THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

ENGINEERING
DEPARTMENT OF ELECTRICAL,
COMPUTER, AND SOFTWARE ENGINEERING

2023 Part IV Research Project

Final Report

Investigating the Integration of Pepper Robot and ChatGPT: A Study on Enhancing User Experience and Engagement

Prepared by: Xiaohui Chen (cxia813)

Project number:	91
Project partner:	Katherine Luo
Main supervisor:	Mahal Nejati
Co-supervisor:	Trevor Gee

Declaration of Originality

This report is my own unaided work and was not copied from anyone or anywhere nor written in collaboration with any other person.

Author: Xiaohui Chen

9th October 2023

Abstract

Humanoid robots are designed to execute human interaction tasks. However, the current communication abilities of these robots, such as Pepper from SoftBank Robotics, do not meet human expectations. The appearance of Large Language Model (LLM) demonstrates a potential to address the communication limitation for humanoid robotics. This paper details the integration methodology of Pepper-GPT, the system integrating with the Whisper Automatic Speech Recognition (ASR) system, the gpt-3.5-turbo language model and the Pepper robot. The results of subsequent usability investigations prove that 88% of the 25 human participants thought the system did improve UX. It is evidenced that although some challenges are required to be solved, like improving the robot's facial tracking capabilities and reducing the long processing time to generate replies, users generally responded positively to the system, feeling like they were engaging in conversation with a real human.

Acknowledgement

First and foremost, I would like to express my sincere appreciation to my supervisor, Dr. Mahla Nejati, and co-supervisor, Dr. Trevor Gee, for their selfless support, guidance, and mentorship throughout this part 4 project. They have provided me with constructive feedback and professional suggestions that enhanced the quality of this research project.

I also want to thank my project partner, Katherine Luo, for her consistent and fantastic contribution to this project. We have worked collaboratively towards the project goals. Reflecting on our journey together, I am grateful for the shared experiences and the growth we've undergone as a team.

Contents

List of Figures

List of Equations

Acronyms

AI Artificial Intelligence. 8, 9, 13, 23

API Application Programming Interface. 6, 7, 10, 22, 23

ASR Automatic Speech Recognition. 1, 6–11, 14–19, 22, 23

HRI Human Robot Interaction. 5, 8, 16, 24

LLM Large Language Model. 1, 8

NLP Natural Language Processing. 6, 8

UX user experience. 1, 5–7, 11, 20–24

WER Word Error Rate. 3, 7, 10, 14, 15, 17–19, 23

1 Introduction

As technology rapidly advances, Human Robot Interaction (HRI) becomes increasingly prevalent. People show substantial enthusiasm and a growing desire to communicate and collaborate with robots more naturally and effectively seamlessly [1]. However, current technologies are still under expectation as humans cannot interact with robots effectively [2], preventing users from getting the complete benefit from these systems.

According to Hardson and Pyla et al.'s study, user experience (UX) is defined as the total effect of interacting with a system, including usability, usefulness, and emotions during and after, shaping the overall impact on a person [3]. It is crucial to ensure a good UX during the interaction between humans and robots, as this increases the acceptance rate of robots in human society and brings practical value to people's lives [4][5].

Pepper is a well-known humanoid social robot created by SoftBank Robotics. It is celebrated for its versatile interactive features, including speech recognition, gestures, pre-coded dialogue, etc. However, if the robot is expected to become more human-like and adaptable during human interactions, its current capabilities may fall short of expectations [6]. Existing studies have found that the Pepper robot still faces challenges when performing tasks even though its developers have added versatile functions to it. Considerable delays and errors still exist in the language processing of Pepper during responses, leading to a notable impact on participants and making users feel challenged to remain engaged during subsequent sessions [7].

The poor communication performance of the Pepper robot limits its interaction ability. Pepper relies on its built-in dialogue-based tool for communication. This tool allows developers to embed scripted response contents into the robot and trigger them by capturing keywords. Poorly designed reply scripts resulted in Pepper being unable to answer users' questions completely or leading to misunderstandings during queries [8].

Additionally, users encountered challenges as the Pepper robot struggled to understand speech content effectively, requiring users to make multiple attempts for the robot to understand [9]. The Pepper robot's integrated speech recognition API is limited to recognizing predefined phrases, making its capability to understand natural speech insufficient.

Utilizing a high-performance Natural Language Processing (NLP) model, such as ChatGPT, to generate responses or analyze instructions is a potential solution to address these problems. These systems can provide detailed responses according to the instructions in the prompt [10]. Given its exceptional performance, it is hypothesised that compared to the current dialogue tool of Pepper, users should be able to engage in more natural and contextually relevant conversations with ChatGPT API.

Furthermore, it is essential to improve the speech recognition abilities of the Pepper robot to ensure precise and complete transcription of the user's audio input. This enhancement can minimise the robot's misinterpreting of users, leading to better overall performance and promoting the UX.

This paper outlines the methodology of the integration system (Pepper-GPT), which embedded the Whisper Automatic Speech Recognition (ASR) and GPT API into the Pepper robot, and the results from subsequent usability investigations with human participants, evaluating the performance of the Pepper-GPT.

2 Related Work

Current studies usually evaluate the UX with humanoid robots via human trials. Participants interact with these robots in specific real-life scenarios, such as the healthcare facility [11], and answer questionnaires to gather their feedback, including their feelings, perceptions, attitudes, etc. [9], utilised for robot performance evaluation.

In addition, existing research is devoted to improving social robots, such as Pepper's speech recognition and natural language processing skills, to enhance their communication capabilities. Google Cloud Speech Recognition is one of the solutions used to enhance the speech recognition skill of the Pepper robot [6][12]. For instance, in a front desk application, the average processing time of Google Cloud Speech Recognition API was 1.031 and 0.887 seconds in noisy and quiet environments, respectively [6]. However, their study did not use the Word Error Rate (WER) to measure the accuracy of Automatic Speech Recognition (ASR) systems, leaving the gap for further research.

According to current comparisons of ASR systems' WER, three ASR models stand out: Google ASR, Google Cloud ASR and Whisper ASR. Powered by the Python speech recognition library, Google ASR demonstrated an average WER of 20.63% [13]. Google Cloud ASR is a cloud-based service with an average WER of 12.16% [14]. Whisper ASR was developed by OpenAI and performed best during Radford et al.'s comparison of different datasets in long-form transcription, ranging from 3.523% to 19.6% [15].

However, it is notable that there is a shortage of direct quantitative comparisons among these three ASR systems, especially with a wide range of accent tests. Even though Radford et al.'s study involves the comparison between the Whisper ASR model and other commercial

ASR models in long-form transcription, the companies' names are anonymous, and the datasets they used are not labelled with different accents [15]. Therefore, a comparison experiment of these three ASR models was designed to evaluate their performance and find the optimal solutions for the Pepper-GPT. The results confirm that the Whisper ASR model is far ahead of the other ASR systems' accuracy and processing time.

In addition to enhancing the speech recognition feature, the existing methods also try to improve the language processing ability of Pepper via GPT-3 [12], enabling the Pepper robot to make open conversations with various topics. However, OpenAI has published other state-of-the-art LLM, including GPT-3.5 and GPT-4.0, allowing more fluent and contextually relevant conversations between AI and users. GPT-3.5 is the extension of the GPT-3 model with experienced further enhancements in performance and efficiency [16]. Furthermore, compared to the GPT-4 model. The gpt-3.5-turbo model maintains competitive performance and balances efficiency and affordability [17]. This makes the gpt-3.5-turbo the optimal solution for the design of Pepper-GPT.

In short, integrating ASR and NLP models into the Pepper robot makes it more attractive and adaptable, providing valuable interaction and help to users. The design of this integration system brings considerable possibilities for further improvements of the robots' interaction abilities, which becomes a potential solution to solve the bottleneck of communication barriers in HRI fields.

3 Methodology

In this section, the methodology utilised in the design of Pepper-GPT is sketched (Fig. 1). The design framework comprises two central programs, namely Black Box and Pepper Controller. Furthermore, a data transmission mechanism between these two programs relies on a client-server model employing the TCP/IP protocol.

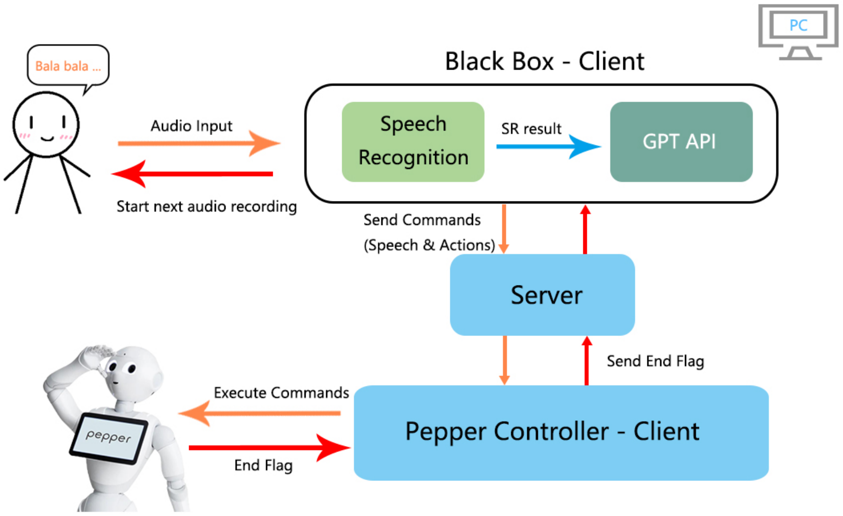


Figure 1: The Workflow of Pepper-GPT.

Black Box includes the Whisper ASR system and the gpt-3.5-turbo language model, facilitating responsive and contextual-relevant communication. Meanwhile, the Pepper Controller serves as the control hub for the execution of physical action and speech commands.

The Pepper-GPT addresses the limitation that AI models cannot physically interact with humans because they lack physical bodies. In Appendix A, additional details of the flow chart and pseudo code for modules in the Black Box are provided.

3.1 BlackBox

Black Box consists of two core modules: speech recognition and the GPT module. Audio recorded by the Black Box is transcribed to text by the speech recognition module. Subsequently, the GPT module analyses the text content and generates accurate action commands or relevant replies (speech commands) via the gpt-3.5-turbo model.

3.1.1 Speech Recognition Module

Compared to Pepper's built-in speech recognition API, the outstanding performance of the Whisper ASR system makes it a more appropriate selection to enhance the robot's speech recognition ability. The study of Radford et al. demonstrates that the Whisper Small model excels in balancing processing time, resource use, and transcription accuracy among all Whisper models [15]. Comparative analysis of model sizes, WER, and time consumption confirm that the comprehensive performance of the Whisper Small model exceeds both larger or smaller sizes of Whisper ASR models, positioning it as the optimal choice for Pepper-GPT's speech recognition module.

The flow chart in Appendix A presents the process in the speech recognition module. In addition to the Whisper Small model, the speech recognition module employs a microphone for audio recording and the Silero VAD model [18] for human voice detection. The microphone would activate when it detects sounds and stop when silence appears. As the Whisper ASR system would transcribe the empty audio into contents like "Thank you.", it is essential to use the Silero VAD model to detect whether the recorded audio contains a human voice to solve this problem.

After recording, the Whisper ASR system transcribes the audio file into text. If the Whisper ASR system makes successful transcription, the results are sent to the GPT module to analyse users' speech comprehensively and generate responses. However, there may be transcription failure under some circumstances, and the Whisper Small model returns an empty transcription. To continue the conversation, the system would proactively respond, sending a speech command to the Pepper robot, which makes the robot encourage users to rephrase their words. This feature prevents the users from facing the occasional technical problems and ensures smooth interactions between humans and robots.

3.1.2 GPT Module

The core of the design architecture relies on the gpt-3.5-turbo model, functioning as the engine to generate responses. The choice of the gpt-3.5-turbo model is driven by its exceptional comprehension and text generation abilities, enabling the creation of responses that mimic a natural human-like style. This capability aligns with the goal of generating highly diverse content in the Pepper-GPT design. The model's proficiency in comprehending user inputs and producing relevant, accurate, and captivating conversations highlights its crucial role in enhancing the overall UX.

The pseudo-code in Appendix A demonstrates the command generation process in the GPT module. After obtaining transcribed text from the speech recognition module, the GPT module initially analyses the user's dialogue content. Then, it will enter distinct processing modes based on the analysis results, referred to as the action and speech modes, respectively.

The main goal of the GPT module in the action mode is to transform users' input into action commands for the Pepper robot to carry out. The GPT-3.5 model detects keywords related to actions in the dialogue and converts them into explicit physical directives. The GPT-3.5 model only analyses the input content during the process, not generating contextual-relevant replies.

Conversely, users' input is intended to either start or continue a conversation with the Pepper-GPT during the speech mode. In this mode, the GPT-3.5 model resumes its role as a chat robot, receiving the dialogue content and generating contextually relevant responses to maintain engaging and meaningful user interactions.

It's important to highlight that the GPT-3.5 model's analysis may sometimes lead to misunderstandings, such as interpreting a physical action command as a speech command or vice versa. Therefore, a double-check function has been implemented as a solution. This function reviews the GPT-3.5 model's response and intervenes if it detects that the user's input should actually enter the action mode. In such cases, it extracts relevant actions and transitions to action mode to generate precise action commands.

This complicated user input dealing process ensures accurate command generation, including physical action commands and speech commands. The design of the GPT module places emphasis on prioritizing user intent and the robot's ability to respond appropriately, eliminating the limitation of dialogue content topics and thereby enhancing the interaction between the user and the robot.

3.2 Pepper Controller

The main function of the Pepper Controller is to act as the central command hub for the Pepper robot, carrying out various physical actions or speaking out the replies content matching with the instructions and data transmitted from the BlackBox. By bridging the gap between the virtual world of AI models and the physical world of the robot, the Pepper Controller ensures that the Pepper robot can appropriately respond to user interactions.

The Naoqi ALAnimatedSpeech agent handles both action and speech commands. In the case of speech, the Pepper Controller transforms the reply text from the Black Box into spoken words. Additionally, the initialisation of the Pepper robot employs the ALSpeakingMovement agent and is set in the contextual mode, enabling the robot to do specific gestures according to the keywords in reply content. This configuration enhances the Pepper robot's dynamism and charm to users.

Furthermore, a dedicated dataset is created for action commands, containing all the predefined physical action agents for the Pepper robot to execute. The Pepper Controller locates the corresponding agents based on the received physical action commands and executes them accordingly. Recognize that the time required to transcribe the speech input and generate the response from the GPT model is not negligible. Therefore, when the Pepper robot still does not receive a reply from the GPT module after a certain period of time, it should perform a transition animation, such as a thinking action, to maintain a smooth interaction process.

3.3 Data Transmission

The Pepper-GPT system uses distinct Python versions for the Black Box (Python 3) and Pepper Controller (Python 2.7) as the Naoqi Python SDK [19], the interface for remote robot operation, only supports by Python 2.7, and AI models in Black Box can only work with Python 3. Different development environments highlight the significance of data transmission in the Pepper-GPT system.

A client-server model is employed for communication between the Black Box and the

Pepper Controller. Both programs have their clients in the system for different functionality enactments. In the system, commands from the Black Box are sent to the server for forwarding to the Pepper Controller. Subsequently, the Pepper Controller finds the corresponding action agents based on these commands and sends back an end flag after the robot finishes executions. This end flag signs the end of one small task and indicates the Black Box to prepare for the next audio recording (the beginning of the next task). In addition, to distinguish the destination of data transmission in the server, different prefix symbols are used in the transferred commands for identification.

To prevent potential data loss, the Pepper-GPT utilize the TCP/IP protocol during the data transmission, ensuring it is stable and reliable. Compared to the UDP protocol, TCP minimizes data loss as it resends the data if there is no confirmed signal sent out by the sender within a reasonable round-trip time, turning it into an ideal choice for transmitting data.

4 Evaluation

This section outlines a comparative approach of the three cutting-edge ASR systems (Google, Google Cloud, and Whisper) and the human trial designed to evaluate the performance of the Pepper-GPT.

4.1 Speech Recognition Comparison

A comparative analysis of three current advanced ASR systems is designed to select the optimal ASR systems for the speech recognition feature of the Pepper-GPT. It uses the WER and processing time as the measurement metric for the quantitative evaluation. And the 'Speech Accent Archive' dataset on Kaggle and the 'daily-dialogue' dataset on Hugging Face are utilised for testing.

4.1.1 Evaluation Metrics

The speech recognition study typically evaluates system accuracy based on the WER metric, as it is a widely accepted measure for such systems [20]. It is defined as the Eqn. 1.

$$WER = \frac{S \ + \ D \ + \ I}{N_1} \tag{1}$$

The meaning of each variable is shown below [13]:

- S = Number of Substitution Errors (other words replace the original word)
- D = Number of Deletion Errors (the original word is not transcribed)
- I = Number of Insertion Errors (unintentional addition of words)
- N_1 = Number of Reference Words

Notably, innocuous differences, such as a recognized name spelled differently but pronounced the same as the practical result, are ignored as it does not influence transcription content.

In addition to WER, processing time for speech recognition is utilised to evaluate the performance of ASR systems. The human interactions are real-time, and less processing time means more fluent and natural communication. Therefore, comparing the processing time is crucial for the ASR systems' assessment.

4.1.2 Dataset

Two tests evaluate the performance of ASR systems. The first test employed the 'Speech Accent Archive' dataset on Kaggle [21]. The dataset includes recording audio of the same sample content from 177 countries' speakers in English. The wide range of accent samples makes the dataset an optimal selection for assessing the adaptability of the ASR systems. In the accents test, the chosen accents comprised those from 5 English-speaking countries (Australia, Canada, New Zealand, United Kingdom, and the United States) as well as accents from 7 non-native English-speaking countries/regions (Africa, Arabic, China, France, India, Philippines, and Spain). Evaluating the ASR systems' adaptability with different accents helps widen the range of users and enhances its practicality and globality in the HRI field.

The second test utilised the 'daily-dialogue' dataset on Hugging Face [22] to evaluate ASR systems' proficiency in daily dialogue transcription. Five scenarios were chosen to ensure model excellence in practical, real-world dialogues.

4.2 Experiment Design

My teammate Katherine was responsible for designing the experiment to evaluate the Pepper-GPT performance with human interactions. The experiment was tested by 25 participants (11 males and 14 females) recruited randomly on campus. Before the experiments, participants would receive confirmation letters with the Participant Information Sheet and the Consent Form. After signing the consent forms, they agreed to attend the experiment to interact with the Pepper-GPT, and we arranged the sessions for them to come.

Before the experiment started, one of the researchers briefed participants on the integrated system's features and conversation guidelines. The other student researcher was in the room for technical support, discreetly seated in a corner during the whole experiment. A microphone was connected to the system to provide clear speech input. Subsequently, participants engaged in open-ended conversations for 5 to 10 minutes with the Pepper-GPT and filled out two questionnaires after the conversation to collect their personal information and feedback for evaluation.

5 Results

This section outlines the results of the ASR systems comparison and the human trial with the Pepper-GPT.

5.1 Speech Recognition Results

In the accent test, two sample data from each country/region are randomly selected for identification, resulting in a total of 24 test samples. Each sample was repeatedly recognized three times by Google, Google Cloud and Whisper ASR models, resulting in 216 transcribed texts (which will be manually labelled to calculate their WER) and the running time used for transcription. The test results will be re-divided into 12 groups according to the samples' source (country/region), and their average WER and recognition time will be calculated for evaluation, shown in Fig. 2.

Lower WER means higher accuracy of transcription results, while shorter processing time

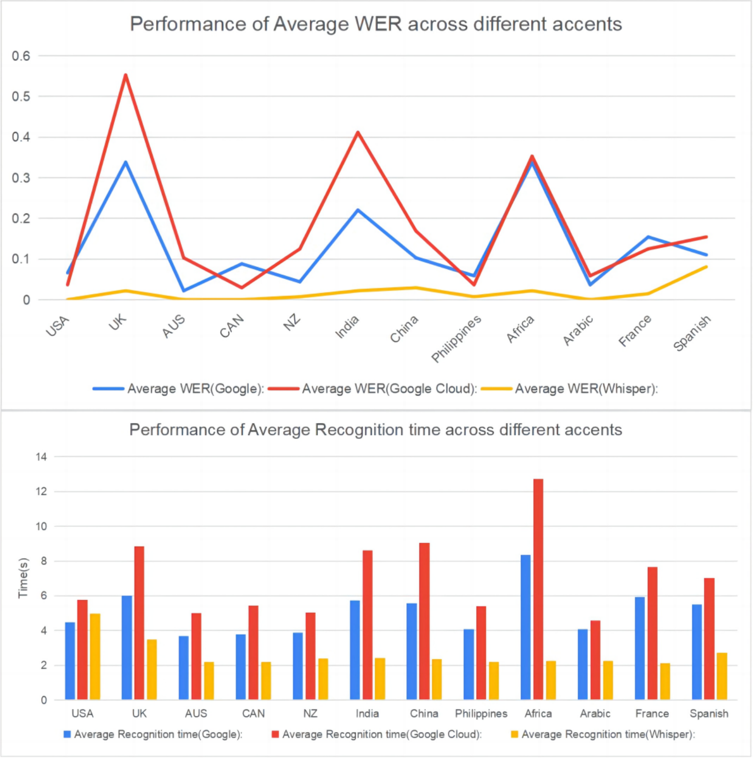


Figure 2: The Results of the Accent Test.

means higher efficiency of the ASR model. According to the experiment results, the Whisper ASR model got the lowest WER and shortest recognition time among the three ASR systems, confirming its outstanding performance as the optimal selection for the Pepper-GPT system.

Among non-English-speaking countries, Indian accents pose the highest challenge for recognition (average WER=0.218), while Arabic accents are more apparent to transcribe (average WER = 0.032). However, it is remarkable that compared to the non-native English-speaking countries, English-speaking countries typically got a lower average WER, excluding the UK. This situation is even more apparent when using Google and Google Cloud ASR models for speech recognition, with average rates of 0.338 and 0.553, respectively. In the meantime, American accents consistently demonstrated the lowest WER compared to other English-speaking countries across the experience. In addition, the Whisper ASR system has fewer differences in average recognition time among diverse countries, but the USA (4.976 s) and the UK (3.494 s) have the longest average recognition time.

Similar to the accent test, the daily-dialogue experience tests 5 selected conversation topics from the ’daily-dialogue’ dataset three times to calculate its average WER and recognition time. As displayed in Fig. 3, it is confirmed that Whisper consistently achieved the lowest WER of 1.72% and the shortest average recognition time of 5.258 seconds.

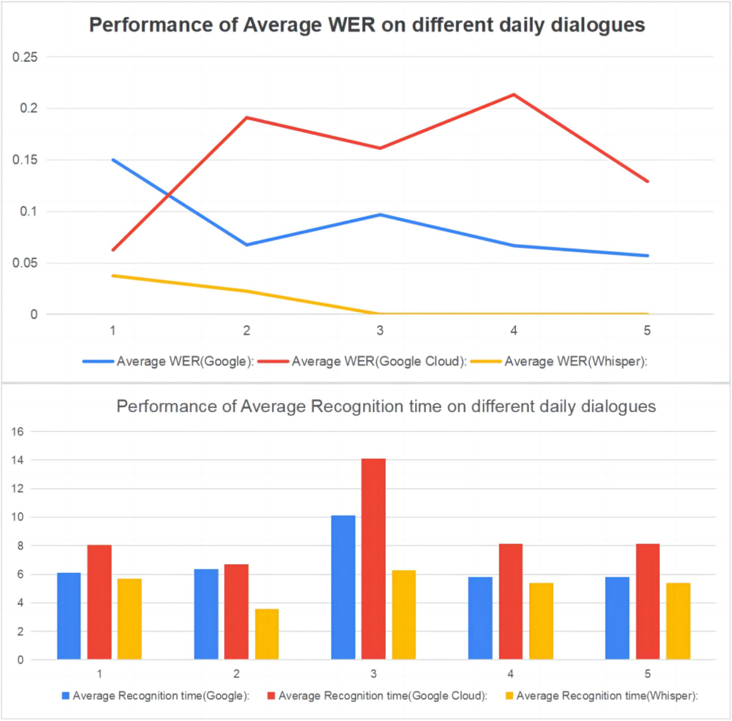


Figure 3: The Pseudo Code for the GPT module.

According to the results, the performance of the Whisper ASR model has consistently exceeded the other two ASR models, verifying its potential and appropriability in the Pepper-GPT design.

5.2 Human Trial Results

The study involved 25 individuals, with a noteworthy majority of 52% belonging to the 18 to 23 age range. According to the participants' information (Fig. 4), the majority of participants are Chinese (48%), and most come from the Engineering faculty (76%).

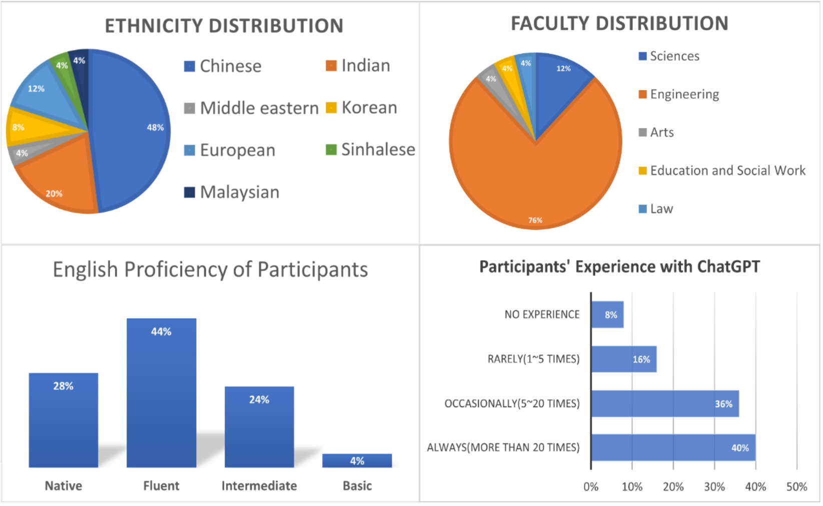


Figure 4: Participants' Information.

It is worth mentioning that 72% of the participants evaluated their oral English level as fluent/native, representing their ability to interact smoothly with the Pepper-GPT, as the system is currently designed to communicate only in English. Furthermore, 40% of experimenters stated that they often use ChatGPT. Participants' familiarity with ChatGPT allows them to compare the Pepper-GPT with ChatGPT more directly during the experiment and evaluate whether the system can improve UX.

According to the feedback from participants, 56% of participants considered the performance of the Pepper-GPT "excellent", while the rest ranked this system as "good". In addition to the general ranking, participants were required to comment on their feelings, including the excitement level and the ease of interactions with the Pepper-GPT, the appropriateness of the

Pepper-GPT's gestures during the conversation, and the performance of the comprehension and response abilities of the Pepper-GPT (Fig. 5).

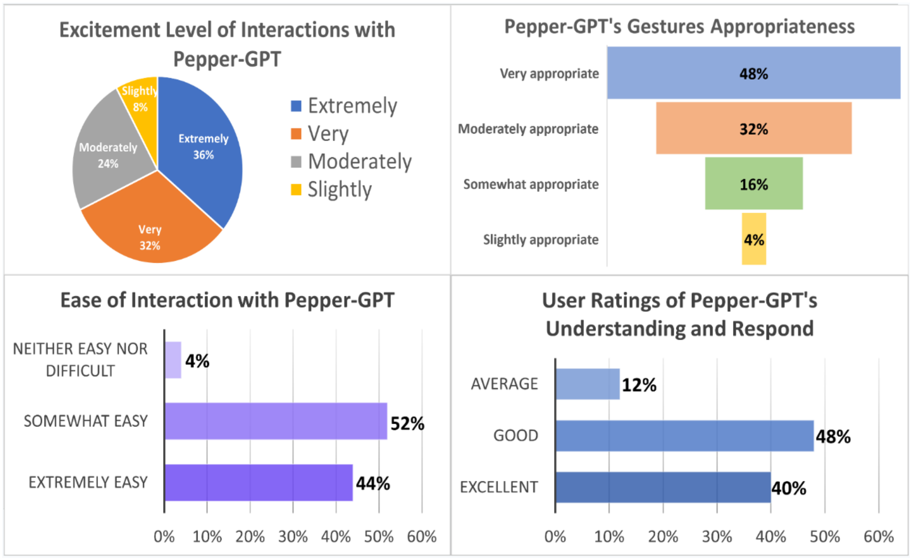


Figure 5: Human Trial Results.

The overall user feedback is positive. More than half of the participants were excited about interacting with Pepper-GPT. All participants confirmed Pepper-GPT's ability to understand and respond, with 88% rating its performance above average. In addition, only 4% of users believed that the gestures displayed by the robot during communication were not appropriate enough. It is worth mentioning that most users (52%) rated the ease of use of this system at a medium level. Even though it is still user-friendly, it indicates that future designs should be optimized in this aspect.

According to participant comments, some users found that Pepper embedded with ChatGPT can make conversations more engaging and human-like. However, some still favoured text-based communication with ChatGPT directly for its convenience. In conclusion, integrating ChatGPT with the real robot markedly improved the UX, creating a more natural and enjoyable interaction with users.

6 Discussion

In general, the ASR performance comparison results are consistent with the hypothesis that the recognition accuracy in English-speaking countries is expected to be higher than in non-native-English-speaking countries. The issue of poor UK-accent recognition may be because the tested ASR models were all developed by American companies. Their training data may mainly focus on American accents and lack enough British accent samples, resulting in the system's poor recognition of this accent.

The recognition time of the Google Cloud ASR model is always longer than other models in the accent and daily dialogue tests. The recognition process for the Google Cloud ASR system is to send the audio file to the cloud server and await the replies from the cloud while the other two ASR models are processed locally. Therefore, the network status significantly affects the recognition time of Google Cloud ASR, resulting in its extended recognition time.

Participants' English levels significantly impacted the overall UX during the experiment. Although the Whisper ASR system is the current cutting-edge speech recognition API and performed well in the speech recognition experiment, speakers with low fluent oral English abilities typically faced challenges communicating with the robot. The Pepper-GPT cannot accurately recognise audio content, which results in a misunderstanding and generating wrong commands. These participants sometimes must rephrase their content to make the robot understand correctly.

Additionally, some participants needed clarification about the engagement time with the robot as they could not distinguish whether it was listening to users or processing commands. This feature decreases the extent of system user-friendliness, demanding to enhance in the future.

In addition, participants with extensive prior experience with ChatGPT had higher expectations for the Pepper-GPT than less experienced users. This over-expectation may cause less satisfaction after the experience. However, most participants expected more physical actions designed for the Pepper-GPT in the future to make a further interaction.

Moreover, Pepper's poor facial tracking capabilities prevent it from maintaining eye contact with users during communication. Some participants tried to seek the robot's attention to keep the robot in a face-to-face state during the interaction, which declined their UX.

In summary, current results did not present any relationship between WER and recognition time. The high WER may emerge because of the data bias caused by lacking the UK accent training data. Both users and the robot collectively influence the ultimate UX, including the users' oral English proficiency and initial expectations and the robots' comprehension, response and facial tracking abilities. All the mentioned robot-relevant factors should be improved for further UX enhancement.

7 Conclusion and Future Work

In conclusion, integrating Whisper ASR and GPT-3.5 APIs with the Pepper robot significantly enhances UX. This integration system bridges the gap between virtual AI and physical robots.

Among the comparisons for speech recognition performance, the Whisper ASR system outperforms Google and Google Cloud ASR systems with a 1.716% WER and 2.639s processing time, improving Pepper-GPT's comprehension. In addition, human participant feedback shows over 90% find Pepper-GPT user-friendly, with 50% approving of the robot's gestures.

Pepper-GPT enables user content analysis, contextually relevant responses, and diverse physical interactions. The results indicate strong potential for the Pepper-GPT in HRI, with participants expressing enjoyment and interest, and predicting future interactions with the system.

There is no doubt that the Pepper-GPT significantly enhance the UX. However, several limitations still exist and need improvements for more natural, highly efficient and harmonious interactions of the Pepper-GPT in the future:

- **Listening Hint:** Users expect clear prompts to help them distinguish the state of the Pepper-GPT to know whether to communicate with the system or wait for the system to complete the task execution.

- **Multilingual Ability:** Current language configurations limit multilingual conversations. Future work should focus on enabling seamless language switches for diverse user preferences.

- **Fine-tuning:** The processing time for the GPT module to analyse and generate responses is extended. Therefore, Fine-tuning from OpenAI can reduce expenses and facilitate faster request response times [23].

- **Facial Tracking Enhancement:** The Pepper robot's facial tracking falls short, affecting natural interactions. Improving this feature is crucial for more engaging conversations.

References

[1] D. Mukherjee, K. Gupta, L. H. Chang, and H. Najjaran, “A survey of robot learning strategies for human-robot collaboration in industrial settings,” *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102231, 2 2022.

[2] R. Sharma, V. I. Pavlovic, and T. S. Huang, “Toward multimodal human-computer interface,” *Proceedings of the IEEE*, vol. 86, pp. 853–869, 1998.

[3] R. Hartson and P. S. Pyla, *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. 2012.

[4] S. Khan and C. Germak, “Reframing hri design opportunities for social robots: Lessons learnt from a service robotics case study approach using ux for hri,” *Future Internet*, vol. 10, 10 2018.

[5] B. Alenljung, J. Lindblom, R. Andreasson, and T. Ziemke, “User experience in social human-robot interaction,” *International Journal of Ambient Computing and Intelligence*, vol. 8, pp. 12–31, 4 2017.

[6] A. Gardecki, M. Podpora, R. Beniak, and B. Klin, “The pepper humanoid robot in front desk application,” Institute of Electrical and Electronics Engineers Inc., 8 2018.

[7] M. MatulÃk, M. Vavrecka, and L. VidovicovÃ¡, “Edutainment software for the pepper robot,” 2020.

[8] M. E. Foster, “Natural language generation for social robotics: Opportunities and challenges,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, 2019.

[9] A. Corrales-Paredes, D. O. Sanz, M. J. TerrÃ^3*n* − *L*Ã^3*pez, andV. Egido* − *Garc*Ã*a, “Userexperiencedesignforsocialrobots* : *Acasestudyinintegratingembodiment,” Sensors, vol.* 23, 2023.

[10] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q. L. Han, and Y. Tang, “A brief overview of chatgpt: The history, status quo and potential future development,” *IEEE/CAA Journal of Automatica Sinica*, vol. 10, pp. 1122–1136, 5 2023.

[11] D. Hebesberger, T. Koertner, C. Gisinger, and J. Pripfl, “A long-term autonomous robot at a care hospital: A mixed methods study on social acceptance and experiences of staff and older adults,” *International Journal of Social Robotics*, vol. 9, 2017.

[12] E. Billing, J. RosÃšn, and M. Lamb, “Language models for human-robot interaction,” 2023.

[13] F. Filippidou and L. Moussiades, “Î benchmarking of ibm, google and wit automatic speech recognition systems,” vol. 583 IFIP, 2020.

[14] R. P. MagalhÃ£es, D. J. R. Vasconcelos, G. S. Fernandes, L. A. Cruz, M. X. Sampaio, J. A. F. de MacÃªdo, and T. L. C. da Silva, "Evaluation of automatic speech recognition approaches," *Journal of Information and Data Management*, vol. 13, 2022.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2023.

[16] S. R. Shihab, N. Sultana, and A. Samad, "Revisiting the use of chatgpt in business and educational fields: Possibilities and challenges," *BULLET : Jurnal Multidisiplin Ilmu*, vol. 2, 2023.

[17] OpenAI, "Chat completions vs completions - openai documentation," 2023.

[18] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier." `https://github.com/snakers4/silero-vad`, 2021.

[19] Aldebaran, "Python SDK - Overview - Aldebaran 2.5.11.14a documentation," 2023.

[20] J. V. Egas-López and G. Gosztolya, *Predicting a Cold from Speech Using Fisher Vectors; SVM and XGBoost as Classifiers*, vol. 12335 LNAI. Springer, 2020.

[21] R. TATMAN, "Speech Accent Archive," 2017.

[22] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," 2017.

[23] OpenAI, "Fine-tuning guide." `https://platform.openai.com/docs/guides/fine-tuning`, 2023. Accessed: Oct. 11th, 2023.

A Pepper-GPT Details

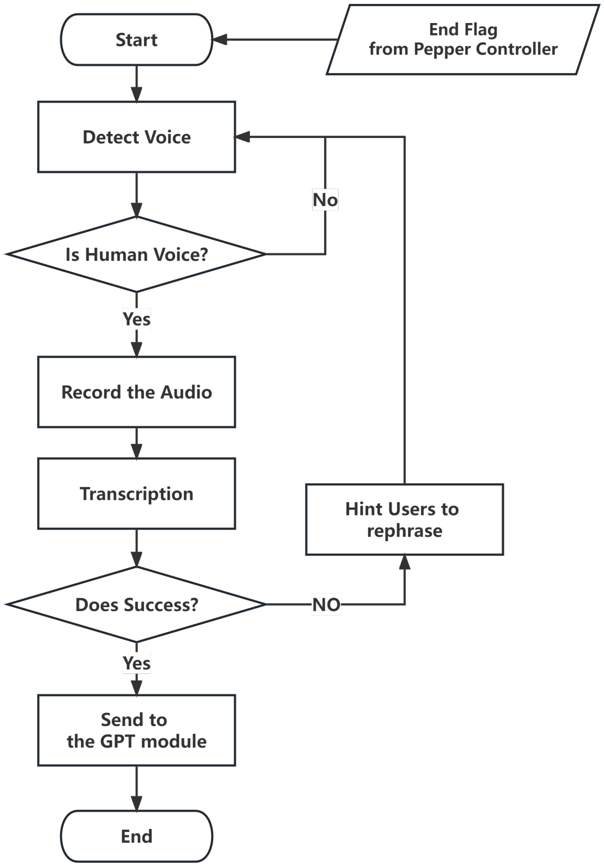
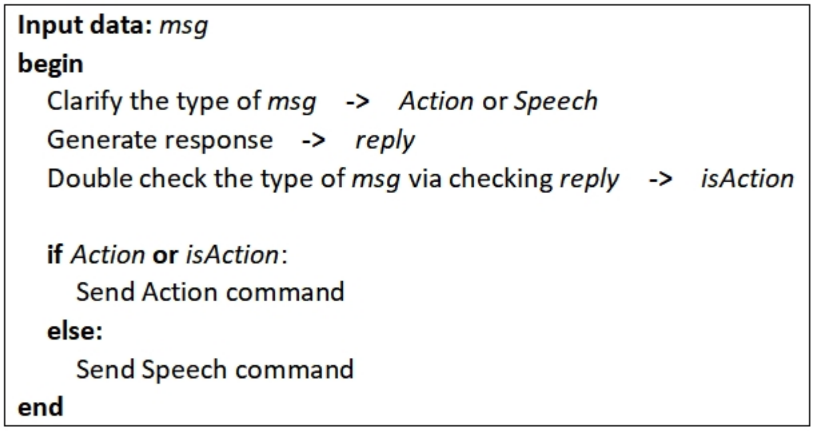


Figure 6: The Flowchart of the Speech Recognition Module.



```
Input data: msg
begin
  Clarify the type of msg  ->  Action or Speech
  Generate response  ->  reply
  Double check the type of msg via checking reply  ->  isAction

  if Action or isAction:
    Send Action command
  else:
    Send Speech command
end
```

Figure 7: The Pseudo Code for the GPT module.