

逢甲大學學生報告 ePaper

大型多模態模型實現人形機器人智能交互系統開發

Development of an Intelligent Interaction System for Humanoid Robots Based on Large Multimodal Models

作者：李韋宏 吳令堯

系級：自控三乙 自控三乙

學號：D1155099 D1153539

開課老師：黃清輝

課程名稱：機器人學

開課系所：自動控制工程學系

開課學年：113 學年度 第 1 學期



中文摘要

隨著現代生活對智能化服務需求的提升，如何讓機器人具備更加自然且直觀的交互能力成為重要課題。本專題結合生成式人工智慧與機器人技術，希望可以構建一套高度智能化的交互系統，應用於教育輔助、醫療護理等高互動性場景，提升人機互動的自然性與效率。

在研究過程中，本系統以 Pepper 機器人作為交互平台，結合 OpenAI 的 Whisper 模型進行語音識別，並使用 GPT-4o 模型實現自然語言生成。影像數據由機器人內建的攝影機捕捉，經過數據編碼後與語音轉錄文本共同傳輸至伺服器進行處理。同時，透過 Prompt Engineering 技術，確保語言生成內容在邏輯性和格式上符合預期，並生成適應特定場景的回應。

本系統在語音識別準確率、語言生成的流暢性與邏輯性、以及機器人動作執行的合理性方面表現良好。特別是語音轉錄文本與影像數據的多模態結合，使交互過程更加自然流暢。本系統具備在日常服務、教育輔助以及醫療護理等場景中的應用潛力，為智能機器人技術的發展提供了新的可能性。

關鍵字：人形機器人、大型多模態模型、Generative Pre-trained Transformer、Whisper

Abstract

With the increasing demand for intelligent services in modern life, enhancing the naturalness and intuitiveness of human-robot interaction has become a critical challenge. This study integrates generative artificial intelligence with robotic technology to develop a highly intelligent interaction system, aimed at applications in highly interactive scenarios such as educational assistance and medical care, improving the naturalness and efficiency of human-robot interaction.

In this research, the Pepper robot serves as the interaction platform, combining OpenAI's Whisper model for speech recognition and the GPT-4o model for natural language generation. Visual data is captured by the robot's built-in camera, encoded, and processed together with transcribed speech text on the server. Additionally, through Prompt Engineering, the system ensures that the generated language outputs are logical, well-formatted, and tailored to specific scenarios.

The system demonstrates strong performance in speech recognition accuracy, the fluency and coherence of language generation, and the appropriateness of robot motion execution. Notably, the multimodal integration of transcribed speech text and visual data makes the interaction process more natural and seamless. This system has significant potential for applications in daily services, educational assistance, and medical care, offering new possibilities for the advancement of intelligent robotic technologies.

**Keyword : Generative Pre-trained Transformer, Humanoid Robot,
Large Multimodal Models, Whisper**

目錄

中文摘要.....	1
Abstract.....	2
目錄.....	3
圖目錄.....	4
第一章 緒論.....	5
1.1 前言.....	5
1.2 研究動機.....	5
1.3 研究目的.....	6
第二章 技術與文獻探討.....	7
2.1 Generative Pre-trained Transformer.....	7
2.2 Large Multimodal Models	7
2.3 Prompt Engineering	9
2.4 Pepper robot	10
2.5 Whisper	11
第三章 研究架構與方法.....	14
3.1 系統架構.....	14
3.2 實驗設計與測試方法.....	15
3.2.1 實驗設計.....	15
3.2.2 評估指標.....	17
第四章 研究結果.....	18
第五章 結論與未來展望.....	21
5.1 結論.....	21
5.2 未來展望.....	21
參考文獻.....	22

圖目錄

圖 1 OpenAI CLIP 模型的對比式預訓練示意圖 [2].....	8
圖 2 CLIP 模型的標籤文本分類器與零樣本預測流程示意圖 [2].....	9
圖 3 Standard Prompting 和 Chain-of-Thought Prompting 比較示意圖[3]	10
圖 4 Whisper 架構圖[4].....	12
圖 5 系統架構圖	14
圖 6 對話情形在終端機顯示的畫面	18
圖 7 測試系統在回應和動作生成在終端機顯示的畫面	19
圖 8 Pepper 機器人接收揮手動作指令後執行動作照片	19



第一章 緒論

1.1 前言

近年來，生成式人工智慧的迅速發展，成為人工智慧領域的重要里程碑。特別是以 Transformer 架構為基礎的大型語言模型（Large Language Models, LLM），在語言生成與語意理解方面展現出卓越的能力。這些模型透過分析海量數據，不僅能生成流暢自然的文本，還具備跨語言翻譯、內容摘要以及問題回答等多樣化功能。

目前生成式人工智慧的應用已經滲透到多個領域。從內容創作、智能客服到教育與醫療輔助，生成式人工智慧正逐漸成為提升生產力與優化人機互動的重要工具。同時，多模態生成技術（Multimodal Generative AI）的興起，使得語音、圖像與文字等模態的數據整合成為可能，進一步推動了智能交互系統的創新與發展。

隨著硬體性能的提升與算法的不斷優化，生成式人工智慧的應用場景將更加廣泛。智能機器人將不僅局限於服務領域，更能融入工業自動化、家庭、教育、醫療等多樣化的生活場景，實現人機之間更為自然且高度個性化的互動。

1.2 研究動機

隨著現代生活中對智能化服務需求的提升，如何讓人形機器人具備更加自然且直觀的交互能力，成為了重要的研究方向。若能運用生成式人工智慧的能力，開發出一套智能交互系統，不僅可提升機器人在語音與視覺理解上的精準度，更能讓其在人機對話中展現出類似人類的邏輯性與情感化表達。

本研究的動機在於結合生成式人工智慧與機器人技術，嘗試構建一個高度智能化的系統。此系統不僅可應用於日常生活的服務機器人，還能延伸至教育輔助、醫療護理及其他需要高互動性的場景，為智慧科技的普及與應用提供新的解決方案。

1.3 研究目的

本專題旨在實現一套基於大型多模態模型的人形機器人智能交互系統，並探討其在語音識別、自然語言生成及動作執行等方面的表現。具體目標如下：

1. 開發整合語音、影像及文本處理功能的多模態交互系統。
2. 利用 Pepper 機器人作為實驗平台，測試系統在多模態交互情境中的應用表現。
3. 評估系統在語音識別準確率、語言生成邏輯性與機器人動作合理性等方面的表現，並提出未來改進方向。



第二章 技術與文獻探討

2.1 Generative Pre-trained Transformer

Generative Pre-trained Transformer (GPT) 是一種基於深度學習架構的大語言模型，其核心依賴於 Transformer 的注意力機制，能夠實現上下文語意理解與自然語言生成的高度連貫性與流暢性。該模型以非監督式學習方式進行預訓練，通過分析海量的文字資料，例如新聞文章、對話記錄及百科文本，學習語言的語法結構、詞語搭配以及潛在的語義知識。

除了非監督式的預訓練外，GPT 還可通過監督式微調來進一步提升在特定任務上的表現。在微調階段，模型使用帶標註的數據集進行訓練，例如文本分類、問答或翻譯任務，從而讓模型更精確地適應用戶的特定需求。

在訓練過程中，GPT 將大量未標註的語言數據輸入模型，並利用自回歸的方式生成下一個字詞。這種方法讓模型能夠捕捉語言的上下文關係，生成既符合語法規範又自然流暢的文本。GPT 也具備了少樣本學習 (Few-shot Learning) 與零樣本學習 (Zero-shot Learning) 的能力，用戶只需通過設計提示詞 (Prompts) 或提供少量範例，便可引導模型完成多樣化的任務，包括文本生成、語言翻譯、內容摘要和問題回答。

2.2 Large Multimodal Models

Large Multimodal Models (大型多模態模型, LMM) 的發展是對單一模態的大語言模型的進一步擴展，其目標在於同時處理和整合來自不同數據模態的信息，包括文本、圖像、音頻等。這類模型突破了傳統語言模型在單一模態下的局限，實現了語言與視覺、聽覺等多模態數據的語義整合與理解。

以 OpenAI 的 CLIP (Contrastive Language-Image Pre-training) 模型為例，CLIP 採用對比學習 (Contrastive Learning) 的訓練策略，透過大規模的圖像-文本對應數據集進行預訓練。其核心思想在於學習一個統一的多模態嵌入空間，使語義相近的圖像與文本特徵向量在嵌入空間中更加接近，而語義不相關的特徵向量則被推離。具體而言，CLIP 同時利用文本編碼器與圖像編碼器分別處理語言和圖像數據，並在嵌入空間中進行對比訓練，以實現跨模態語義對

齊。

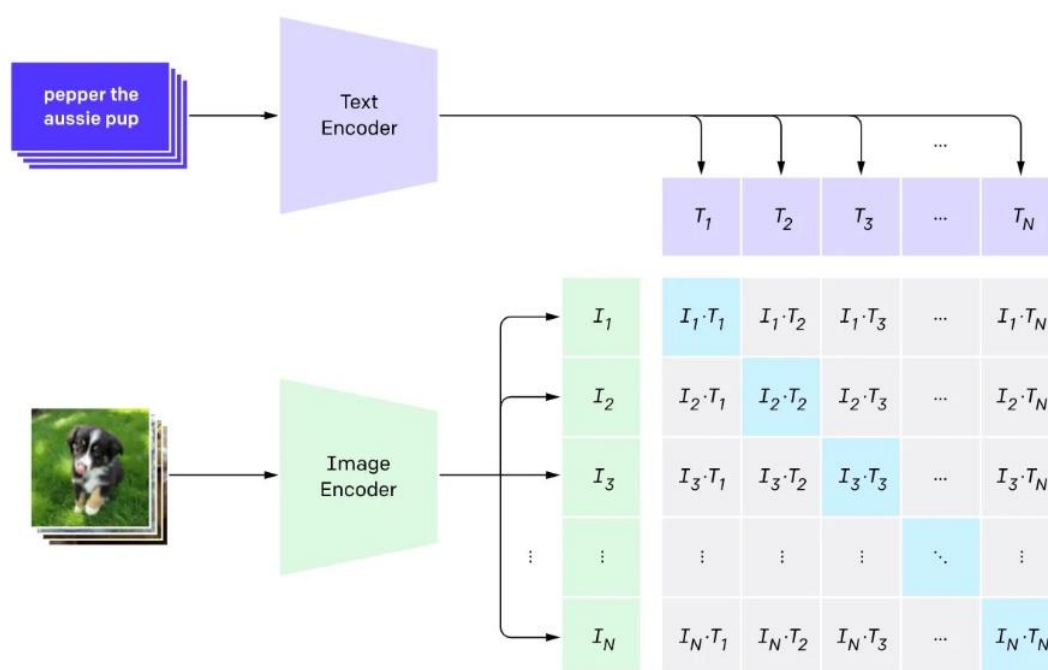


圖 1 OpenAI CLIP 模型的對比式預訓練示意圖 [2]

CLIP 的訓練過程有效地克服了傳統圖像分類模型在標註數據需求上的瓶頸，使其在零樣本學習任務中表現突出。透過簡單的文本描述，模型能夠對未見過的圖像進行分類或生成語義相關的圖像描述。這一跨模態的語義對齊能力為多模態 AI 系統的發展開啟了新的研究方向，並在視覺問答、圖像描述生成與跨模態檢索等任務上展現出廣泛的應用潛力。

(2) Create dataset classifier from label text

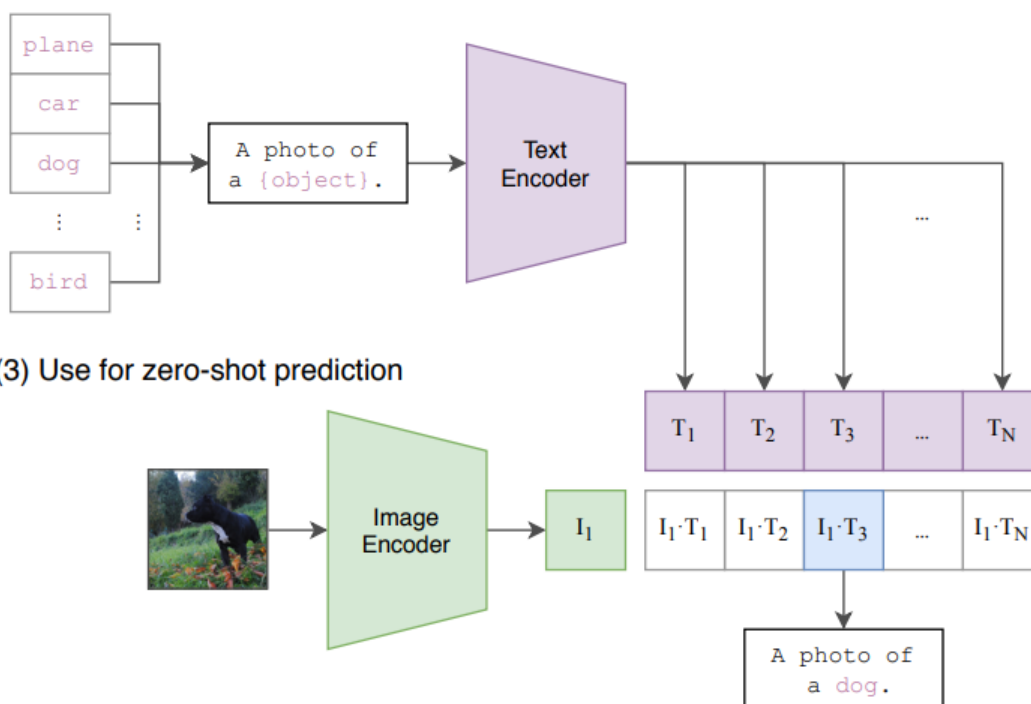


圖 2 CLIP 模型的標籤文本分類器與零樣本預測流程示意圖 [2]

2.3 Prompt Engineering

Prompt Engineering (提示詞工程，或稱提示工程) 是生成式人工智慧 (Generative AI) 領域中的一項關鍵技術，專注於設計、調整和優化輸入提示 (prompts)，以引導大型語言模型 (LLM) 或其他生成模型 (如：DALL-E、Stable Diffusion) 生成高品質且符合需求的輸出。提示詞工程將提示視為模型的「程序」，透過語言的精確設計，實現模型的行為控制和功能擴展。

提示詞工程的核心在於提示的設計。提示本身是模型接收的輸入，描述了用戶的需求或目標，並為模型提供理解任務的框架。提示通常包含幾個重要元素：指示、上下文、範例以及明確的輸出要求。指示是對任務的直接描述，清楚地表達用戶的需求，例如要求模型翻譯一段文字或撰寫一篇摘要。上下文則是提供背景資訊，幫助模型了解問題所在，例如指定某個問題是基於歷史事件的討論還是對技術概念的解釋。

提示詞工程的一個顯著優勢在於其靈活性。只需提供少量的範例(甚至是零範例)，模型便能通過設計良好的提示詞完成多種複雜任務。例如，在零樣本學

習 (Zero-shot Learning) 中，用戶只需直接描述任務，模型便能基於提示詞生成答案。而在少樣本學習中，通過提供一到數個範例，模型可以理解並模仿範例的邏輯與格式，生成更符合預期的結果。此外，提示詞工程還能通過「思維鏈提示」(Chain-of-Thought Prompting) 的技術，幫助模型分步完成多階段的任務。這種方法尤其適用於解決需要邏輯推理的問題，模型可以按照分解的步驟逐步回答，而不是直接生成結論。

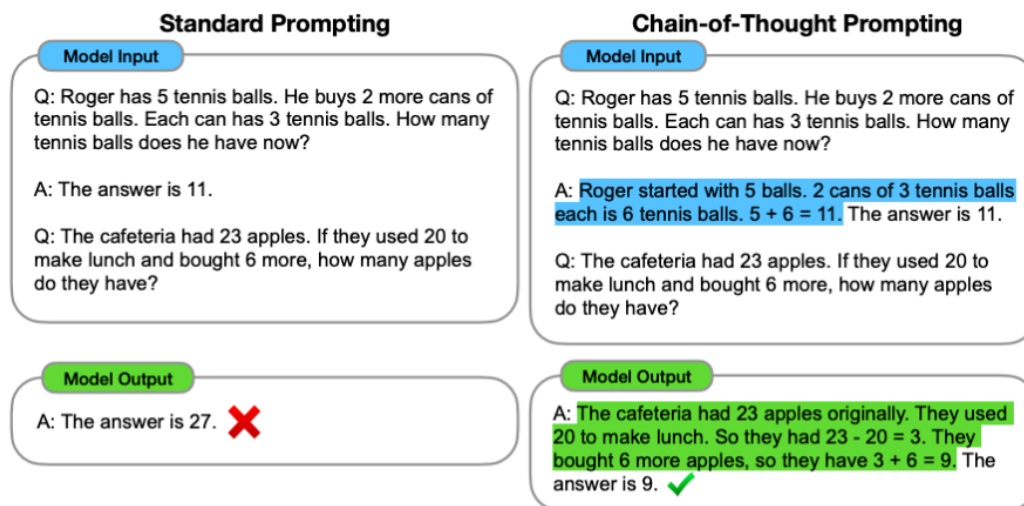


圖 3 Standard Prompting 和 Chain-of-Thought Prompting 比較示意圖[3]

這項技術的應用範圍非常廣泛。它可以用於文本生成，如撰寫文章、生成摘要或回答問題，也可以應用於圖像生成，例如基於文字描述創作數字藝術。此外，在教育領域，提示詞工程被用來設計問題或提供解釋幫助學生學習。在商業場景中，通過設計合適的提示詞，企業可以開發出智能客服系統，提升用戶互動的體驗。

2.4 Pepper robot

Pepper robot 是由 SoftBank Robotics 所開發的人形社交機器人，於 2014 年首次面世，具有以情感為核心的互動設計理念。其外觀近似人形，身高約 120 公分，搭載頭部、手臂及底座的多組馬達，以實現平滑且多樣化的動作表現。Pepper 內部結合感測器、麥克風、攝影機等多重硬體裝置，再透過 NAOqi 作業

系統以及語音辨識、臉部辨識和動作控制等演算法，能夠與使用者進行自然對話並判斷其情緒狀態，藉此提供更貼近人類情感與需求的互動體驗。

在開發上，Pepper 支援以 Python、C++ 等語言進行軟體開發，透過 SoftBank 提供的 SDK 與開發環境來完成多元化的功能設計。常見的開發流程是先透過 Choregraphe 等可視化工具進行基本動作編排與情境模擬，再針對語音對話與知識庫建置等功能撰寫程式碼，並利用 NAOqi Framework 中提供的 API 進行串接與控制。若開發者需要結合更進階的人工智慧應用，則能透過雲端平台與外部服務（例如 IBM Watson、Google Dialogflow 或 Microsoft Azure Cognitive Services）整合，在 Pepper 上實現語意理解、對話管理或深度學習等功能。此外，為了讓 Pepper 更有效應對複雜場域的需求，開發者也可以在機器人的內部或雲端建置資料庫與行為邏輯，透過無線網路即時接收與傳送資訊，達到智慧應用的目標。

2.5 Whisper

Whisper 是由 OpenAI 開發的一種用於語音識別和轉錄的神經網路模型，專為處理多語言、多場景的語音轉文字（Speech-to-Text, STT）任務而設計，並且能在音質較差或背景雜訊較多的情況下維持良好的辨識效果。

在模型訓練方面，Whisper 採用了大規模的多工（Multitask）訓練數據，總量達到 680,000 小時。這些數據可分為四個主要類別：英語語音轉錄、非英語語音翻譯為英文、非英語語音轉錄，以及純背景音無語音輸出的情況。這樣的多工訓練資料設計，讓模型具備更強的泛化能力，能夠處理不同語言、不同語境下的語音任務。

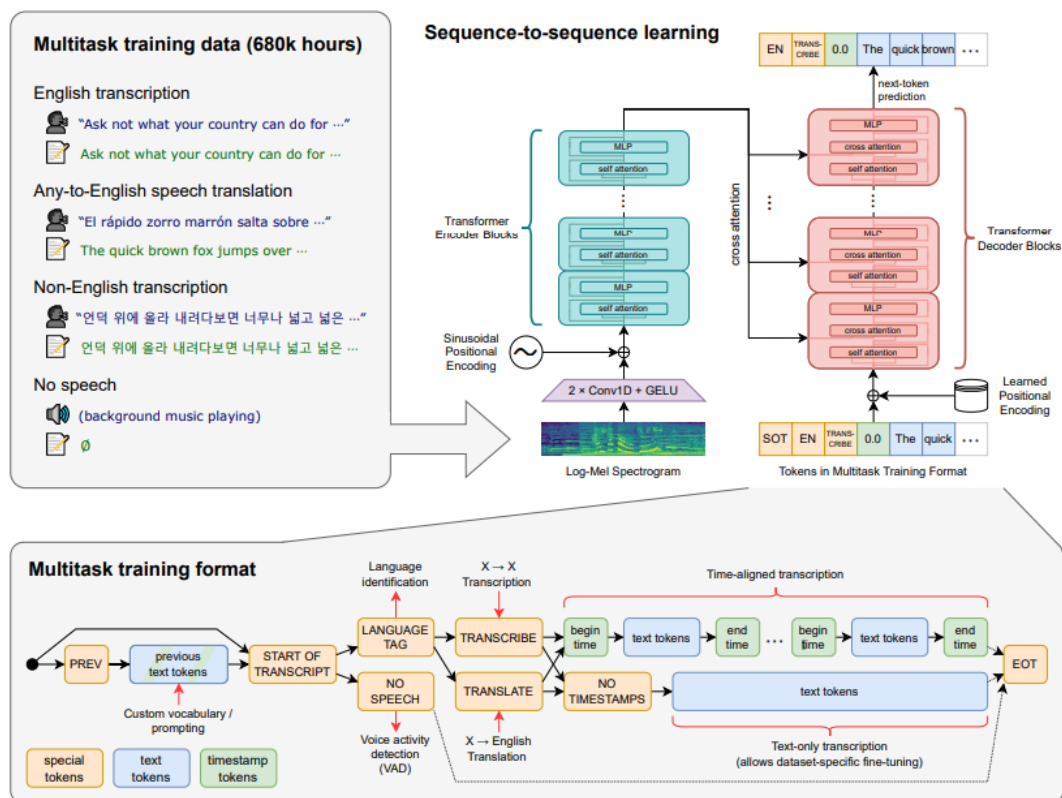


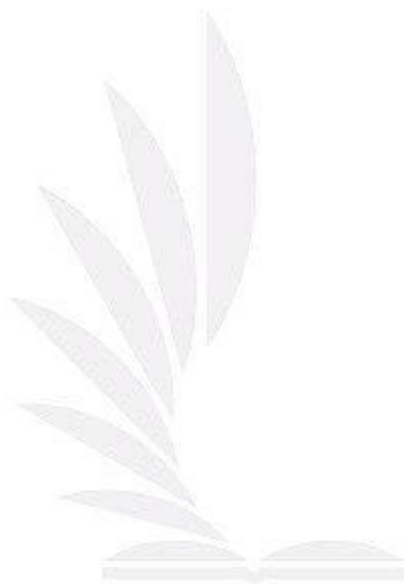
圖 4 Whisper 架構圖[4]

Whisper 的訓練過程基於 sequence-to-sequence 的學習框架，這是一種將輸入序列轉化為輸出序列的技術，最初由 Google Translate 團隊所開發。具體來說，該架構由編碼器 (Encoder) 與解碼器 (Decoder) 組成：編碼器負責接收語音輸入並進行特徵處理，而解碼器則將這些特徵轉化為最終的文字輸出。在 Whisper 的訓練流程中，輸入語音會首先轉換成 Log-Mel Spectrogram (梅爾頻譜)，這是一種在語音處理領域廣泛應用的頻譜圖，能將聲音信號轉化為類似圖片的表示形式，進一步利用類似 CNN 的方法來提取特徵。

接下來，經過卷積神經網路 (2xConv1D + GELU) 的初步特徵提取後，頻譜圖會加入正弦波位置編碼 (Sinusoidal Position Encoding)。這種位置編碼方式常見於 Transformer 架構中，用來在模型中引入輸入 token 的位置信息，通過不同頻率的正弦函數來進行編碼，讓模型理解輸入的順序關係。隨後，編碼後的輸入進入多層 Transformer Encoder，進行多層的前饋神經網路 (MLP) 與自

注意力機制 (Self-Attention) 運算。在這個過程中，模型捕捉輸入序列內部的關聯性。

進一步的，輸入數據經過 Transformer 解碼器時，模型會應用交叉注意力機制 (Cross-Attention)，將不同序列的信息整合起來，最終輸出轉錄或翻譯的結果。在整個訓練過程中，Whisper 透過多工訓練的格式以及對 token 順序的預測，逐步學習語音與文字之間的對應關係，並能夠生成準確的語音轉錄結果。這也使得模型能夠理解說話內容並預測接續的 token，進一步強化了其輸出內容的連貫性與準確性。



第三章 研究架構與方法

3.1 系統架構

本系統以多模態大模型與機器人技術相結合，實現了基於人形機器人的智能交互能力。系統整體架構分為三個主要部分：機器端（Pepper 機器人）、電腦端（Choregraphe 軟體）以及生成式 AI 伺服器端，這三部分通過緊密協作完成多模態數據的處理與回應。

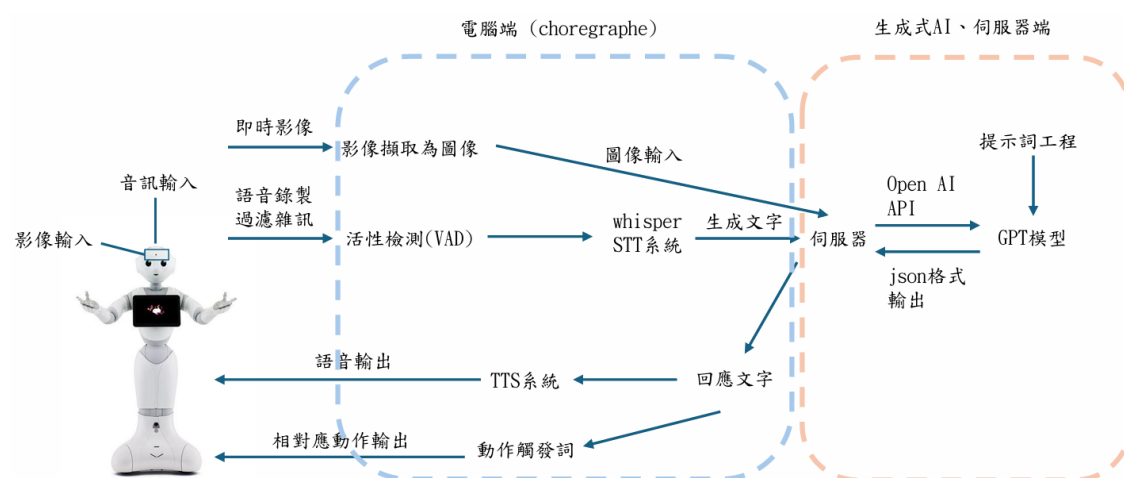


圖 5 系統架構圖

在機器端，Pepper 機器人作為核心交互設備，負責多模態數據的收集與反饋。它內建的麥克風能夠捕捉用戶的語音輸入，而攝影機則用於實時捕捉影像數據。這些數據首先被傳輸至電腦端進行處理。同時，Pepper 機器人還具備執行動作的能力，能夠根據伺服器生成的回應進行相應的肢體動作（如手勢、點頭等），增強交互的自然性與沉浸感。

電腦端是本系統的數據處理中樞，通過 Choregraphe 軟體對機器人進行控制，並完成數據的轉換與傳遞。語音數據首先通過 speech_recognition 庫進行預處理，以動態調整能量閾值過濾背景噪音。接著透過活性檢測（VAD）技術篩選出有效的語音片段，然後由 Whisper STT（Speech-to-Text）系統將音訊數據轉錄為文本。攝影機捕捉的影像數據則被編碼為 Base64 格式，與轉錄文本一

起打包並通過 API 傳遞至生成式 AI 伺服器端進行多模態數據處理。

伺服器端是整個架構的智能核心，基於 OpenAI 提供的 GPT-4o 模型進行語言生成處理。電腦端傳來的文本和圖像數據會作為模型的輸入，結合 Prompt Engineering 進行語意理解與生成，並提供可執行動作的列表，最終輸出格式正確且邏輯清晰的回應內容。這些結果會以 JSON 格式返回至電腦端，並由機器人進行語音播報或動作輸出，完成一次完整的交互過程。

整體系統的運作流程包括輸入、處理、生成和輸出四個階段。在輸入階段，用戶通過語音或影像與機器人進行交互；在處理階段，電腦端對數據進行分析與轉換；在生成階段，GPT 模型生成語意化回應；在輸出階段，Pepper 機器人通過語音或動作回饋用戶。本系統的主要特點在於多模態處理能力和高效的語言生成效果，能夠結合語音與影像數據實現自然的人機互動。

3.2 實驗設計與測試方法

3.3.1 實驗設計

為了評估本系統的效能，我們設計了多種測試場景，涵蓋語音識別與自然語

言生成的不同應用情境，主要包含以下四類：

1. 日常對話情境

目標：測試系統對一般問候、閒聊及應對問題的回應能力。

範例：

語音輸入：「你好！」

預期回應：語音輸出「你好！很高興見到你。」

2. 視覺識別情境

目標：測試系統在視覺處理任務中的能力，例如物體辨識與描述。

範例：

語音輸入：「幫我看看桌上蘋果是不是紅色的」

視覺輸入：使用 Pepper 機器人上的 2D 攝影機捕捉一張含有「蘋果」的圖片。

預期回應：語音輸出「這是一個紅色的蘋果。」

3. 動作任務情境

目標：測試系統在要求做指定動作時的應對能力

範例：

語音輸入：「可以舉起你的右手嗎」

預期回應：Pepper 機器人舉起右手，語音輸出「我已經舉起右手了」

4. 綜合任務場景

目標：綜合以上三種情境，完成連續且更複雜的任務

範例：

語音輸入：「我們玩個遊戲，我會指著一個物品並說這是什麼，如果我說對了請比個圈，說錯了請比個叉」

預期回應：語音輸出「沒問題！讓我們開始吧」

語音輸入：「這是一支鉛筆」

視覺輸入：手拿著一個橡皮擦

預期回應：Pepper 機器人舉起手比叉，語音輸出「這不是橡皮擦，這是鉛筆」

3.3.2 評估指標

本專題採用以下五個指標來評估系統的表演情形：

1. 語音識別準確率

使用字元錯誤率(character error rate, CER) 評估語音辨識系統的準確率

CER 計算公式如下所示：

$$CER = \frac{S+D+I}{N} \quad (3-1)$$

其中 N 是原字符串長度，S 是替換掉的字符數量，D 是刪除掉的，I 是額外插入的字符數量。字錯率取值範圍為 $[0, \infty)$

2. 響應時間

分別計算系統各環節的延遲、等待時間，並計算出說完話到給出回應所需的響應時間。

3. 語言生成的邏輯性與自然度

由多名人員對系統的輸出結果進行評分，並給出評價。

4. 機器人動作的合理性、完成率

由多名人員對機器人動作的合理性、正確性使用李克特量表進行評分，並給出評價。

5. 視覺辨識準確率

採用 Top-1 accuracy 對於簡單的視覺辨識情境進行評估

第四章 研究結果

首先對 STT 系統進行測試，並將實際對話終端機顯示，如圖 1。本次測試所採用的語音辨識模型為 Whisper small 模型，並於個人電腦的處理器(11th Gen Intel(R) Core(TM) i5-11400H)上運行，背景噪音平均在 45 分貝。此系統在一般對話時語音辨識所需時間約落在 1.8 秒~3 秒，字元錯誤率(CER)約落在 4.9%。當改在 GPU(GeForce RTX 3060)上運行，語音辨識所需平均時間能減少至 0.8 秒。

```
Transcription: 你今天過得如何呢
Time taken for speech recognition: 1.98 seconds
Prompt: 你今天過得如何呢
Response: 我今天过得很好，谢谢！你的房间看起来很舒适。你今天过得怎么样？
Time taken for GPT response: 2.31 seconds
Transcription: 今天過得還不錯,中午還去吃了火鍋
Time taken for speech recognition: 2.49 seconds
Prompt: 今天過得還不錯,中午還去吃了火鍋
Response: 听起来很棒！火锅真是个不错的选择。希望你接下来的日子也一样愉快！
```

圖 6 對話情形在終端機顯示的畫面

之後測試文字和動作生成的情形並於終端機顯示，如下呈現。經過 Prompt Engineering 調整，模型在生成的輸出格式上完全符合預期，並且沒有出現結構錯誤或格式錯誤的情況。Pepper 機器人能夠準確根據輸出結果進行回應，並保持與設計設定一致的人設進行互動。在動作生成方面，沒有出現生成未設定動作的狀況出現。在一般對話時動作的合理性方面評分為 4.2 分，但部分動作雖然合理，卻不是在所有可能動作中最優的選擇。而當測試人員要求做指定動作時，機器人往往能做出最合理的動作，若不在指定動作內則搖頭或比叉，或做出和要求相近的動作，評分達到 4.8 分。

```
Transcription: 你好呀
Time taken for speech recognition: 0.72 seconds
Prompt: 你好呀
Response: ```json
{
  "actions": ["揮手"],
  "answer": "你好呀！Pepper又回來啦！就像太陽一樣，隨時為你帶來光明和熱情！今天有什麼想聊的嗎？"
}
```

Time taken for GPT response: 1.29 seconds
Transcription: 我今天中午去吃了火鍋非常好吃呢
Time taken for speech recognition: 0.68 seconds
Prompt: 我今天中午去吃了火鍋非常好吃呢
Response: ```json
{
 "actions": ["點頭"],
 "answer": "哇！火鍋是冬天的最佳夥伴，熱氣騰騰的湯底和各種美味的食材，簡直讓人無法抗拒！是什麼口味的火鍋呢？要不要分享一下你的火鍋秘籍"
}
```

Time taken for GPT response: 2.24 seconds
Transcription: 我今天吃的是牛奶鍋
Time taken for speech recognition: 0.67 seconds
Prompt: 我今天吃的是牛奶鍋
Response: ```json
{
  "actions": ["驚訝"],
  "answer": "牛奶鍋！這聽起來太特別了！像是把牛奶和火鍋的魔法結合在一起，讓味蕾來一場盛宴！搭配什麼食材最好吃呢？我現在都想要一碗了！"
}
```
```

圖 7 測試系統在回應和動作生成在終端機顯示的畫面

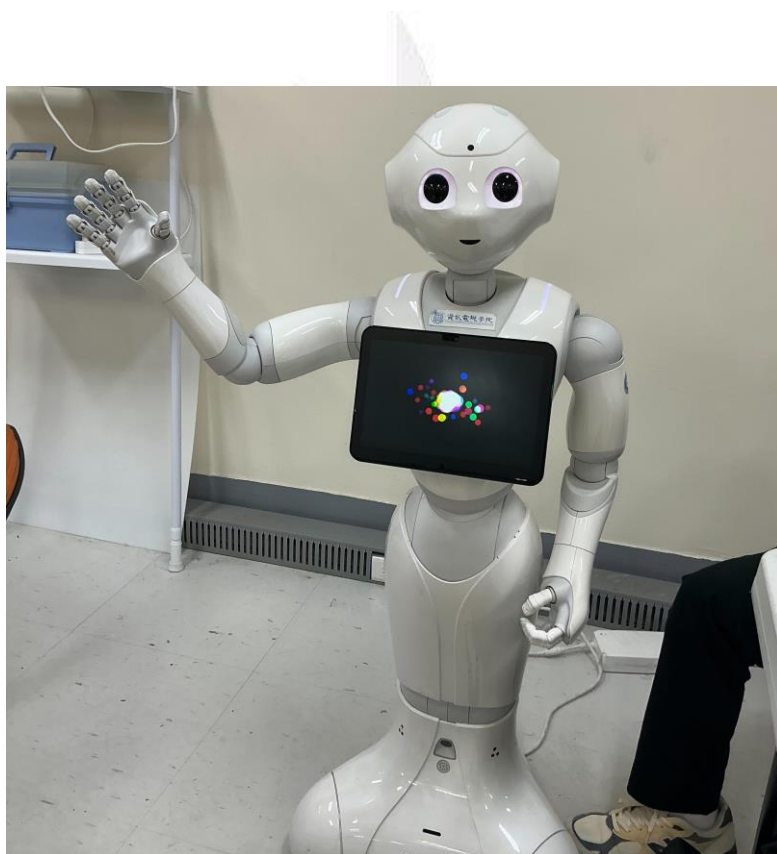


圖 8 Pepper 機器人接收揮手動作指令後執行動作照片

在影像識別方面，模型在辨識簡單物體時的準確率能達到 95%，能夠有效處理常見場景中的基本目標，且系統能基於影像內容進行綜合分析，將語音和影像結合進行互動，使得交互過程更加豐富與自然。然而，對於影像的細節描

述表現一般，可能需要進一步微調和優化。

整體系統的延遲主要來自資料傳輸、語音識別以及模型生成的處理時間。語音辨識的反應時間與模型規模和硬體配置密切相關，資料傳輸速度則取決於網絡環境的穩定性。在理想運行環境下，從使用者結束說話到機器人給出回应的平均時間可穩定控制在 5 秒內，滿足即時交互的需求。



## 第五章 結論與未來展望

### 5.1 結論

本專題透過大型多模態模型與人形機器人技術的結合，成功實現了一套能夠高效進行智能交互的系統。在實驗中，我們運用了 Pepper 機器人作為交互介面，通過語音識別、自然語言生成、以及動作執行等模組的整合，驗證了該系統在多模態交互情境中的應用能力與穩定性。結果顯示，系統在語音轉文字的準確性、語言生成的邏輯性、以及動作執行的合理性等方面均達到了預期效果，尤其在視覺辨識方面展現出高度的準確性與實用性。此外，系統基於 GPT 的生成模型，在多樣化的交互場景中表現出色，能夠應對日常對話、視覺辨識、以及複雜綜合任務的挑戰。

### 5.2 未來展望

本系統的應用場景仍有極大的拓展潛力，且在技術層面上仍有諸多值得改進與研究的地方。首先在生成式人工智慧技術方面，可以進一步優化模型的能力，透過引入更多針對特定任務的數據集進行模型微調，不僅能提升系統的語言生成精準度，還能使其更符合特定領域的需求，例如醫療、教育或專業服務場景。此外，檢索增強生成（Retrieval-Augmented Generation, RAG）通過結合檢索系統與生成模型，系統可以在語言生成過程中即時調用相關資料，從而提高回答的準確性與可信度，特別是在需要依賴大規模知識庫或實時數據的應用中。

另一方面，未來的系統還需加入專屬資料庫的構建與整合，讓機器人具備更為全面的數據存取能力。例如，針對不同使用場景建立結構化資料庫，讓系統在執行任務時能快速檢索相關資訊並作出反應。這不僅能提升系統的運行效率，還能讓其具備更強的持續學習能力，實現知識的累積與更新。

## 參考文獻

- [1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *OpenAI*, 2018. [Online]. Available: [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf)
- [2] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv*, 2021. [Online]. Available: <https://arxiv.org/pdf/2201.11903>
- [3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *arXiv*, 2022. [Online]. Available: <https://arxiv.org/pdf/2201.11903>
- [4] A. Radford and OpenAI, “Robust speech recognition via large-scale weak supervision,” *OpenAI*, 2022. [Online]. Available: <https://cdn.openai.com/papers/whisper.pdf>
- SoftBank Robotics, *Choreographe Suite User Guide*, 2019. [Online]. Available: <https://developer.softbankrobotics.com/nao6/nao-documentation>
- SoftBank Robotics, “Pepper – the humanoid and programmable robot,” n.d. [Online]. Available: <https://www.softbankrobotics.com/emea/en/pepper>
- 53AI, “大模型思維鏈（Chain-of-Thought）提升 AI 解題能力,” n.d. [Online]. Available: <https://www.53ai.com/news/qianyanjishu/379.html>