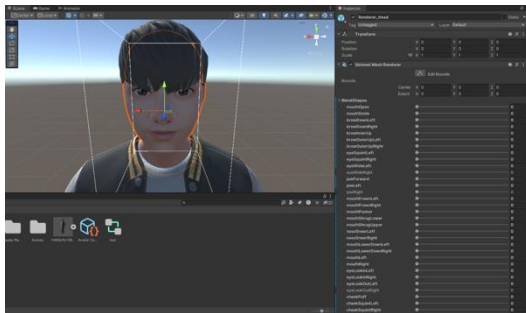
圖六.

Ready Player Me 所提供的動畫庫



圖七.

Adobe 的 Mixamo 所提供之肢體動作



圖八.

在 Unity 中調整 BlendShape 之值

4. 主要成果與評估

- 選擇自定義人物模型或是選取現有之四種模型。
- 使用語音輸入並且選取語音輸入之時間，以避免等待過於冗長。
- 選擇希望使用之 ChatGPT 模型。
- 將語音輸入內容顯示於畫面上。
- 即時回應使用者所輸入之內容。

- 人物將以語音及文字之形式回應，並表現出使用者之相應情緒的臉部及肢體動作。
- 現有人物之語音回覆為分身之原型之聲紋。
- 人物模型可任意上下移動及放大縮小。

5. 結語與展望

此計畫整合了 OpenAI 的 ChatGPT、Whisper 語音辨識、Coqui TTS 語音合成以及 Ready Player Me 的 3D 人物模型，實現了一個能與使用者自然地對話的虛擬分身系統。此系統不僅能夠辨識並回應使用者的情緒，更能依對應分身輸出本人聲音，降低虛擬分身的機器感，使虛擬分身能夠展現更真實的情感交流，拉近與使用者之間的距離。

在專題進行過程中，我們面臨了許多技術挑戰。如同前面所述，在語音辨識和合成方面，人們傾向真實人聲而非文字回覆^[3]，因此，將 ChatGPT 的回覆轉換為語音成為此計畫的關鍵步驟之一。此外，我們期望能讓使用者將語音與分身模型聯繫，藉由真實人類的語音樣本，合成近似於此人的聲音。透過不斷地查找和測試，最終達成此目標。

未來我們將進行更進一步地調整與擴充：

- ◆ 群組功能：增加使用者與虛擬分身的人數，達到同時多人對話。
- ◆ 情緒切割：豐富情緒、以及臉部/肢體動畫的種類，精準表達使用者的感受。
- ◆ 個性化表現：新增更多元的

功能，例如：更換聊天室背景、文字框主題等，讓使用者可自由選擇更為舒適的對話環境。

- ◆ 自訂聲紋：加入使用者即時自訂聲紋功能，彷彿與親友對話。

6. 銘謝

感謝指導教授的悉心教導與指正，以及實驗室中學長姐的幫助與解答。

7. 參考文獻

- [1] S. Shaikh, S. M. Daudpota, S. Y. Yayilgan and S. Sindhu, "Exploring the potential of large-language models (LLMs) for student feedback sentiment analysis," 2023 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2023, pp. 214-219, doi: 10.1109/FIT60620.2023.00047.
- [2] O. Banimelhem and W. Amayreh, "The Performance of ChatGPT in Emotion Classification," 2023 14th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2023, pp. 1-4, doi: 10.1109/ICICS60529.2023.10330544.
- [3] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human-Agent Interaction: A Survey. ACM Comput. Surv. 54, 4, Article 81 (May 2022), 43 pages.
- [4] F. Limano, "Realistic or Iconic 3D Animation (Adaptation Study with Theory Uncanny Valley)," 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC), Bandung, Indonesia, 2019, pp. 36-41, doi: 10.1109/ICSECC.2019.89071

