2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) | 979-8-3503-3531-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOMET57998.2023.10099242

Introducing VAIFU: A Virtual Agent for Introducing and Familiarizing Users in VR

Muhammad Hasham Qazi
Electrical Engineering
Habib University
Karachi, Pakistan
mq05497@st.habib.edu.pk

Muhammad Palize Qazi
Business Administration
Institute of Business Administration (IBA)
Karachi, Pakistan
m.qazi.24027@khi.iba.edu.pk

***Abstract*—This paper introduces VAIFU, a Virtual Agent for Introducing and Familiarizing Users in Virtual Reality. VAIFU is an interactive embodied conversational agent that allows for both speech interaction and physical interactions in Virtual Reality. The goal of VAIFU is to introduce and familiarize new users with the landscape of virtual reality. The paper discusses the system design, agent development, interaction methods, and natural language processing techniques that may be used in order to develop such an agent for immersive human-like social interactions and how these social features may be beneficial to users. The paper concludes with a pilot study and future work for further iterations of the project.**

***Index Terms*—Human-Computer Interaction, Virtual Reality, Embodied Conversational Agents, Natural Language Processing Applications**

I. Introduction

Virtual Reality (VR) has seen massive growth in its software, hardware, ease of access, available usable applications, entertainment options, and userbase in the last few years [1]. The increasing popularity of immersive media and VR technology has expanded the user base and lowered the barriers to entry in terms of cost and community support. However, some new users may lack technical expertise, creating a need for new strategies to reduce the barrier to entry even further [2]. To achieve this goal and onboard users in a much more social and efficient manner, this paper proposes 'VAIFU', a 'Virtual Agent for Introducing and Familiarizing Users in VR'. VAIFU is an interactive Embodied Conversational Agent (ECA), a Human-Computer Interaction (HCI) technique that utilizes an intelligent virtual avatar to converse with the user, one that appears in a user's virtual home environment and acts as an assistant to help new users navigate the VR landscape. Example VAIFUs are as shown in Figure. 1.

A reason for choosing to use an ECA for this purpose is the multitude of benefits brought about by the use of ECAs as virtual assistants. Further information on these benefits may be found in [3] and [4]. As a system, VAIFU is currently able to: (1) provide explanations for different hardware and software features to get new users accustomed to the terminologies involved in the field, (2) provide navigation support for reaching and changing different software settings, (3) provide recommendations for different new and popular media such as games, applications, virtual tours, etc. Using an embodied conversational agent that is interactable through voice communication and physical interactions allows new users to be able to socialize in their new VR environment, even before they are exposed to VR social applications such as VRChat. This is important because socializing and collaboration are important aspects of virtual reality [5]. Furthermore, such a system allows new users to talk to the agent and use its functionalities without any prior knowledge of navigating VR systems. This helps create an environment that allows the users to familiarize themselves with the technology in a much more engaging manner.

(a) (b)

Fig. 1. Example VAIFU ECAs with different ethnic characteristics. (a) Represents VARA, an Asian Female VAIFU. (b) Represents ABE, a Black Male VAIFU.

This paper is structured in the following manner. Section 1 introduces the project and the motivation for its design. Section 2 covers existing literature in the domain of using embodied conversational agents in Extended Reality (XR), i.e. both Virtual Reality (VR) and Augmented Reality (AR), situations [6]. Section 3 covers the methodology and system design. Section 4 includes a small preliminary study and discusses

how this system may be understood from a human-computer interaction perspective. Section 5 mentions the limitations of this paper and future work to be done. Finally, Section 6 concludes the paper.

II. Literature Review

Existing work in this domain includes methods regarding creating effective embodied conversational agents, their different use cases, the effects of their use on people, etc. ECAs have been previously researched and used for diverse use cases in XR.

A. AR Social Robot 'Hupo'

In [7], Leo-Liu and Wu-Ouyang explored user experiences of human-robot interaction when employing a social robot with stylized humanistic (and anime) characteristics in its agent. This was done using a combination of augmented reality (AR), artificial intelligence, and anime [7]. Their results concluded that people used the AR anime humanoid agent due to their interests in (1) anime, (2) new technology, and (3) to reduce their loneliness and social segregation [7]. Users in the study also noted that while the AI features were lacking in the social robot, and were a cause of frustration, the human (anime) embodiment delivered some satisfaction to counteract their frustrations [7].

B. Exploring Virtual Agents for AR

In [3], Wang, Smith, and Ruiz explore different types of embodiments for ECAs. They discuss how the embodiment of a virtual assistant is extremely beneficial for human-computer (or human-robot) interaction, as it allows for both verbal and non-verbal communication (in the context of AR) and helps improve the presence and relatability of the agent. The paper then discusses the different types of agents (each with a different type of embodiment) deployed in AR for a user study to see which type of agent(s) generate the most user engagement and emotion while testing (playing a game) [3]. The paper concludes that users preferred the life-size and small-size human agent as compared to no-agent and an inanimate object agent when being provided with help and assistance [3].

C. Young Merlin: An ECA in VR

In [8], Gris et al. discuss the system for their ECA 'Merlin' which uses speech interactions to interact with humans in VR for storytelling purposes. The system, built on the Unity 3D software, allows the ECA to take in speech input and generate both speech and gesture outputs based on triggers and the decisions made in-game [8]. However, the paper uses a very basic interaction system for their agent, one which uses markup language and not Natural Language Processing (NLP) for carrying out conversations, which limits the number of interactions, thus disallowing truly dynamic conversations and interactions [8].

D. Conversational Agents and Mindfulness Assistant in VR

In [9], V. Häggström Fordell proposes a system for developing conversational agents in VR for the purposes of "interactive machine teaching" and as a "mindfulness assistant". This is done for a use case in the area of medicine [9]. The paper discusses why such a system is needed for rehabilitation and mindfulness purposes, how it may be achieved, and a case study compiling results from user testing [9]. While the paper does discuss the onboarding of users into VR, it does so from a "mindfulness coach" perspective where the utilized conversational agent is not embodied [9]. This, however, contrasts research findings in [3] and [10] which note that a human-like social agent may be preferable for new user onboarding.

E. Area of Further Research

While quite a lot of literature exists on how to create effective ECAs and how they are currently being used in junction with XR, the authors could, however, not find any literature addressing the use of ECAs as a tool for introducing VR to new users of the technology and for familiarizing them with the VR landscape. Systems that are inefficient in user acceptance are often abandoned for other alternatives [11]. While the current VR onboarding methods may be adequate for an average user, users who are not tech-savvy or not under the supervision of facilitators, especially users who are senior citizens, may find it difficult to use and navigate such equipment on their own [11], [12]. As such, ECAs may be used for this purpose as they also have a history of being used for similar pedagogical and navigational tasks [4], [13].

III. Methodology

A. Effective Embodied Conversational Agent Design

In order to design an effective ECA certain guidelines must be followed in order to make the conversational agent appear more human and thus trigger a greater social response. This will allow the user to develop more empathy for the agent and as such achieve much more cognitive stimulation [14]. One aspect which must be considered is the ECA's personality. As per [15], audiences preferred a young, cheerful, and friendly agent. In addition, the use of a "cuteness" effect may encourage users to overlook the ECA's shortcomings such as its limited AI functionalities [7], [16]. However, in certain serious situations, a much more stern, assertive and calm agent may also be used [15]. The authors propose the use of a friendly and cheerful ECA for the aforementioned scenario where it will be deployed. Creed, Beale, and Cowan, in [17], found that participants in their study strongly preferred an emotional agent over a non-emotional agent and that participants appreciated subtleties in the emotions of the agent such as even showing "negative emotions" (e.g. sadness, anger, etc.) when needed. Based on this, the authors propose that the conversational agent may also display negative emotions in addition to positive ones in order to create more immersion for the user. Furthermore, agents must be interactive in not only speech but other mannerisms as well such as body

language, expressions, etc. [17], [18]. As per [19], users also preferred gender-marked agents as compared to gender-ambiguous agents, with more users preferring female agents for assistant roles. One more aspect that may benefit human-agent interaction is the use of "racial mirroring" i.e. creating an agent of a race or culture that the user may be able to relate to [20]. [20] and [21] found that users were much more accepting of virtual agents if the agents appeared to be of their own ethnic, cultural, or racial groups. Based on this, agents of different racial and cultural features may be created in order to further engage the new users. This is shown in Figure. 1. In addition to these, there are countless other design features and guidelines that must be considered when designing effective ECAs. Some of these are mentioned in and are referred from [22].

B. Software Used

VRoid Studio was used to design the 3D Models for the virtual agents and allows for multiple customizations such as race, etc. [20], [21]. Blender 3D, a 3D modeling software, was used to develop custom 3D assets. Mixamo was used to animate the agent using pre-captured motion capture data, thus helping create much more human-like movements. Unity 3D was used to develop the VR simulation environment. This software (with the help of some open-source GitHub repositories [23]) allowed for all the 3D models, animations, and interaction methods to be incorporated into one VR application. The Unity XR Interaction Toolkit was used to develop the VR interactions. The Meta Voice SDK (built using Wit.ai) was used for conversational NLP functionalities and speech recognition tasks. This allowed the users to converse with the virtual agent using natural language dialogue. The project was developed for and tested on the Meta Quest 2 HMD [24].

C. System Architecture

The proposed architecture for the design of VAIFU is shown in Figure. 2. To start, the VAIFU must first be summoned. This is as shown in Figure. 3 and is achieved by using a custom button mapping on the controller, hereafter termed the "Agent Spawn Button". Once the VAIFU is summoned, the user may start a conversation using an initiation prompt. Since the agent used in this iteration is named 'VARA', the use of the phrase *"Hello VARA"* prompts the agent. Once done, the agent goes from the *Idle* state to the *listening to user* state and stays there while the user is speaking. This is shown in 3. Once the user stops speaking, the speech is converted into text using the Meta Voice SDK. This text is then sent to Wit.ai to process and extract user intents and send the relevant triggers to the *VAIFU Action Manager*. The VAIFU Action Manager uses these triggers to execute the relevant animation and dialogue set, hereafter known as the *"task set"*, through the agent. These task sets are pre-made for each of the questions that the agent can currently answer, and are composed of (1) displaying the required explanation background screen, (2) playing the relevant animations, and (3) speaking the relevant explanation dialogue. For example, as seen in Figure. 6, if a user asks VARA for the top rhythm game to play, the intent is processed from the input speech, the relevant triggers are sent to the VAIFU Action Manager, and VARA consequently displays the relevant board, plays the relevant animation set, and speaks the dialogue suitable for telling the user about the game. Once a set is complete, VARA then returns back to her *Idle* state which plays the idle animation and VARA speaks no dialogue. During the entirety of this simulation, VARA may be despawned at any moment if a user desires and this may be achieved by pressing the Agent Spawn Button again. In future iterations of the project, an emotion analysis system will also be added, as seen in Figure. 2. This is further discussed in Section 5. It is also to be noted that the use of physical interactions overrides any other states that the VAIFU may currently be on and such interactions load the relevant responsive *Physical Interaction State*, as seen in Figure. 4. Furthermore, as currently, only a preset number of answer states exist, we are able to use pre-recorded voice dialogue, with the answer *"I am afraid I cannot answer that."* as a default for any questions that the agent cannot answer. In future iterations, a fully conversational agent using the OpenAI API will be developed, and a text-to-speech engine will also be required. The Meta Voice SDK is planned to be used for this purpose as well.

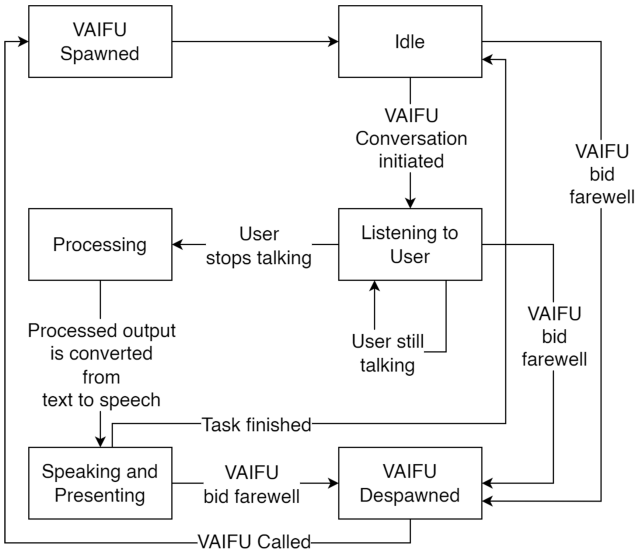


Fig. 2. State Transition diagram of VAIFU.

D. Media Presentation and Physical Interactions

As the main purpose of VAIFU is to introduce VR to new users and familiarize them with its landscape, the agent incorporates relevant features to accomplish these purposes. The first of these features is the use of a background screen that is used to display graphics to the user. This allows the user to not only hear about but also see what the agent is explaining in a highly visual manner. This may be seen in Figure. 5 where VAIFU explains the controller layout to the

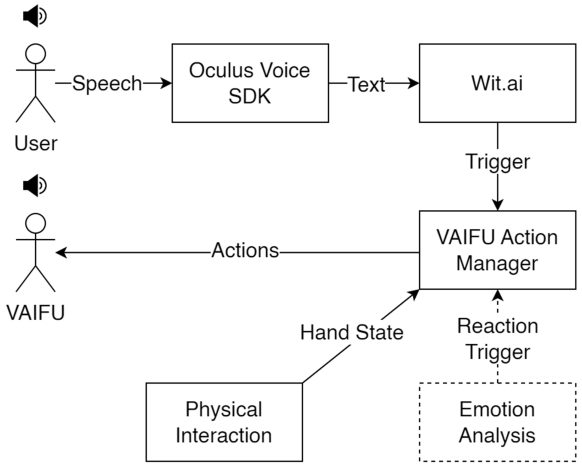
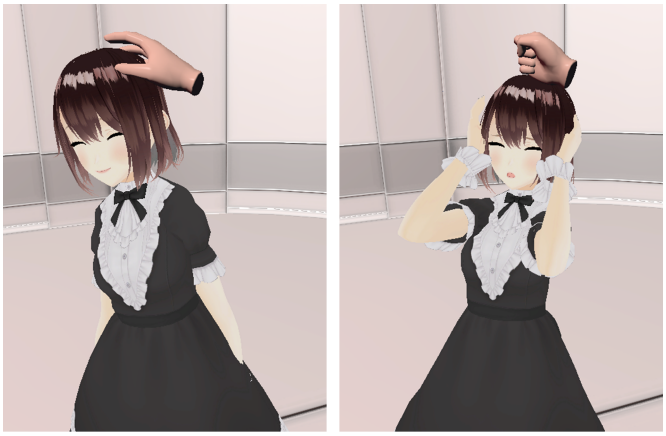


Fig. 3. System Architecture of VAIFU.



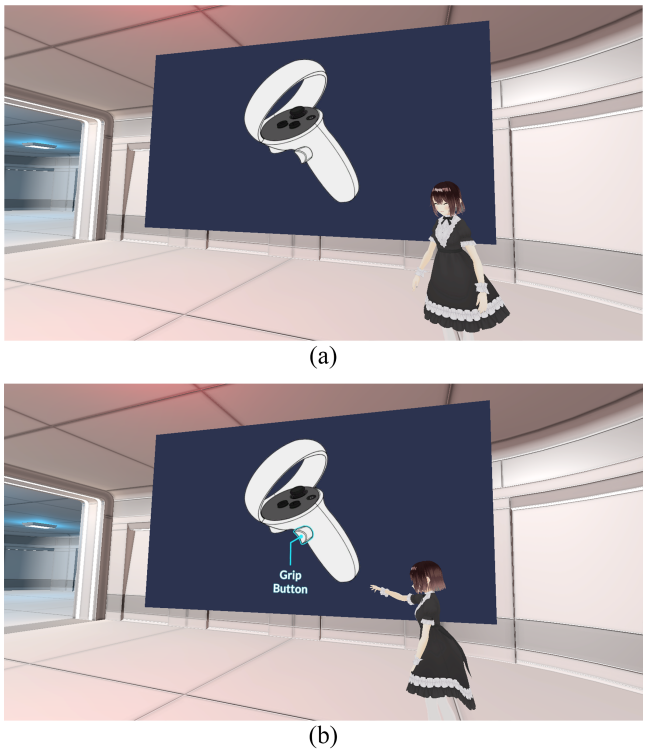
(a) (b)

Fig. 4. Physical Interactions with VAIFU. (a) Positive bashful reaction to an open-hand interaction. (b) Negative anguish reaction to a fist interaction.

user. This task set may be executed in response to a variety of user prompts such as the user asking about the controller layout, the user asking what the 'Grip" button is, etc.

Once the VAIFU completes answering the asked question, it returns back to the *Idle* state. Currently, a specific number of question areas have been identified for executing their relevant task sets. These include the *Guardian System*, *controller layouts*, *interactable objects*, *full-body tracking*, *new and popular games (genre may be specified)*, *new and popular applications* and other *entertainment*, how to alter settings such as *brightness*, *volume*, *muting*, *device-specific features*, *passthrough mode*, *hand-tracking*, *personalization features*, *accessibility settings*, and more.

While it may be seen how VAIFU may be used to introduce users to VR, another goal for this project is to also familiarize them. This may include things such as explaining different VR terms, explaining how to change certain settings, recommending games to play, etc. An example of how this may take place can be seen in Figure. 6. Here, VAIFU was prompted by the user to tell them about good rhythm games to play in VR.



(a) (b)

Fig. 5. VAIFU Introducing the VR controller to the user. Images (a) and (b) are frames taken from the task set sequence.

VARA was then able to provide answers on popular rhythm games that may be played.

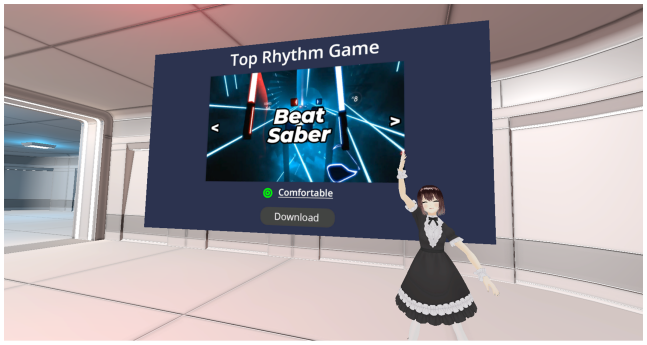


Fig. 6. VAIFU providing a game recommendation.

VAIFU is also able to allow the user to perform physical interactions. These interactions can be both positive and negative in nature as seen in Figure. 4. For the purposes of this project, only the head was configured to be able to recognize physical interactions. This was accomplished using a collider on the head of the VAIFU model as seen in Figure. 7. When the user's hand models collide and intersect with the VAIFU head collider, a trigger is sent to the VAIFU Action Manager based on the hand's configuration. If the grip button is pressed, a fist is formed. This is used to simulate the action of punching and the trigger sent is one which executes the negative reaction. However, if an open hand is used, this is used to simulate

the use of a soft hand. This triggers a positive reaction. In future iterations, the velocity of the hand may also be used to determine the trigger that may be sent to the VAIFU Action Manager to alter which reaction is executed. This will allow distinctions to be made between different physical actions with similar hand states such as soft head pats versus slapping.

Fig. 7. Head Collider used for Physical Interactions.

E. Voice Interaction and Natural Language Processing

The Meta Voice SDK was used to convert the user's speech to text. This was then sent to Wit.ai. Wit.ai uses utterances to derive intents which are then used to understand what the user wishes to accomplish. These utterances, intents, and entities are then used to train NLP models which allow the user to use conversational speech in order to achieve specific outcomes. For example, learning how a certain component works or where a certain setting may be found are questions that the user may have, and as such, the model must be trained using sample utterances (and their synonyms) in order to derive user intent at runtime. Keywords such as *What*, *Why*, *How*, *When*, *Where*, *Who*, etc. were used as prompts, alongside the objects of the questions, such as the controller as shown in Figure. 8. This was done for different types of questions that the user may ask.

Future iterations of this project will use the OpenAI API in junction with Wit.ai to allow for much higher conversational capabilities.

IV. Preliminary Study and Discussion

Recent studies suggest that VR has the potential to assist various individuals, including those with mental health issues, disabilities, social isolation, etc. [14], [25]. VAIFU may prove useful to achieve this purpose as it allows new VR users to interact with human-like ECAs to get comfortable with VR technology in an immersive and social manner [16], [18]. To test the efficacy of VAIFU, a pilot study was conducted with 10 participants. The participants consisted of 7 males and 3 females. In order to benchmark and compare VAIFU to existing methods, first-time VR users participated in a single-blind test. Half (n=5) of the total participants were selected

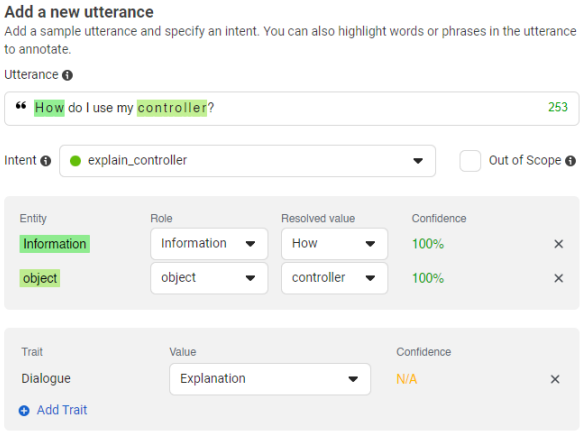


Fig. 8. Sample Screen from Wit.ai.

randomly and were made to use the First Steps for Quest 2 application (hereafter referred to as First Steps) first and then the VAIFU application [26]. In order to have a counterbalanced study, the remaining half of the total participants (n=5) were made to use the VAIFU application first followed by the First Steps application. All participants were asked to provide qualitative feedback on their experience using both applications which showed that the initial hypothesis was validated to a certain degree. The participants' feedback showed that they enjoyed interacting with VAIFU and preferred having an embodied agent as compared to solely a voice in the environment (as in First Steps). However, participants also noted that having an interactive environment such as in First Steps allowed them to explore VR in a more immersive manner. Participants were also amused by the physical interactions with VAIFU, however, some participants noted that a variety of other reactions, in addition to the current ones, may help make VAIFU even more lifelike. Similar to previous literature, the participants of this study also mentioned the limitations of the VAIFU's conversational abilities, particularly noting the limited voice interactions that were available with VAIFU and the issue of improper lipsyncing. According to the participants, the combination of these flaws resulted in lower fidelity of interaction with the agent. In addition to this, general feedback showed that participants found it confusing to prompt VAIFU to explain each feature separately. Therefore, future iterations may contain a list of possible prompts that may be presented to the user to create more engagement. The study's participants found the VAIFU familiarization features, such as game recommendations, to be significantly more efficient than navigating through the Meta Quest Store when initially looking for new applications to experiment with.

V. Limitations and Future Work

To assess the viability of VAIFU as a system for introducing and familiarizing users with VR, the initial prototype was designed with the fundamental features required to achieve this objective. A small pilot study was carried out to identify potential areas for improvement. Due to the limited scope of the