Barra, Paola, Cantone, Andrea Antonio, Francese, Rita, Giammetti, Marco, Sais, Raffaele, Santosuosso, Otino Pio, Sepe, Aurelio, Spera, Simone, Tortora, Genoveffa, Vitiello, Giuliana (2024): A Task-oriented Multimodal Conversational Interface for a CSCW Immersive Virtual Environment. In: Proceedings of the 22nd European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centered Computing on the Design of Cooperation Technologies - Exploratory papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.48340/ecscw2024_ep05

A Task-oriented Multimodal Conversational Interface for a CSCW Immersive Virtual Environment

Paola Barra[1], Andrea Antonio Cantone[2], Rita Francese[2], Marco Giammetti[2], Raffaele Sais[2], Otino Pio Santosuosso[2], Aurelio Sepe[2], Simone Spera[2], Genoveffa Tortora[2], and Giuliana Vitiello[2]

[1]Department of Science and Technology, University of Naples Parthenope, Italy
[2]Department of Computer Science, University of Salerno, Italy

Abstract. In CSCW immersive Virtual Reality environments, users may be uncomfortable when interacting with a two-dimensional menu. Multimodal conversational interfaces may enhance the interaction enabling users to communicate with the system in different modalities. In this paper, we investigate the use of an embodied multimodal chatbot for improving interaction in a Virtual Reality (VR) environment simulating a working context. In particular, we adopt a User-Centered Design approach to build a multimodal conversational interface, named Muxi, in which a task-oriented voice avatar is enhanced with an interactive board for supporting meeting organization in VR. Users were involved in all the development phases, from task definition to iterative user testing. To assess the usability of the proposed interface, we conducted a controlled experiment involving 32 participants to compare the use of Muxi with a traditional menu-based interface in a CSCW environment. We performed quantitative analysis, concerning efficiency and effectiveness assessment, and qualitative analysis, related to participant cognitive load and perceived usability. Results revealed that our multimodal interface increases usability by greatly alleviating cognitive load and improving user performance, representing a good alternative to a menu-based interface.

1 Introduction

Since the advent of the Graphical User Interface (GUI), menus have been recognized as essential tools for computer users. They help users navigate through various items and select one of them. They are also adopted in 3D VR environments, where they require the user to select the menu item by using gestures or controllers to indicate the object and confirm the selection (Mundt and Mathew, 2020; Wang et al., 2021).

In VR, however, the user is immersed in a three-dimensional, spatial environment, and it may be uncomfortable to have to interact with a two-dimensional menu. Multimodal conversational interfaces may enable the user to communicate with the computer in different modalities, such as speech, text, gesture, image, video, and sound. Introducing them in VR environments may improve the system's usability. VR voice assistants are generally implemented by using an avatar (Zhao et al., 2022) to increase the presence perception and engagement of the users by providing a more realistic interaction.

The design of a multimodal conversational interface is not an easy task (Crovari et al., 2020; Francese et al., 2022). It requires the choice of the most appropriate interaction modalities for the user, the task, and the context. In addition, multiple modalities have to be integrated coherently and consistently, providing clear and intuitive feedback to the user (Sebillo et al., 2009). In multimodal conversational interfaces, interaction may also depend on the chatbot type, ranging from service chatbot useful for customer support (Mohamad Suhaili et al., 2021), to task-oriented chatbot helping users complete tasks in specific domains, to Personal Assistants, serving the user continuously, to general purpose chatbots (Følstad et al., 2019). Chatbots are also adopted to support collaborative work and learning in VR environments (Trappey et al., 2022; David et al., 2019; De Lucia et al., 2009).

In this paper, we equipped the multi-user VR CSCW Environment MetaCUX (Barra et al., 2023a,b) with a multi-modal conversational interface, named Muxi, for helping users in tasks related to the setting of a working environment, such as creating a meeting room.

The main contributions of the paper are the following:

- We describe the User-Centered Design (UCD) approach we followed to design a multi-modal task-oriented CSCW conversational interface.
- The proposed interface enhances the vocal interaction provided by an embodied avatar with a board GUI.
- We conduct a user study involving 32 participants aiming at assessing the impact of the use of a multimodal task-oriented chatbot versus a menu-based interface on user performance and perception when interacting in an immersive virtual environment.

The paper is structured as follows: Section 2 discusses related work. Section 3 describes the MetaCUX system and Section 4 describes the UCD methodology used for the development of the multimodal conversational interface Muxi. Section 5 describes the experimental user study while in Section 6 results are reported and discussed. Finally, Section 7 concludes the paper.

2 Related work

In this section, we discuss the research efforts that have been devoted to the use and assessment of menu-based and Chatbot interfaces in VR environments, and the support offered by task-oriented chatbots in VR CSCW environments.

2.1 Menu-based interface in VR environment

Das and Borst (2010) compared different types of design choices for Menu in VR: layout (pie vs. linear list), placement (fixed vs. contextual), and pointing method (ray vs. pointer-attached-to-menu) reporting the pros and cons of each of them. They involved 34 participants and compared time and errors. Mundt and Mathew also assessed the use of several types of pie-menu (Mundt and Mathew, 2020) with 24 participants, assessing usability, user experience, presence, error rate, and selection time.

The authors in (Lipari and Borst, 2015) integrated touch menus into a cohesive smartphone-based VR controller. Users transitioned between the menu interaction area and the other for spatial interactions such as VR object navigation areas. The study involved 20 participants and compared touch menu selection and ray-based selection by measuring time, errors, and user satisfaction. Results showed that both techniques have advantages and disadvantages.

Wang et al. (2021) assessed the use of handled menus in VR that follow the users to move, without obstructing their vision. They compared two types of menu interfaces (fixed menu and handheld menu) and three selection techniques. The choice of the solution depends on the contexts of use and end-users.

2.2 Chatbot in VR environment

The amount of experiments on chatbot usability has increased in the literature. Generally, it is assessed with experiments measuring usability based on effectiveness, efficiency, and satisfaction (Ren et al., 2022). The study proposed in (Nguyen et al., 2022) investigated disparities in user satisfaction between a chatbot and a menu-based interface system related to a mobile app. The research findings reveal that the use of the chatbot results in a decreased level of perceived autonomy and increased cognitive load compared to menu-based interface systems, ultimately leading to lower user satisfaction. This study suggests that advanced technology may not always be the optimal solution to organizational problems, which could lead to unintended negative consequences if user concerns are not adequately addressed.

Concerning the usability assessment of chatbots in VR, few works performed this kind of analysis. Indeed, Trappey et al. (2022) introduced a VR chatbot trained to answer frequently asked questions (FAQs) from a power transformer manufacturer. They assessed only the performance of the NLP models, which achieved an accuracy rate exceeding 91%. No user study has been conducted. In (Xie et al., 2023), chatbots are integrated into a university platform to assist both students and teachers with various tasks. Also in this study, no user study has been conducted.

Pick et al. (2017) compared speech-based and pie-menu-based interaction for the control of complex VR applications. They conducted a user study with 20 participants and assessed their performance in terms of time and errors and perceived usability. It resulted that on one side, speech is faster but on the other side pie menus are less error-prone.

2.3 Supporting CSCW with task-oriented chatbots

Task-oriented chatbots in VR are designed with specific purposes. They focus on assisting users in achieving well-defined tasks within the working VR environment. Examples include managing virtual meetings, coordinating complex projects, or providing real-time information.

Wang et al. (2023) provided guidance for online retailers to design chatbots with appropriate communication styles for effective service recovery in electronic commerce. Trappey et al. (2022) considered the context of industrial equipment manufacturing, involving customized design, assembly, installation, and maintenance services for electric power transformers. These services cater to the specific needs of customers. They proposed a VR-Enabled Chatbot for intelligent engineering consultation. The chatbot provides VR users with highly interactive and realistic graphical views during engineering counseling sessions.

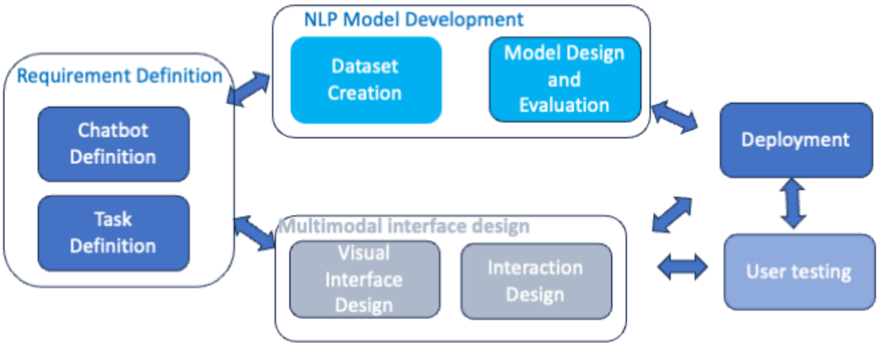


Figure 1: The adopted development process.

Unlike previous works that investigated the use of chatbot or menu-based interfaces, we propose a multimodal conversational interface enhancing a vocal chatbot represented by a virtual avatar with an interactive board. The novelty of this paper lies also in adapting UCD principles to chatbot development in a CSCW VR meeting environment, ensuring that these AI-driven interfaces truly enhance user experiences and seamlessly integrate into their daily interactions. We also assess the multimodal interface usability by comparing it with an already existing menu-based interface.

3 The MetaCUX system

The MetaCUX system is a multi-user CSCW VR immersive environment (Barra et al., 2023b,a). It allows users to choose a customized avatar, offered by Meta, for navigating different virtual rooms.

A user can play two roles: *the organizer*, enabled to create a new public or private room, select the scenario, and manage the creation and scheduling of various activities, such as meetings, interviews, etc. or *the participant*, enabled to enter rooms and perform activities organized by other users. Whenever the organizer changes the room scenario, all users in the room can view the change in real-time. The new scenario is automatically loaded for everyone.

4 Enhancing MetaCUX through a UCD approach

The goal of introducing a multimodal assistant into MetaCUX is to try to simplify the interaction. To this aim, we adopted a UCD methodological approach consisting of the steps summarized in Fig.1. We involved the users in the various development phases.

4.1 Requirement definition

Muxi is expected to assist the user in performing the meeting management inside the MetaCUX environment.

4.1.1 Chatbot definition

The chatbot is task-oriented (*chatbot type definition*) to be used in a CSCW VR environment. For this type of chatbot, we select a user-driven dialogue (*dialog type definition*): the chatbot has to identify the user intent, serve it, and provide feedback on the result. The relation is short-term (*relation-type definition*): the chatbot considers the user as a newcomer, it does not remember the past interactions (Følstad et al., 2019).

4.1.2 Task definition

To identify the tasks that require greater support we performed a preliminary study on the management of a meeting in MetaCUX, involving three HCI experts. We first let them freely use the original system and then asked them to use the features for creating rooms, changing the scenarios, scheduling and participating in a meeting, writing on the whiteboard, and so on. Then, immediately after leaving the experience, we conducted a focus group (Kontio et al., 2008) by following a discussion template previously prepared by the authors of this paper (Cassell et al., 2004). During the focus group meeting, which lasted 30 minutes, participants had to reveal the positive and negative aspects of their experience. From the discussion it emerged that the most critical interaction aspects were found for the following tasks:

- Creating a new room;
- Scheduling a new meeting;
- Changing the room scenario.

4.2 NLP model development

In this study, targeted data collection was conducted to develop two Deep Learning models: one for intent recognition and the other, Named Entity Recognition (NER), capable of understanding and interpreting users' intentions and recognizing named entities within a voice request. We associated each interaction task with an intent Muxi has to detect to accomplish the task.

4.2.1 Dataset creation

To create a dataset for training the NLP model implementing the chatbot Muxi, we first studied possible human-based dialogues for performing those tasks. Thus, we conducted a survey, which provides an example of the three intents the chatbot has to execute and requires two sentences for each of them. This small number of sentences has been chosen to avoid overloading the user. A group of 31 volunteer users (Computer Science students) were involved. They were asked to simulate the inquiry of a voice assistant for performing the considered tasks and fill out the form with their requests. We collected 186 sentences.

4.2.2 Data Augmentation

We removed the duplicated sentences. Then, the original dataset, consisting of 157 sentences collected with their corresponding intent labels was input to ChatGPT which was required to generate similar sentences. It was also tasked with altering the structure of existing sentences. This approach resulted in the generation of a new dataset consisting of *900 sentences*, divided into 300 sentences for each of the three intents.

To create a training dataset for intent detection the data collected were pre-processed as follows.

1. *Data Cleaning*, consisting in removing duplicate and inconsistent sentences.
2. *Tokenization*, the method of splitting a large text into tokens, which are shorter texts.
3. *Stopwording*, consisting in the removal of commonly used terms, such as "a", "an", and "the".
4. *Lemminization*, consisting in reducing the words to their root, e.g., "running" is reduced to "run".
5. *Vectorization and Transformation*, the text data were converted into a numeric format so that it can be used as input for NLP tasks for BERT.

4.2.3 NER dataset creation

The tagging of the datasets of NER models was done manually. In particular, we defined the tags for intent as follows:

- *Create rooms*:
  - Scenario type: B_TYPE_SCEN;
  - Number of participants: B_NUM_PART;
  - Room name: B_NAME
- *Create meeting*:
  - Name meeting: B_NAME-MEETING;
  - Meeting description: B_DESCR;
  - Day: B_DAY;
  - Month: B_MONTH;
  - Start time: B_HOUR-START;
  - Finish time: B_HOUR-END;

- *Change scenario*:
  - Scenario name: B_NAME_SCEN;

The dataset was divided into sentences and words, along with their respective named entity labels. Additionally, missing labels were filled using the forward-fill method to ensure dataset consistency. The labels were converted to uppercase for uniform formatting.

4.3 Model development

For *intent recognition*, we adopted the pre-trained BERT model[1]. In particular, we adapted the BERT model for the specific task of intent recognition by including a dropout layer to prevent overfitting, and an output layer for the three intent classifications. Also, the *NER model* is based on BERT, utilizing the implementation provided by the "simpletransformers" library[2].

Both models were validated with K-fold cross-validation, for K=5. Their performance was assessed by using on the test set standard multiclass evaluation metrics, such as Macro Average Precision (MAPrecision), Macro Average Recall (MARecall), and Macro Average F1 (MAF1) (Berger and Guda, 2020), reported in Table I for both the models and computed as follows, where p_i and r_i are precision and recall computed on the multiclass Confusion Matrix on the i-th class, for $i = 1 \ldots 3$. These measures are computed by assessing the Task Completion Success.

$$MAPrecision = \frac{\sum_{i=1}^{3} p_i}{3}, MARecall = \frac{\sum_{i=1}^{3} r_i}{3})$$

$$MAF1 = 2 * \left(\frac{MAPrecision * MARecall}{MAPrecision + MARecall} \right)$$

Table I: Model performance

Model	**MAPrecision**	**MARecall**	**MAF1**
Intent rec.	92.72	91.52	92.12
NER	89.24	88.07	88.65

When the accuracy of all three intents is lower than 75% we consider that the chatbot does not understand the question or it is inappropriate.

4.4 Multimodal interface design

The design phase is concerned with both the visual appearance of the interface and the interaction modality the interface offers.

[1] https://huggingface.co/docs/transformers/index

[2] https://simpletransformers.ai/

Table II: The adopted usability guidelines (Crovari et al., 2020)

ID	Guideline
P1	Show, don't tell.
P2	Separate feedback from support
P3	Show information only when necessary
P4	Design a light interface — emphasize content
P5	Show one modality at a time
P6	Do not overload multiple modalities beyond user preferences and capabilities
P7	Use multi-modality to resolve ambiguities

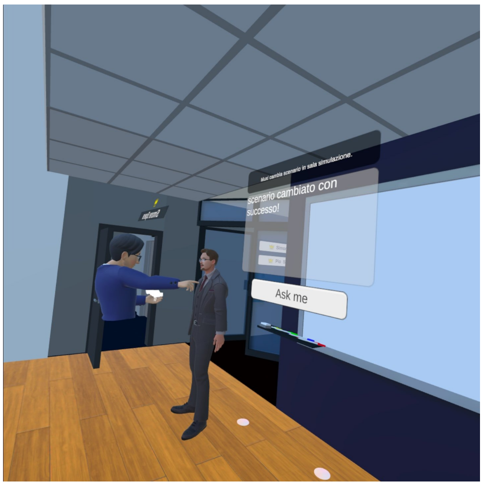
4.4.1 Visual Interface design

To make the user experience more engaging and realistic in a task-oriented interaction, we decided to represent the chatbot with an avatar. In some cases, visual interaction may be preferred to the vocal one, e.g., when a list of the available actions is provided. Thus, we decided to offer a multimodal interface consisting of the chatbot avatar equipped with an interactive board. The appearance of the avatar and the board should be appropriate for the type of environment in which they are introduced, a working setting in our case. The final result is shown in Fig. 2, where the user is on the left (with the label of his name) and the avatar is on the right, near the board. We animated the avatar with movements that resemble a person gesturing while speaking.

4.4.2 Interaction design

There is a need to design how the different communication approaches have to combine the two elements (chatbot and board) to avoid overloading or confusing the user, also considering the wide space of the virtual environment. We followed the design guidelines (Crovari et al., 2020) summarized in Table II while Table III describes how the guidelines have been applied to the Muxi design.

As shown in Fig. 2, the user avatar starts the interaction with the chatbot by pressing the "Ask me" button on the board. In particular, the P1 guideline is related to providing the user feedback on what the chatbot has understood of the user request. The user may pronounce a sentence, such as *"Create a meeting room for twenty people called job interviews."* Visual feedback is provided on the higher part of the board, where the text of the user command is displayed. This is useful to permit the user to give again the command in case of misunderstanding. In the case the conversation is out of the three individuated topics the chatbot vocally specifies that it does not understand the question and shows on the board a description of the task it may perform (P2).



Muxi cambia scenario in sala simulazione.

scenario cambiato con successo!

Figure 2: The multimodal conversational interface.

Table III: Application of the usability guidelines to the Muxi interface design

ID	Guideline application
P1	Use the visual interface to display the user sentence after the pronunciation.
P2	Feedback on the performed operation is vocally provided, while support (e.g., what the user can or should do in the next interactions) is visually shown.
P3	The GUI changes according to the conversation.
P4	The vocal interface provides only essential information.
P5	One modality at a time is adopted to provide information.
P6	Feedback is vocally provided, list of actions is provided in the support visual interface.
P7	Both the conversational and the visual interfaces produce a message when a task is successfully executed.

4.5 Deployment

The multimodal conversational interface is deployed on a client-server system and communicates with the client via an API call. The interaction between the user and the bot has been implemented as follows.

- **Speech-To-Text:** for Speech-To-Text (STT) we adopted the *wit.ai*[3] NLP platform, which provides various tools and services to build conversational interfaces.
- **Bot intent elaboration**: the translated text is sent to the trained NLP models that recognize the intent of the sentence and the related attributes.
- **Text-To-Speech Result:** the opensource TTS engine eSpeak[4] has been selected.
- **Avatar implementation**: The Muxi avatar performs the required action and provides feedback to the user. It was created using the ready player sdk[5] and has lip and body movements synchronized with the bot's voice during conversation. It is included in MetaCUX.

4.6 Pilot user testing

We performed two iterations. In the first, we involved three users (the same participating in the task definition phase) to experiment with the first Wizard-of-Oz prototype, in which the avatar speech was pronounced by one of the authors, and another author managed the room changes and the display. One of the participants suggested better highlighting the "Ask me" button. In the first iteration, the avatar was a futuristic man with a head-mounted display. An avatar more appropriate to a working setting was preferred, such as the man dressed formally shown in Fig. 2.

We performed a second iteration with the same users and a running prototype. Participants suggested displaying the user sentence (see the top of the board in Fig. 2) and adding the Chatbot feedback when the intent is not understood. We enhanced the final prototype with these last suggestions and then performed a user study with real users, as described in the following.

5 Evaluation Planning and design

The experimental design and other measures were approved by Computer Science Ethics Board of the University of Salerno. Participants joined the study voluntarily, and they could leave at any time without having to justify their decision.

[3] https://wit.ai/

[4] https://espeak.sourceforge.net

[5] https://readyplayer.me/

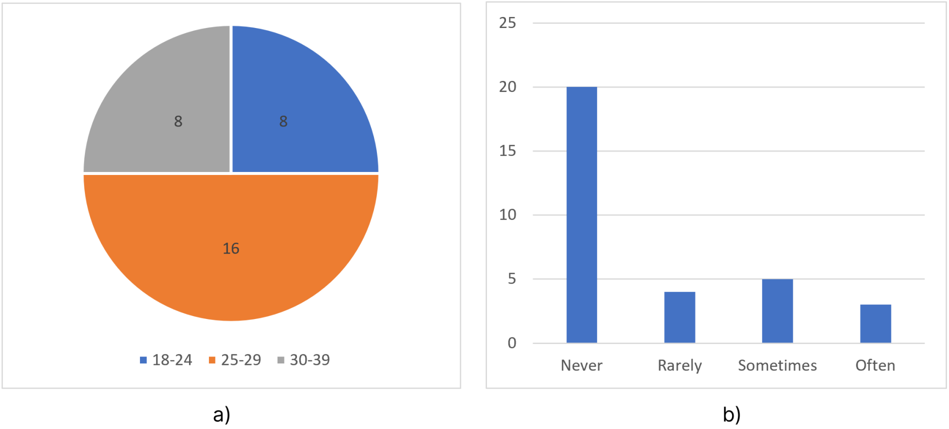


Figure 3: (a) Age and (b) experience with VR device of participants

5.1 Goal

The goal of the study is the following (Basili and Rombach, 1988):

Experiment with the interaction in a CSCW VR environment **in order** to evaluate the impact of the use of a task-oriented multimodal conversational interface when compared with a menu-based interface **with respect to** usability **from the point of view** of end-users **in the context of** a meeting management setting.

Starting from this goal we formulated the following Research Question (RQ):

> RQ
>
> *When immersing in a virtual environment, is there a difference in usability whether using a multimodal conversational interface or a menu-based interface?*

5.2 Participants

We involved 32 participants from the University of Salerno. There were 22 males and 10 females. Their age was distributed according to Fig.3(a) while their previous experience with the use of VR with head-mounted devices is summarized in Fig.3(b).

5.3 Tasks

We identified the following two tasks:

- *T1*: create a new room and change the scenario room;
- *T2*: schedule a meeting.

In particular, for task T1, we asked participants to perform these activities: "Create an interview room for 20 people called Job Interview" and "Change the environment in a meeting room"; for task T2: "Schedule a meeting with the development team on December 17th from seven to eight o'clock".

5.4 Study design

Participants performed two tasks, namely T1 and T2 described in the previous section. They were randomly grouped into two groups named Group1 and Group2, except for participants experts in VR use, who were equally distributed. All performed two tasks T1 and T2, and were exposed to two treatments: Menu, when the user interacts with a menu-based interface, and Chatbot, when the interaction occurs with Muxi. To avoid bias due to task ordering we adopted a crossover design (Vegas et al., 2016), where the Menu treatment is provided in T1 for Group1 and in T2 for Group2. Vice versa for the Chatbot treatment. Figure 4 shows the study design.

5.5 Variables and Measurements

As *independent variable*, we considered the two treatments Menu and Chatbot.

To assess the two considered interaction modalities we measured the following *dependent variables* representing usability, grouped in performance measures and user perceptions.

Performance measures, measuring performance in terms of:

- *Efficiency*: Time. It measures the time to perform a task.
- *Effectiveness*: Errors. It measures the number of errors committed during the execution, e.g., the number of times the chatbot failed in the Chatbot treatment and the number of user errors in the Menu treatment.

Users' perceptions, measuring user satisfaction in terms of:

- *Cognitive load*, measured through the NASA Task Load Index (NASATLX) questionnaire (Hart, 2006). It consists of six subscales representing the following factors: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. To make it simple, the NASA (raw) TLX version was applied (Hart, 2006), where each sub-scale can get a score that goes from 0 (low) to 100 (high), except Performance that goes from 0 (Good) to 100 (Poor). The final score is the mean of the individual scores; a smaller score means lower cognitive load experienced by people performing a task.
- *Perceived usability*, measured through the System Usability Scale (SUS) questionnaire (Bangor et al., 2008), a usability tool based on a ten-item survey, a widely used method. The SUS was evaluated following the standard approach: all items were rated on a 1–5 Likert scale, all items with positive wording were transformed as (xpos-1) (adjusting them to 0–4) while all items with negative wording were transformed as (5-xneg) (reversing the scale and adjusting to 0–4). The SUS score was then computed by summing all items and multiplying them by 2.5, resulting in a final score on a scale of 0–100.

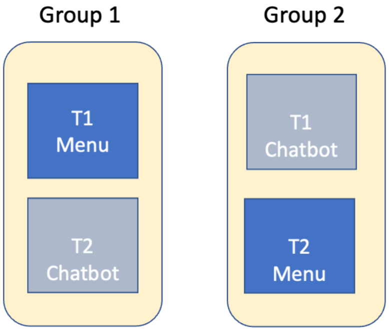


Figure 4: Study Design

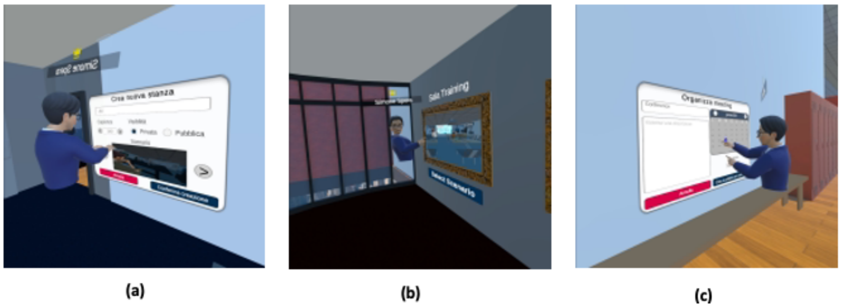


Figure 5: The MetaCUX interface for the Menu treatment

5.6 Experimental objects

The Chatbot and Menu treatments are conducted in a virtual environment ad-hoc developed and hosted on the MetaCUX platform, exposing the two interfaces depicted in Fig. 2 and 5, respectively. To interact with the former interface users have to speak with the avatar while they have to adopt controllers for performing the tasks when using the latter.

5.7 Procedure

During the experiment, we followed this procedure:

- *Recruitment.* Participants were recruited at the University of ***. After collecting their consensus form, they filled in a pre-test questionnaire, collecting their demographic information and their experience concerning of use of VR technology.
- *Assignment.* Considering the results of the pre-test we randomly distributed the participants with and without previous experience in VR use in the two groups (Wohlin et al., 2012).
- *Training.* Participants received training on how to use the Meta Quest 2 device and its controllers to engage with the virtual environment. The duration of this training session was twenty minutes.

- *Operation.* The participants individually performed the two tasks according to the study design in Fig.4. At the end of each task, they filled in the questionnaires described in Section 5.5. A Post-task single open question is also proposed: *What are the positive and negative aspects of this mode of interaction?*

5.8 Analysis procedure

We aim to assess the effect of one factor - the interaction modality - on the dependent variables Effectiveness and Efficiency. For quantitative variables, a t-test for normally distributed data or, otherwise, a Wilcoxon Signed Rank Test may be adopted as our factor has only two levels. We also measure the effect size by using Cohen's distance in the case of normally distributed data and Cliff's (Cliff, 2014) effect size, otherwise. We fixed the significance level (α) at 0.05.

Concerning the user perception analysis related to cognitive load and perceived usability, since all questions are measured on a Likert ordinal scale we analyze the questionnaire results by analyzing the median and adopting nonparametric tests.

6 Results

6.1 Performance analysis

Descriptive statistics of the dependent variables by Treatment are reported in Table IV. It is possible to see that the Chatbot-based interaction modality reached better time performance (Median=40.5 sec.) when compared to the menu-based interaction (Median=61 sec.). This is confirmed by the statistical analysis: Time was not normally distributed, thus, we applied the Wilcoxon Signed Rank Test and it resulted in a statistically significant difference with a large negative effect size (see Tab. V). Similarly, results are also reached by Errors. The boxplot in Fig.6 shows the boxplots of the time by analyzing the two tasks, which confirms this trend for both tasks. A similar trend occurs also for Errors, see Fig.7.

Table IV: Some descriptive statistics for the dependant variables

Variable	**Treatment**	**Mean**	**SD**	**Min**	**Median**	**Max**
Time (sec.)	Chatbot	43.63	16.25	27	40.5	99
	Menu	59.69	15.52	31	61	98
Errors	Chatbot	0.59	1.01	0	0	4
	Menu	2	1.19	0	2	4

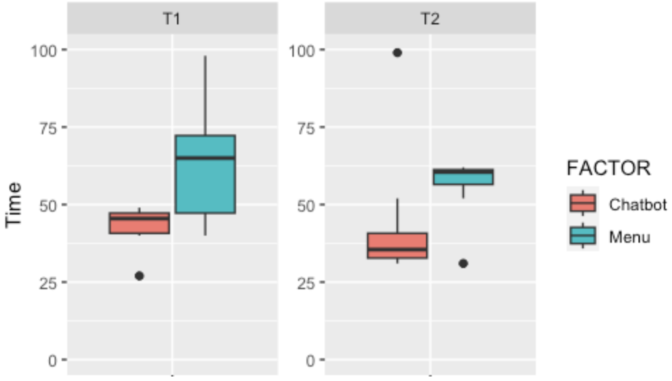


Figure 6: Boxplot of the Times to accomplish the two tasks

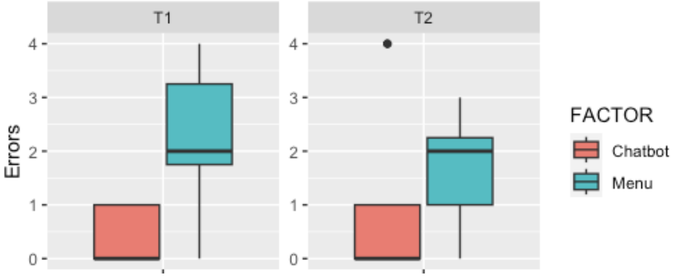


Figure 7: Boxplot of the Errors to performed during the two tasks

Table V: Results of statistical analysis of quantitative variables

Variable	**p-value**	**Cliff's delta**
Time	8.303e-06	-0.6484375 (large)
Error	2.69e-06	-0.6484375 (large)

Table VI: Median for NASA-TLX questionnaire (Lower values have less load and best performance)

Variable	Chatbot	Menu
Mental Demand	35	70
Physical Demand	20	65
Temporal Demand	25	50
Frustration	15	35
Performance	15	70
Effort	30	60
NASA TLX total score	33.33	58.33

Table VII: SUS Score

Variable	Chatbot	Menu
SUS score	82.5	48.75

6.2 User perception analysis

6.2.1 Cognitive load

As shown in Table VI, all the median of the NASA-TLX scales related to the Chatbot treatment always outperforms the Menu treatment ones. This is confirmed by the statistical analysis (Table VIII): users perceived less cognitive load in all the scales when using chatbots with a large negative effect size.

6.2.2 Perceived usability

We assessed the SUS score for each participant and treatment. Fig. 8 shows the results of the single questions. Globally, the Chatbot interface is always better perceived (note that questions with pair numbers have been reversed). The SUS score is 54.06 and 82.19 for the Menu and Chatbot treatments, respectively, as shown in Table VII.

Table VIII: Results of statistical analysis of user perception variables (Chatbot vs Menu)

Variable	p-value	Cliff's delta
Mental Demand	3.168e-08	-0.79 (large)
Physical Demand	1.635e-07	-0.75 (large)
Temporal Demand	3.646e-11	-0.9492188 (large)
Frustration	1.49e-05	-0.6152344 (large)
Performance	6.007e-07	-0.71875 (large)
Effort	2.535e-07	-0.7363281 (large)
Total	6.695e-10	-0.8955078 (large)
SUS score	1.573e-07	0.7617188 (large)

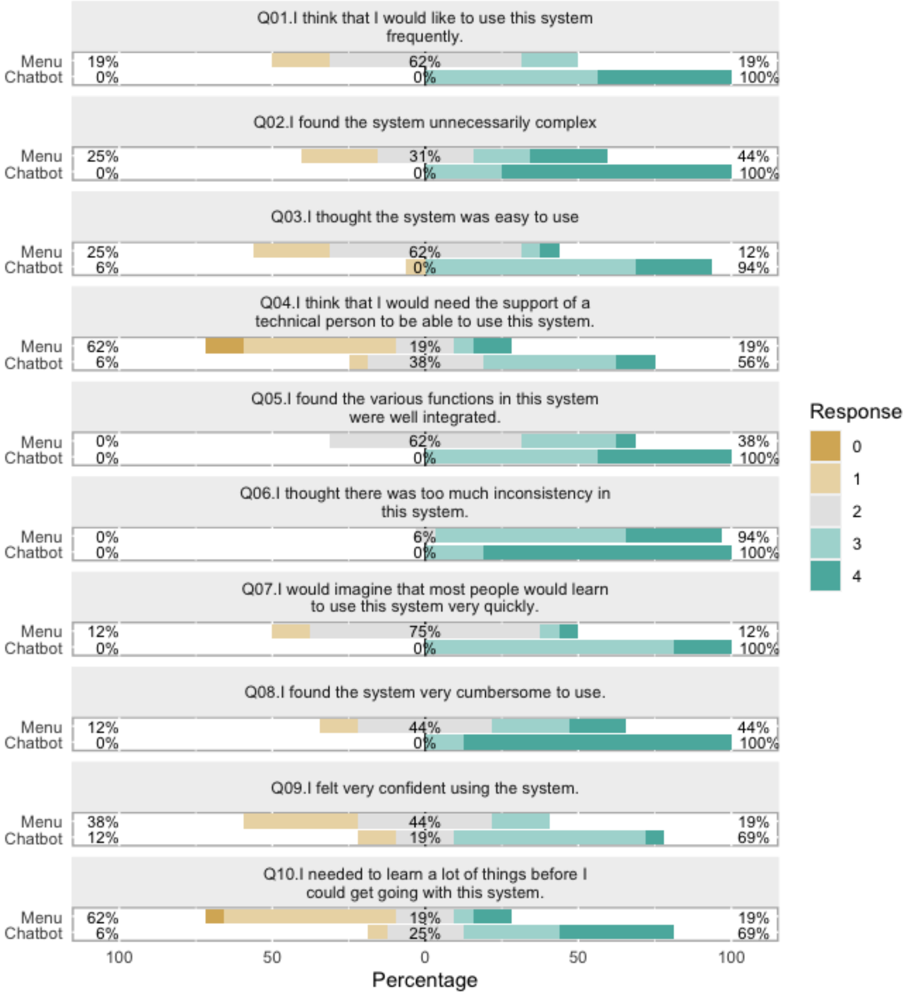


Figure 8: SUS Likert scores summarized for both the treatments (negative answers are reversed)

6.3 Discussion

Performance measures revealed that to support task-oriented activities in a CSCW VR environment a multimodal conversational interface may represent a good alternative to the menu-based interface based on controllers. Indeed, the new interface performed better in both tasks, as shown in Figures 6 and 7.

According to Bangor et al. (2008), a SUS score higher than 77.8 is in the fourth quartile. This indicates that Muxi, which scored 82.5, has no relevant usability problem when compared with the menu approach of the previous version of the system, which scored 48.75.

NASA TLX scores highlight that the cognitive load was significantly lower for all the subscales. In particular, Physical and Mental demand, and Performance seem more affected by the different interaction types.

The performance of NLP models has been further assessed in the experiment obtaining good results: Muxi user performance and perceptions were far better than the menu ones with a large effect size. This may suggest that the proposed UCD development approach has successfully met the users' needs.

Concerning the open questions related to the chatbot experience, an expert participant wrote: *"Using an avatar allowed me to do what was asked of me quickly and in a short time. In this way, however, I had less interaction with the virtual environment in general."*. A non-expert user commented *"It was easier to use your voice rather than the headset controllers to experiment. I did the task much faster."*

Comment of an inexpert user related to the menu interface: *"The negative aspect is that this mode of interaction, for those who are less accustomed to the use of viewers or technology in general, can cause frustration."*

These comments may indicate that interfaces based on chatbots may be particularly useful for non-expert users to start to familiarize themselves with the environment. Only two experts participated in the experiment, which had the same trend for all the factors except for Performance and Physical effort: they both scored better on the menu Performance than Chatbot and signaled a reduced Physical effort in the Menu case.

6.4 Threats to validity

To address the threats that may affect the validity of our findings we follow the recommendations by Wohlin et al. (2012).

External validity. We conducted our experiment with a few participants having different abilities in the use of the technology, which may pose a threat to the interaction of selection and treatment (i.e., the findings may not apply to all the people with the same skills). We tried to limit this threat by uniformly distributing the most skilled participants between the two groups. Furthermore, the adopted multimodal conversational interface was designed to be appealing and easy to use, but we acknowledge that our findings may not apply to a different setting *interaction of setting and treatment*. The selected tasks were also associated specifically with the MetaCUX environment. We formulated the two tasks in such a way as to have about the same duration, to avoid different cognitive loads in the Menu treatment.

Internal validity. The voluntary participation may introduce a *selection threat* because volunteers are usually more motivated than the whole population.

Construct validity. We mitigated the social threats. In particular, participants have not evaluated (*evaluation apprehension*), and we did not communicate the experiment's aim to avoid influencing their opinion (*Experimenter expectancy*).

Conclusion validity. The threat of violated assumptions of statistical tests may exist. To mitigate this threat we adopted non-parametric tests and distances for data that was not normally distributed and qualitative data.

7 Conclusion

In this paper, we described the User Centered Design process we adopted to create the task-oriented multimodal conversational interface in a CSCW VR environment named MetaCUX. A vocal chatbot embodied by an avatar is enhanced by an interactive board for supporting meeting management by easing interaction concerning the original menu-based interface and showing additional content. The empirical investigation involving 32 users aimed to compare the usability of a menu-based interface with the proposed multimodal interface. Both the performance and user perception analyses revealed that performance and user perceptions of the multimodal modalities obtained better results in all the considered aspects. Thus, the proposed multimodal interface may constitute a valid solution for designing task-oriented chatbots in CSCW VR environments.

References

Bangor, A., P. T. Kortum, and J. T. Miller (2008): 'An empirical evaluation of the system usability scale'. *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594.

Barra, P., A. A. Cantone, R. Francese, M. Giammetti, R. Sais, O. P. Santosuosso, A. Sepe, S. Spera, G. Tortora, and G. Vitiello (2023a): 'MetaCUX-a multi-user, multi-scenario environment for a cooperative workspace'. In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. pp. 1–3.

Barra, P., A. A. Cantone, R. Francese, M. Giammetti, R. Sais, O. P. Santosuosso, A. Sepe, S. Spera, G. Tortora, and G. Vitiello (2023b): 'MetaCUX: Social Interaction and Collaboration in the Metaverse'. In: *IFIP Conference on Human-Computer Interaction*. pp. 528–532.

Basili, V. R. and H. D. Rombach (1988): 'The TAME Project: Towards Improvement-Oriented Software Environments'. *IEEE Transactions on Software Engineering*, vol. 14, no. 6, pp. 758–773.

Berger, A. and S. Guda (2020): 'Threshold optimization for F measure of macro-averaged precision and recall'. *Pattern Recognition*, vol. 102, pp. 107250.

Cassell, C., G. Symon, and N. King (2004): *Using Templates in the Thematic Analysis of Text*, pp. 257 – 270. SAGE Publications London.

Cliff, N. (2014): *Ordinal methods for behavioral data analysis*. Psychology Press.

Crovari, P., S. Pidó, F. Garzotto, and S. Ceri (2020): 'Show, don't tell. reflections on the design of multi-modal conversational interfaces'. In: *International Workshop on Chatbot Research and Design*. pp. 64–77.

Das, K. and C. W. Borst (2010): 'An evaluation of menu properties and pointing techniques in a projection-based VR environment'. In: *2010 IEEE Symposium on 3D User Interfaces (3DUI)*. pp. 47–50.

David, B., R. Chalon, B. Zhang, and C. Yin (2019): 'Design of a collaborative learning environment integrating emotions and virtual assistants (chatbots)'. In: *2019 IEEE 23Rd international conference on computer supported cooperative work in design (CSCWD)*. pp. 51–56.

De Lucia, A., R. Francese, I. Passero, and G. Tortora (2009): 'Development and evaluation of a system enhancing Second Life to support synchronous role-based collaborative learning'. *Softw. Pract. Exp.*, vol. 39, no. 12, pp. 1025–1054.

Følstad, A., M. Skjuve, and P. B. Brandtzaeg (2019): 'Different chatbots for different purposes: towards a typology of chatbots to understand interaction design'. In: *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*. pp. 145–156.

Francese, R., A. Guercio, V. Rossano, and D. Bhati (2022): 'A Multimodal Conversational Interface to Support the creation of customized Social Stories for People with ASD'. In: P. Bottoni and E. Panizzi (eds.): *AVI 2022: International Conference on Advanced Visual Interfaces, Frascati, Rome, Italy, June 6 - 10, 2022*. pp. 19:1–19:5, ACM.

Hart, S. G. (2006): 'NASA-task load index (NASA-TLX); 20 years later'. In: *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. pp. 904–908.

Kontio, J., J. Bragge, and L. Lehtola (2008): *Guide to Advanced Empirical Software Engineering*, Chapt. The Focus Group Method as an Empirical Tool in Software Engineering, pp. 93–116. Springer.

Lipari, N. G. and C. W. Borst (2015): 'Handymenu: Integrating menu selection into a multifunction smartphone-based VR controller'. In: *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. pp. 129–132.

Mohamad Suhaili, S., N. Salim, and M. N. Jambli (2021): 'Service chatbots: A systematic review'. *Expert Systems with Applications*, vol. 184, pp. 115461.

Mundt, M. and T. Mathew (2020): 'An evaluation of pie menus for system control in virtual reality'. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. pp. 1–8.

Nguyen, Q. N., A. Sidorova, and R. Torres (2022): 'User interactions with chatbot interfaces vs. Menu-based interfaces: An empirical study'. *Computers in Human Behavior*, vol. 128, pp. 107093.

Pick, S., A. S. Puika, and T. W. Kuhlen (2017): 'Comparison of a speech-based and a pie-menu-based interaction metaphor for application control'. In: *2017 IEEE Virtual Reality (VR)*. pp. 381–382.

Ren, R., M. Zapata, J. W. Castro, O. Dieste, and S. T. Acuña (2022): 'Experimentation for Chatbot Usability Evaluation: A Secondary Study'. *IEEE Access*, vol. 10, pp. 12430–12464.

Sebillo, M., G. Vitiello, and M. De Marsico (2009): *Multimodal Interfaces*, pp. 1838–1843. Boston, MA: Springer US.

Trappey, A. J., C. V. Trappey, M.-H. Chao, and C.-T. Wu (2022): 'VR-enabled engineering consultation chatbot for integrated and intelligent manufacturing services'. *Journal of Industrial Information Integration*, vol. 26, pp. 100331.

Vegas, S., C. Apa, and N. Juristo (2016): 'Crossover Designs in Software Engineering Experiments: Benefits and Perils'. *IEEE Transactions on Software Engineering*, vol. 42, no. 2, pp. 120–135.

Wang, S., Q. Yan, and L. Wang (2023): *Task-oriented vs. social-oriented: Chatbot communication styles in electronic commerce service recovery*, pp. 1–33. Springer.

Wang, Y., Y. Hu, and Y. Chen (2021): 'An experimental investigation of menu selection for immersive virtual environments: fixed versus handheld menus'. *Virtual Reality*, vol. 25, pp. 409–419.

Wohlin, C., P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén (2012): *Experimentation in software engineering*. Springer Science & Business Media.

Xie, Q., W. Lu, Q. Zhang, L. Zhang, T. Zhu, and J. Wang (2023): 'Chatbot Integration for Metaverse - A University Platform Prototype'. In: *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. pp. 1–6.

Zhao, Y., J. Jiang, Y. Chen, R. Liu, Y. Yang, X. Xue, and S. Chen (2022): 'Metaverse: Perspectives from graphics, interactions and visualization'. *Visual Informatics*, vol. 6, no. 1, pp. 56–67.