Multimodal Technologies and Interaction

Article

A Study of NLP-Based Speech Interfaces in Medical Virtual Reality

Mohit Nayak , Jari Kangas and Roope Raisamo *

TAUCHI Research Center, Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland; mohit.nayak@tuni.fi (M.N.); jari.a.kangas@tuni.fi (J.K.)
* Correspondence: roope.raisamo@tuni.fi

Abstract: Applications of virtual reality (VR) have grown in significance in medicine, as they are able to recreate real-life scenarios in 3D while posing reduced risks to patients. However, there are several interaction challenges to overcome when moving from 2D screens to 3D VR environments, such as complex controls and slow user adaptation. More intuitive techniques are needed for enhanced user experience. Our research explored the potential of intelligent speech interfaces to enhance user interaction while conducting complex medical tasks. We developed a speech-based assistant within a VR application for maxillofacial implant planning, leveraging natural language processing (NLP) to interpret user intentions and to execute tasks such as obtaining surgical equipment or answering questions related to the VR environment. The objective of the study was to evaluate the usability and cognitive load of the speech-based assistant. We conducted a mixed-methods within-subjects user study with 20 participants and compared the voice-assisted approach to traditional interaction methods, such as button panels on the VR view, across various tasks. Our findings indicate that NLP-driven speech-based assistants can enhance interaction and accessibility in medical VR, especially in areas such as locating controls, easiness of control, user comfort, and intuitive interaction. These findings highlight the potential benefits of augmenting traditional controls with speech interfaces, particularly in complex VR scenarios where conventional methods may limit usability. We identified key areas for future research, including improving the intelligence, accuracy, and user experience of speech-based systems. Addressing these areas could facilitate the development of more robust, user-centric, voice-assisted applications in virtual reality environments.

Keywords: virtual reality; natural language; NLU; NLP; human–computer interaction; medical; speech recognition; interaction techniques

Academic Editor: Arun K. Kulshreshth

Received: 25 March 2025
Revised: 12 May 2025
Accepted: 22 May 2025
Published: 26 May 2025

Citation: Nayak, M.; Kangas, J.; Raisamo, R. A Study of NLP-Based Speech Interfaces in Medical Virtual Reality. *Multimodal Technol. Interact.* **2025**, *9*, 50. https://doi.org/10.3390/mti9060050

1. Introduction

Virtual reality (VR) technologies are increasingly employed in medicine for their ability to render immersive, three-dimensional environments that enhance spatial cognition, anatomical comprehension, and procedural understanding [1,2]. In diagnostic imaging and surgical planning, VR supports precise interpretation and intervention by enabling detailed visualization of complex anatomical data [3–7]. In medical education, interactive and risk-free VR simulations have been shown to improve learning outcomes, spatial awareness, and skill acquisition, underscoring VR's expanding role in both clinical and educational settings [5,8–10].

Despite the potential of VR in medical applications, there are interaction challenges. Traditional input methods, such as controllers, often fail to replicate the precision and intuitiveness of familiar 2D interfaces like a mouse and keyboard, making tasks such as typing and information retrieval cumbersome in 3D environments [11]. Cluttered interfaces and overly complex controls further increase cognitive load, limiting smooth operation and user experience. The absence of standardized interaction paradigms across VR systems complicates skill transfer and limits usability [12]. These challenges arise from limitations in locating and operating tools in VR, coupled with the unfamiliarity of interacting with virtual objects, which contrasts with habitual 2D interface use [13].

Speech-based interaction offers a promising alternative to address these challenges. Voice commands provide a hands-free method to streamline interaction, reduce cognitive burden, and enhance accessibility in VR systems [14]. However, their effectiveness depends on the integration of robust natural language processing (NLP) techniques to ensure accurate speech recognition and intuitive user interfaces tailored to specific use cases. Implementing such systems can significantly improve task efficiency and user satisfaction, paving the way for broader adoption of VR in healthcare [14–18].

This research focused on assistive technologies in VR, particularly the use of spoken voice commands, and it evaluated their usability in medical virtual reality systems. Chatbot-based assistive technologies are not new; they have demonstrated significant benefits across various industries [19]. Recent advancements in natural language processing (NLP), including large language models (LLMs) for generative text, transformer models, and later BERT and other models [20,21], have improved text classification, intent recognition, and information retrieval. Combined with advanced speech-to-text technologies, these innovations enable efficient human–machine interaction, making VR systems more intelligent and adaptive [17,22–25].

The research specifically explored speech commands for tool selection and answering queries in surgical planning settings. To ground our study in a real context, we used a maxillofacial implant surgery use case, which is an example of surgical planning with tool selection for dental implants. This study aimed to design an intelligent speech assistant and evaluate its impact on usability and cognitive load in these kinds of real-time medical applications. We studied the following research questions:

RQ1: How does the use of intelligent speech interfaces affect usability metrics such as ease of control, comfort, accuracy of commands, satisfaction in response, finding controls, learning and adapting, recovery from mistakes, and naturalness and cognitive load metrics like physical demand, mental demand, temporal demand, performance, effort, and frustration?

RQ2: What are the advantages, limitations, expectations, and general opinion of speech interfaces in VR for medical purposes?

For this study, student participants were chosen to allow rapid, scalable evaluation of core interaction mechanisms, uncovering their expectations of modern speech interfaces before engaging medical professionals. This approach helped in refining usability and cognitive load factors without introducing domain expertise bias, and it allowed early-stage validation under practical resource constraints.

2. Background

Several prior studies have explored the advantages of VR in the medical field, demonstrating simulation of complex medical procedures, improving diagnostics, and enhancing surgical planning [3,7,26]. For example, in surgical planning, VR has been shown to improve the accuracy of tumor localization in liver resection [27,28]. Similarly, in head cancer and neck cancer surgery, VR allows a more detailed assessment of tumor extent and surrounding anatomical structures, leading to improved surgical planning and better oncological outcomes [29]. In skull base neurosurgery, the complex anatomy and proximity of critical neurovascular structures make precise planning possible [30]. Each of these studies proves the capabilities of XR (extended reality) technologies such as virtual reality (VR) and augmented reality (AR) in the medical domain.

VR applications suffer from many interaction challenges, as discussed in the Introduction. To address these issues, various studies have explored integrating additional modalities such as haptics [31–33], speech [15,34,35], gesture-based controls [32,36], and gaze [37] to enhance user interaction. These interaction problems are also solved by adopting a multimodal approach to create more immersive, intuitive, and responsive VR experiences [38]. Among these modalities, speech is regarded as the most natural because it mirrors the way humans inherently communicate with one another.

Assistive technologies such as chatbots and speech-based systems such as Alexa have become fairly common in daily activities and various personal and industrial domains [19]. This has been possible through various state-of-the-art NLP techniques enabling efficient speech systems [17,22–25]. Previous research has demonstrated the specific utility of speech-based systems in various medical use cases. An example was the use of a medical decision support system that integrated real-time speech interfaces with deep neural networks (DNNs) to predict suitable therapies based on patient medical history. It reduced manual data entry time and allowed more focus on diagnosis and patient care [17].

Similarly, speech recognition has been leveraged to control virtual tools in immersive environments, improving efficiency, realism, and user engagement through NLP techniques such as intent classification [35]. In VR, voice-controlled mode switching has been shown to be more intuitive and satisfactory than traditional button-based methods, offering a coherent interaction experience for healthcare professionals [34]. Speech interfaces have also demonstrated significant utility in high-pressure medical settings, such as surgical environments, where they improve task execution and retention of clinical skills [18]. For training applications, the DIET (dual intent entity transformer) model has been employed to classify a wide range of intents, creating immersive and effective medical learning environments [15]. Furthermore, natural language understanding has been advanced in virtual standardized patients (VSPs) by integrating speech-recognition and hybrid AI techniques, enabling enhanced history-taking skills and improving simulation fidelity [16]. Another study that used large language models to navigate around medical virtual reality produced various positive remarks [39,40], although using LLM had an observed latency of 3–4 s and 1.5–1.75 s, depending on the tasks, which is not suitable for flawless medical interactions. These studies provide insights into the usage of machine learning principles in medical domains using speech.

However, there is a lack of research exploring the application of speech interfaces in surgical planning scenarios, where such systems could function as virtual assistants to support secondary tasks in a real-time environment. These tasks may include retrieving surgical tools or providing context-specific information about the virtual environment, thereby improving workflow efficiency and user engagement, especially for new users in VR. Even as users switch between the VR systems with different controls, the users can interact with any system in the same way, using natural language, as long as they have knowledge about the tools, which would be expected from medical professionals. A related study investigated speech interfaces for tool switching, but that was limited to static commands [34]. The reliance on static commands introduces challenges related to memory recall, potentially disrupting the workflow and increasing cognitive demands during critical tasks. Furthermore, the lack of adaptive or context-aware design in such systems requires predefined terminology for tools, which may differ from the naming conventions used in real-world surgical practices.

This research addressed these limitations by integrating a dynamic, context-aware verbal assistant capable of interacting naturally with users and adapting to the specific terminology and requirements of the virtual environment. By doing so, it aimed to reduce cognitive load, improve the usability of the system, and to bridge the gap between existing static command-based interfaces and the flexible and intuitive needs of surgical planning and training scenarios.

3. Materials and Methods

The system design prioritizes natural language understanding, so that it can be used in all VR systems as long as the user knows about the tools to some extent, which is to be expected from medical professionals. Although large language models (LLMs) represent state-of-the-art techniques for handling diverse instructions and generate curated answers, they have several limitations, like limited customizability [41], lack of transparency [42], and latency [39,40]. Small delays can significantly hinder user experience in a negative way [43]. Therefore, other approaches were used, such as using intent recognition [15,44]. Focusing more on interaction, cloud-based pre-trained models were chosen, such as speech service from Azure [45].

3.1. System Architecture

The system architecture is designed to integrate natural language processing (NLP) capabilities with a VR interaction framework. Built on the Unity platform and Oculus Meta Quest 3 VR headsets, the architecture features modular components to support overall speech-driven interactions, intent recognition, and real-time user feedback. Figure 1 shows the key architectural elements, including the NLP pipeline, which consists of speech-to-text (STT), intent recognition, question answering, and text-to-speech (TTS) components that facilitate bidirectional communication between users and the VR system:

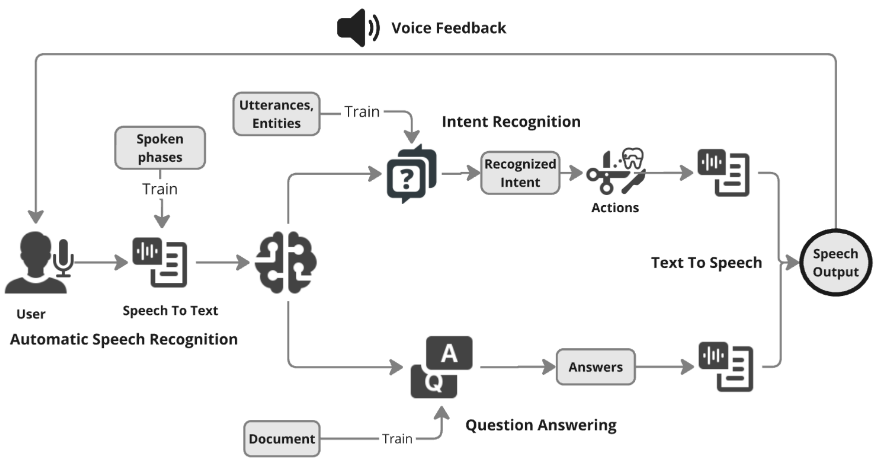


Figure 1. System architecture of usage of NLP-based speech system and its integration with VR environment.

Upon user speech input, audio is captured through the VR headset's microphone. The STT module processes the audio signal and generates corresponding textual data. This text is subsequently analyzed by the intent recognition module to classify the user's command or to detect questions intended for the question answering subsystem. This change between intent recognition and question answering is effected by the user through a toggle button, as displayed in Figure 2 "Help On". If a command is identified, an appropriate system action is triggered by mapping the recognized intent to a corresponding function within the VR application. In the case of a detected question, the question answering module generates an appropriate response, which is then delivered to the user through visual feedback in the form of text (Figure 2) and synthesized speech via the TTS module heard through the speakers of headsets.

A feedback mechanism is implemented, whereby a logger tracks speech recognition outputs, system responses, and error handling, providing real-time feedback to users in the form of audio and visual feedback.

(**a**) (**b**)

Figure 2. (**a**) The button panel as the control interface. (**b**) The empty speech logger.

The architecture was designed with a focus on evaluating usability and cognitive load in an exploratory study of speech models. Therefore, fewer training examples were used, since the sentences were small and this number satisfactorily generated the outputs during the development stages. Another reason was to minimize costs and resources, and limited emphasis was placed on fine-tuning NLP models. This approach prioritized feasibility and preliminary insights over extensive optimization.

3.2. NLP Development

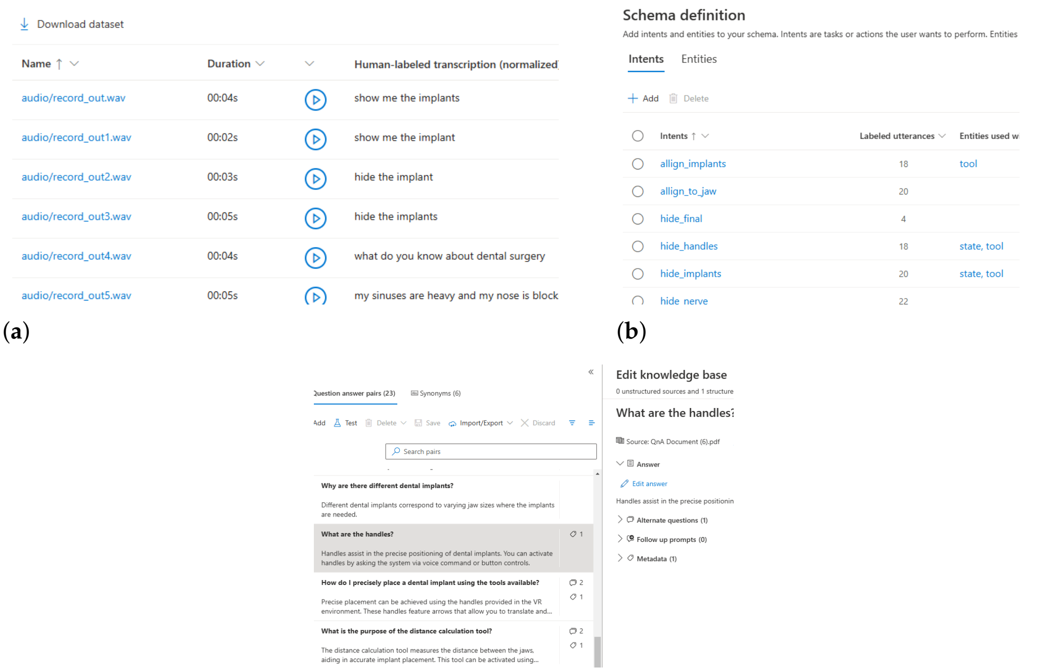
Natural language processing components were developed using Azure cognitive services, focusing on STT, intent recognition, QA functionalities, and text-to-speech (TTS). Various speech services were tested, and Azure was selected for its minimal latency, due to integration within the same Azure environment.

The pre-trained real-time STT model of Azure AI speech services was fine-tuned using 20 training examples from people with different tones and accents to improve recognition of medical terminology and diverse vocal patterns, which was enough for the limited amount of special terminologies we used, like “sinuses”, “jawline”, “X-ray”, “flashlights”, and “Dental Implants”. The rest were common and could easily be handled without training, like “handles”, “undo”,“redo”,“show/hide”, etc. This training resolved issues such as misinterpretation of terms like “sinuses” (previously recognized as “cinuses”) and “X-Ray” (misheard as “exray”), ensuring accurate processing of tools and commands such as “Turn on/off the X-Ray Flash Light”, which was seen commonly while testing the STT. These training examples were in the form of voice recording and its corresponding transcripts. Some examples of training data are “What is a X-ray Flashlight?”, “Turn off the handles”, “Can you tell me about the sinuses?”, “how to use handles?”. These sentences were important to let the training model know how a particular word sounds, thus avoiding misinterpretations.

While speech-to-text (STT) is crucial for accurately recognizing spoken words, correctly interpreting the meaning of the recognized text is equally important. This is where Azure Language Understanding (LUIS) plays a critical role, enabling the system to extract intents and relevant information from the transcribed speech. Azure Language Understanding

(LUIS) was trained with 20 intents, with 15–20 utterances per intent corresponding to different system functionalities. These 20 intents corresponded to all the available features in the project and were necessary to complete the medical tasks. The major challenge here was classification among short sentences that did not have enough information. The sentences “Hide all the dental implants” and “Show all the dental implants”, for example, differed only slightly (“hide” and “show”); these sentences and their related synonyms were contrary to the relatively long sentences used in [15,44]. An iterative approach was used with various combinations of synonyms in the training data. Utterances were categorized into entities such as “tools” (e.g., dental implants, X-ray flashlights, etc.) and “states” (e.g., activate, hide), ensuring robust command interpretation and handling of synonyms and phrasing variations, as shown in Figure 3.

A custom knowledge base enabled dynamic and consistent responses to user queries about system functionality and tools. The QA model supported detailed information retrieval, enhancing user support. Although the training examples used were small in quantity, the models were trained in an iterative manner to achieve the satisfactory results that were observed in the development and pilot testing. Figure 3 shows instances of training in Azure; Figure 3a showing examples of transcripts and voice files uploaded in Azure for training, the user speaking while in the VR, and the user interacting using speech.



(**a**) (**b**) (**c**)

Figure 3. Snapshots of NLP instances trained in Azure: (**a**) Adding recorded files and transcripts to train a speech-to-text (STT) model; (**b**) Creating intents in Language Understanding (LUIS), where the first column shows intent names, the second shows the number of utterances, and the third shows the entities used for categorization; (**c**) Snapshot of a knowledge base used for question answering.

3.3. VR Development

The VR environment was developed using Unity upon Planmeca Romexis software for dental implants and was used on the Oculus Meta Quest 3. See Figure 4 for the overall VR environment. The environment included high-fidelity 3D models, such as a skull for dental implant placement and a dental implant tray with adjustable implants of varying sizes [34]. These models provided a realistic representation of medical tools and anatomy, ensuring a practical training experience. In addition to this, the VR system allowed interaction through both speech and a panel with traditional button controls. The button panel was used as a reference control interface. Speech commands provided a hands-free alternative for the same functions and tools as the button panel interface. The control for speech activation was configured using a primary button on a hand controller to reduce external noise.

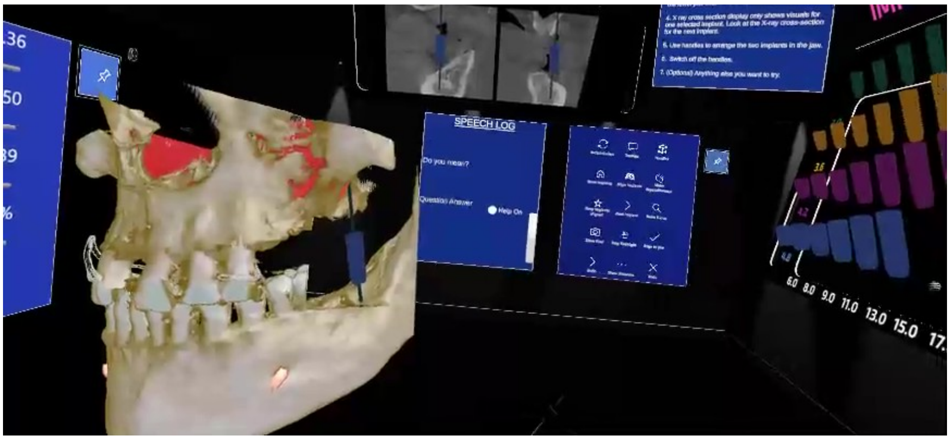


Figure 4. The complete VR setup from a user's point of view. The major components from left to right are the skull model, the X-ray cross-section monitor on top, the speech logger next to the skull model, the button interface panel, the implant tray, and the panel with the list of tasks on top.

A speech interface panel was developed as a logger to provide a visual representation of both spoken words and their recognized intents. This configuration enabled the users to monitor the accuracy of their inputs and generated responses, thereby determining whether their speech was correctly recognized and processed through the system's pipeline. A toggle switch was introduced to switch between question answering and tool selection. Text-to-speech (TTS) converted system responses into audible outputs, improving accessibility, ensuring correct commands, and enhancing the immersive experience. A task panel displayed specific objectives for users to complete, while visual cues and contextual highlights guided interactions with tools and objects. This ensured that the participants, especially those unfamiliar with medical tools, could navigate and perform tasks effectively.

Additionally, error-handling capabilities were added to the VR system, considering the results from the STT and LUIS models. A threshold of a 0.6 confidence score was defined. This was to ask the user if they meant what the system recognized. For example, if the user said "I want the X-Ray Flashlight" but the system recognized it with a low confidence score then the system asked the user for confirmation, such as "Do you mean X-Ray Flashlight?". The visual feedback of the text in the speech logger became very important, in order to know what was being spoken. Another threshold was added for cases in which the confidence is very low in our case, 0.3, when the system asked for a repeat.

3.4. Participants

For exploratory purposes, we selected non-medical student participants, ensuring that the system could be assessed for usability in a manner consistent with everyday technology usage. The study involved 20 participants aged between 21 and 35 years, with an average age of 25.8 years. The group comprised 10 males and 10 females, representing diverse nationalities, of which 19 participants were trying medical VR for the first time and 5 participants had never used VR in any capacity. The participants were briefed about the study's objectives, tasks, and potential risks before providing informed consent. To minimize learning bias, the sequence of interaction modes was randomized, with 10 participants starting with the speech interface and the other 10 beginning with the button interface.

3.5. Experiment Process

Two instructional videos were shown to familiarize the participants with the VR environment. The videos introduced the key components of the VR space, particularly the medical tools that are necessary for dental implant-related activities. These videos demonstrated VR functionalities, including interaction techniques. The participants were trained on operating the hand controllers, including activating speech input via designated controller buttons. They were then given 5–10 min to explore the VR environment freely, manipulating 3D skull models, practicing dental implant placement, and experimenting on some tools provided by the button panel and speech interfaces.

For those who appeared unsure during the freeform try-out phase, minimal guidance was provided to help them perform basic tasks such as moving objects or toggling implant visibility. The participants were also introduced to the task panel and the question answering toggle feature, which they would use to retrieve assistance or complete tasks during the experiment. Figure 5 shows the usage of speech for different operations:

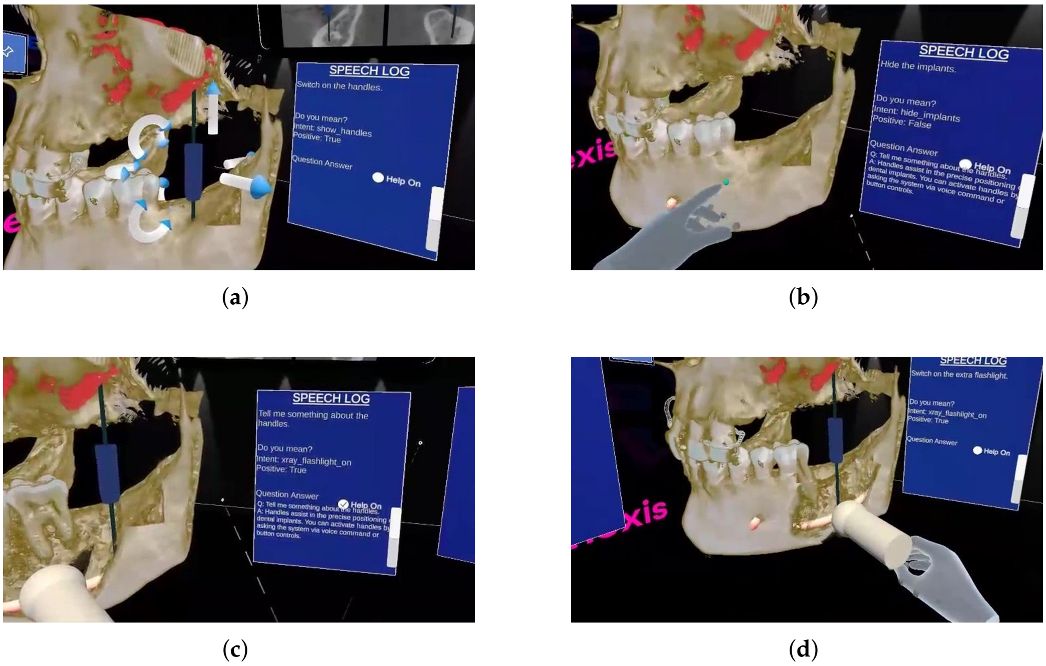


Figure 5. Snapshots of the speech logger with different commands. The spoken text can be seen on top with activated tools under "Do you mean?" and question, answers under the "Question Answering" section: (**a**) Switching on handles for precise positioning of dental implants. (**b**) Hiding dental implants. (**c**) Asking question about handles and getting answers at the bottom of the speech logger. (**d**) Switching on the X-Ray flashlight.

After familiarization, the participants were asked to complete predefined tasks, using both the speech interface and the button interface (Figure 2a). The order of the interfaces was alternated to minimize the order effects. The following seven compulsory tasks were assigned to each user, followed by one optional task:

1. Ask something about the project. For example "tell me about the system".
2. Use the X-Ray flashlight to look for sinuses. (Hint: Search near button of the jawline.)
3. Switch off the X-Ray flashlight.
4. Manually pick up 2 random dental implants from tray on the right, using controller, and place them near the empty spaces in the jawline.
5. If you look at the X-ray cross-section view, it displays a cross-section view for one implant; please change this to display for another implant.

6. Use the handles to adjust the position of the implants in the jaw.
7. Switch off the handles.
8. (Optional) Any other tasks you need to perform.

These tasks simulated the actual tasks that professionals are supposed to complete. At the end of each interaction mode, the participants completed a questionnaire evaluating their experience, specifically focusing on usability and cognitive load metrics.

3.6. Data Collection

The study employed a within-subject design, where each participant interacted with both the speech and button panel interfaces. Usability metrics, including ease of control, comfort, accuracy, satisfaction, and naturalism, were evaluated using a 7-point Likert scale (1 = lowest, 7 = highest). Cognitive load was assessed using the NASA TLX questionnaire, which included components such as mental demand, physical demand, temporal demand, effort, and frustration.

Open-ended questions were also included to gather qualitative feedback regarding challenges, preferences, and suggestions for improving the speech interface. To maintain clarity, the "Performance" component of the NASA TLX scale was reversed during data collection, with higher scores indicating better performance. Adjustments were made during analysis to ensure consistency with the original methodology.

4. Results

The results were categorized into quantitative and qualitative metrics. Quantitative metrics assessed usability and cognitive load. The NASA-TLX framework provides six key dimensions that collectively address various aspects of the cognitive load experienced by the user. However, assessing usability for speech-driven systems requires a different focus compared to traditional system usability evaluation, particularly in light of the evolving user expectations from modern voice interfaces.

Users today are increasingly familiar with interacting with voice systems such as Alexa, Siri, and ChatGPT, which shape their expectations regarding naturalness, responsiveness, accuracy, etc. Accordingly, the usability parameters selected in this study, as outlined in the research questions (RQs), were specifically curated to comprehensively evaluate the unique usability aspects relevant to speech-based interaction systems, distinct with no similar-sounding parameters.

4.1. Quantitative Results

The results for the total, mean (M), and standard deviation (SD) of usability and cognitive load indicate an overall preference for the speech interface in regard to usability but suggest less overall cognitive load with the button panel interface (Table 1, Figure 6):

Table 1. Summary of total, mean, and standard deviation of usability and cognitive load for button and speech interfaces.

	Usability			Cognitive Load		
Mode	**Total**	**Mean**	**Std Dev**	**Total**	**Mean**	**Std Dev**
Button	911	5.69	1.21	288	2.40	1.56
Speech	931	5.82	1.02	301	2.50	1.37

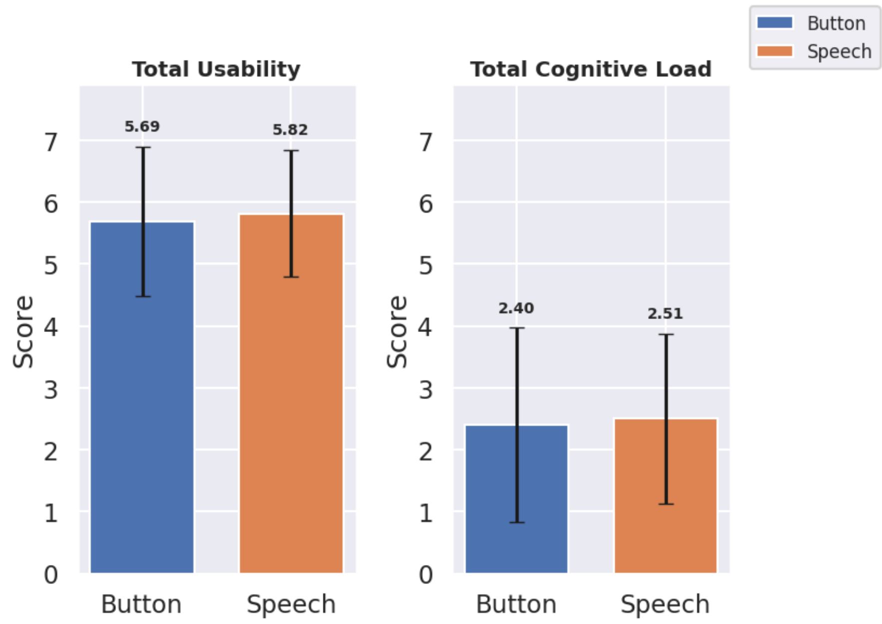


Figure 6. Bar graph for combined usability and cognitive load metrics.

The speech interface achieved an overall similar *total usability* score ($M_S = 914$) in comparison with the button panel interface ($M_B = 911$). The *mean usability score* also demonstrated similar behavior for the speech and button panel interfaces, with scores of ($M_S = 5.71, SD_S = 1.21$) and ($M_B = 5.69, SD_B = 1.21$), respectively. The button panel interface exhibited a lower *mean cognitive load* ($M_B = 2.23, SD_B = 1.69$) than the speech interface ($M_S = 2.34, SD_S = 1.46$), suggesting that speech may slightly increase cognitive load due to recognition inaccuracies. These findings demonstrate the overall results. But for in-depth analysis, the metrics used for usability and cognitive load were analyzed.

4.1.1. Usability Metrics

The usability metrics were evaluated across eight domains for both the button panel and speech interfaces: *Ease of Control, Comfort, Accuracy of Commands, Satisfaction with Response, Finding Controls, Learning and Adapting, Recover from Mistakes, and Natural and Intuitive Use*. The mean scores (M) and standard deviation (SD) provided insights into the user experiences with each interface. Table 2 shows the mean and standard deviation of the button panel and speech interfaces for each usability metric, and Figure 7 shows a bar graph of the same.

Table 2. Usability metrics: mean ± standard deviation for button panel and speech interfaces.

Metric	Button	Speech
Ease of Control	5.70 ± 1.08	5.80 ± 1.01
Comfort	5.75 ± 1.12	6.15 ± 0.93
Accuracy of Commands	5.50 ± 1.24	5.45 ± 0.83
Satisfaction with Response	5.75 ± 1.25	5.60 ± 1.05
Finding Controls	5.55 ± 1.43	6.05 ± 1.05
Learn and Adapt	5.95 ± 1.05	6.05 ± 0.89
Recovery from Mistakes	5.90 ± 1.12	5.80 ± 1.06
Natural and Intuitive Use	5.45 ± 1.43	5.65 ± 1.31

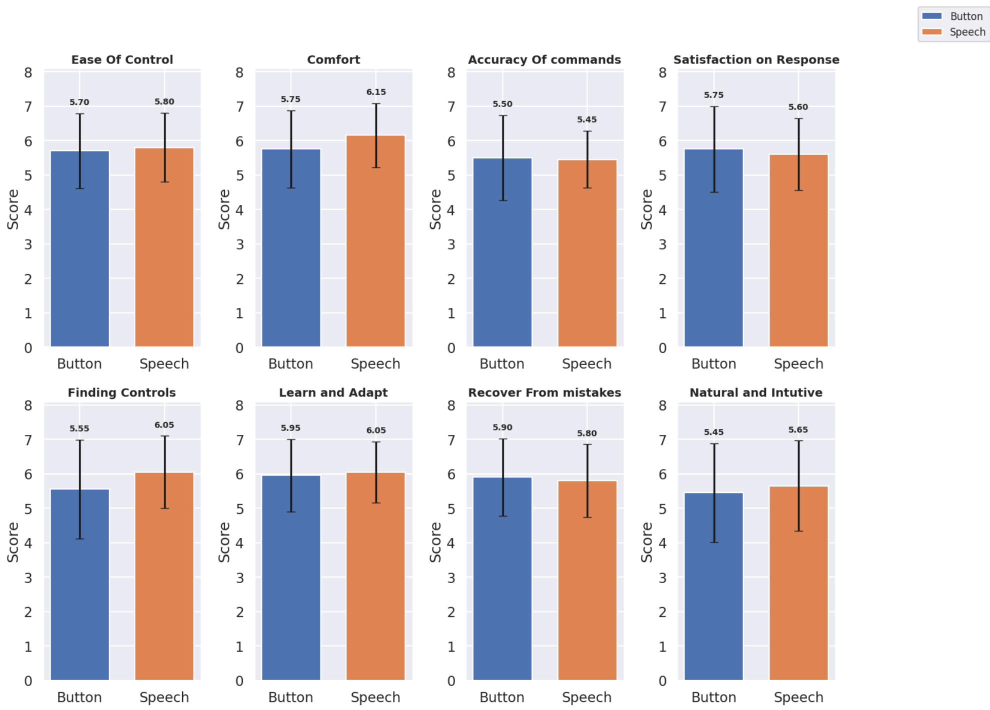


Figure 7. Bar graph showing usability metrics: mean and standard deviation for button panel and speech interfaces.

Based on the updated usability metrics, we observed slight variations between the button panel and speech interfaces across several dimensions:

For *Ease of Control*, the speech interface scored marginally higher ($M_S = 5.80, SD_S = 1.01$) compared to the button panel interface ($M_B = 5.70, SD_B = 1.08$), indicating a slight preference for speech-based control. The *Comfort* ratings also favored the speech interface ($M_S = 6.15, SD_S = 0.93$) over the button panel interface ($M_B = 5.75, SD_B = 1.12$), suggesting that the users found voice commands less physically and mentally demanding. For *Accuracy of Commands*, the button panel interface performed slightly better ($M_B = 5.50, SD_B = 1.24$) than the speech interface ($M_S = 5.45, SD_S = 0.83$). However, in terms of *Satisfaction with Response*, the speech interface received slightly lower ratings ($M_S = 5.60, SD_S = 1.05$) compared to the button panel interface ($M_B = 5.75, SD_B = 1.25$). The speech interface was rated higher in *Finding Controls* ($M_S = 6.05, SD_S = 1.05$) and *Learn and Adapt* ($M_S = 6.05, SD_S = 0.89$), indicating that the users found the speech interface more intuitive for these aspects. However, the button panel interface scored slightly better on *Recovery from Mistakes* ($M_B = 5.90, SD_B = 1.12$) compared to the speech interface ($M_S = 5.80, SD_S = 1.06$). For *Natural and Intuitive Use*, the speech interface scored slightly higher ($M_S = 5.65, SD_S = 1.31$) than the button panel interface ($M_B = 5.45, SD_B = 1.43$), suggesting that the users found the speech interface somewhat more natural and intuitive.

Overall, the usability metrics indicated a nuanced user preference. While the button panel interface excelled in Accuracy of Commands, Satisfaction with Response, and Recovery from Mistakes, the speech interface provided more ease in terms of Ease of Control, Comfort, Finding Controls, Learn and Adapt, and Natural and Intuitive Use.

4.1.2. Cognitive Load Metrics

The cognitive load metrics were evaluated across six domains: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. The mean scores and standard deviations provide insight into the cognitive load associated with each interface (see Table 3 and Figure 8):

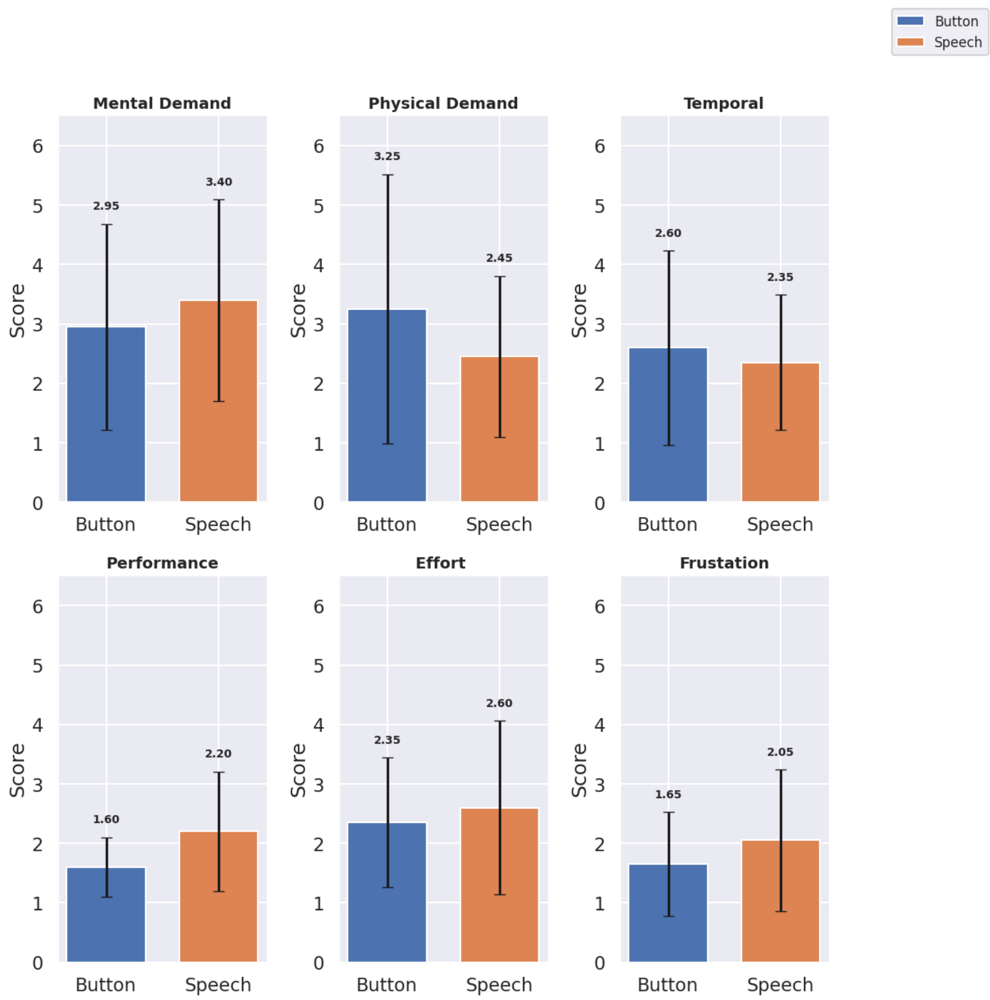


Figure 8. Bar graph showing cognitive load metrics: mean and standard deviation for button panel and speech interfaces.

Table 3. Cognitive load metrics: mean ± standard deviation for button panel and speech interfaces.

Metric	Button	Speech
Mental Demand	2.95 ± 1.73	3.40 ± 1.70
Physical Demand	3.25 ± 2.27	2.45 ± 1.36
Temporal Demand	2.60 ± 1.64	2.35 ± 1.14
Performance	1.60 ± 0.50	2.20 ± 1.01
Effort	2.35 ± 1.09	2.60 ± 1.47
Frustration	1.65 ± 0.88	2.05 ± 1.19

The speech interface showed higher *Mental Demand* ($M_S = 3.40, SD_S = 1.70$) than the button panel interface ($M_B = 2.95, SD_B = 1.73$), though it scored lower on *Physical Demand* ($M_S = 2.45, SD_S = 1.36; M_B = 3.25, SD_B = 2.27$), suggesting that speech may reduce physical effort but increase cognitive processing. *Temporal Demand* was rated similarly across the interfaces, with the button panel interface at $M_B = 2.60, SD_B = 1.64$ and the speech interface at $M_S = 2.35, SD_S = 1.14$, indicating that neither interface added significant time pressure. The button panel interface scored higher on affected *Performance* ($M_B = 1.60, SD_B = 0.50$), reflecting user confidence with button panel interface interactions (here, Performance was evaluated using NASA TLX cognitive load scoring, where a lower score indicates better performance and a higher score means bad performance). The speech interface required more *Effort* ($M_S = 2.60, SD_S = 1.47; M_B = 2.35, SD_B = 1.09$). The speech interface had a higher *Frustration* score ($M_S = 2.05, SD_S = 1.19$) compared to the button panel interface ($M_B = 1.65, SD_B = 0.88$), potentially due to inaccuracies in voice recognition.

4.1.3. Effect of Participant's Past Experience

Out of the 20 participants in the study, 19 were using medical VR systems for the first time, while only 1 had prior experience with such applications. However, when considering general VR usage, 15 participants had used VR in some form before, whereas 5 reported no prior VR experience at all. Due to the imbalance in prior experience with medical VR (19 vs. 1) and the modest size of the dataset, formal statistical testing was not conducted with this. Instead, we performed a preference-based exploratory comparison limited to those participants who had some form of prior VR experience (n = 15), and, later, we took percentages of distribution for normal comparison, allowing for better interpretation within a semi-homogeneous group. Preferences were measured within-subject, for each usability and cognitive load metric. At each decision point, the participant's preferred input method (speech interface or button panel interface) was determined, based on their score for the same metric under both conditions. The goal was to identify which modality was favored per metric by participants who had prior exposure to VR environments.

From Table 4, Participants with prior VR experience showed a preference for speech input in the metrics Comfort and Finding Controls, while button input was favored for the metrics Satisfaction with Response, Recovery from Mistakes, Mental Demand, and Performance. Preferences for other metrics, including Ease of Control, Natural and Intuitive Use, and Effort, were mixed or inconclusive, though such findings remain indicative rather than definitive, due to the data limitations.

Table 4. Percentage-based input method preference by metric and prior VR experience. Preferred input per group is underlined.

Metric	VR: Yes (n = 15)			VR: No (n = 5)		
	Speech	**Button**	**No Pref.**	**Speech**	**Button**	**No Pref.**
Ease of Control	<u>53%</u>	27%	20%	0%	<u>80%</u>	20%
Comfort	<u>47%</u>	20%	33%	<u>60%</u>	40%	0%
Accuracy of Commands	33%	27%	<u>40%</u>	0%	40%	<u>60%</u>
Satisfaction with Response	27%	<u>40%</u>	33%	20%	<u>40%</u>	<u>40%</u>
Finding Controls	<u>47%</u>	20%	33%	<u>40%</u>	20%	<u>40%</u>
Learn and Adapt	20%	27%	<u>53%</u>	40%	0%	<u>60%</u>
Recovery from Mistakes	33%	40%	<u>27%</u>	0%	40%	<u>60%</u>
Natural and Intuitive Use	<u>40%</u>	<u>40%</u>	20%	<u>60%</u>	40%	0%
Mental Demand	13%	<u>47%</u>	40%	20%	<u>60%</u>	20%
Physical Demand	40%	7%	<u>53%</u>	20%	<u>60%</u>	20%
Temporal Demand	27%	13%	<u>60%</u>	0%	0%	<u>100%</u>
Performance	13%	40%	<u>47%</u>	0%	<u>80%</u>	20%
Effort	20%	33%	<u>47%</u>	<u>40%</u>	20%	<u>40%</u>
Frustration	7%	40%	<u>53%</u>	20%	0%	<u>80%</u>

4.1.4. Significance Test

A Shapiro–Wilk test indicated that the data were not normally distributed. Given that the study employed a within-subjects design, where each participant interacted with both interfaces (button panel and speech), related (paired) measurements were obtained for each usability and cognitive load metric. Consequently, the Wilcoxon signed-rank test was selected, as it is appropriate for analyzing paired or matched samples when the assumption of normality is violated. This approach allowed each participant to serve as their own control, enabling a more accurate comparison of score differences between the two interfaces. The resulting W-statistics and corresponding p-values are presented in Table 5.

Usability metrics: No statistically significant differences were observed in the usability metrics between the button panel and speech interfaces, as all comparisons yielded p-values greater than 0.05. The test revealed no statistically significant differences between the button panel and speech interfaces across most of the usability metrics, including *Ease of Control* ($W = 63.0$, $p = 0.788$), *Comfort* ($W = 43.5$, $p = 0.337$), *Accuracy of Commands* ($W = 30.0$, $p = 0.785$), and *Satisfaction with Response* ($W = 35.0$, $p = 0.439$). However, *Finding Controls* approached significance ($W = 20.0$, $p = 0.064$), suggesting a trend in which users might have found the control features easier to locate on one interface than on the other. The other metrics, including *Learn and Adapt* ($W = 20.0$, $p = 0.755$), *Recovery from Mistakes* ($W = 40.5$, $p = 0.712$), and *Natural and Intuitive Use* ($W = 68.0$, $p = 0.680$), also showed no significant differences, indicating comparable ease in learning and adapting, error recovery, and intuitiveness between the interfaces.

Cognitive load metrics: The *Performance* metric showed a statistically significant difference between the button panel and speech interfaces ($W = 9.0$, $p = 0.014$), suggesting a notable variation in the cognitive load associated with performance across the two interfaces. Additionally, *Physical Demand* and *Frustration* demonstrated marginally non-significant p-values ($p = 0.054$ for Physical Demand and $p = 0.084$ for Frustration), indicating a trend approaching a significance but not exactly statistically significant.

Table 5. Results of the Wilcoxon signed-rank test for usability and cognitive load metrics for button panel and speech interfaces.

Metric Category	Metric	W-Statistic	p-Value
Usability metrics	Ease of Control	63.0	0.7879
	Comfort	43.5	0.3368
	Accuracy of Commands	30.0	0.7850
	Satisfaction with Response	35.0	0.4389
	Finding Controls	*20.0*	*0.0638*
	Learn and Adapt	20.0	0.7551
	Recovery from Mistakes	40.5	0.7121
	Natural and Intuitive Use	68.0	0.6798
Cognitive load metrics	Mental Demand	29.5	0.2545
	Physical Demand	*11.5*	*0.0541*
	Temporal Demand	5.0	0.2356
	Performance	**9.0**	**0.0146**
	Effort	27.5	0.6094
	Frustration	6.0	0.0845

4.2. Qualitative Results

The qualitative analysis of the user feedback on the speech interface was conducted by codes and themes, where the code represented more granular detail while theme presented the overall category that the code belonged to. Table 6 (Qp) and Table 7 (Qn) show some distinct positive and negative remarks from the participants, respectively.

Many of the participants highlighted the speech interface's ease of use, specifically noting its capacity to simplify tasks and reduce physical demands compared to button-based interaction. Qp1, Qp2, Qp9 are some examples of those instances. Responsiveness was another aspect that was commented on in multiple instances (Qp1, Qp9). The participants also stated its ease of use through Qp4, Qp5, Qp6. Some also commented about the ease of finding controls and commands by using the speech interface, which sometimes proved difficult with the button panel interface (Qp8). Many also commented about workflow efficiency and the natural flow of things when using the speech interface (Qp10, Qp11).

Table 6. Positive remarks from participants, by codes and themes.

ID	Quotes	Codes	Themes
Qp1	"It is very responsive and its ability to understand instructions in many ways makes it handy and accessible".	responsive, flexible instructions	ease of use, reduced latency
Qp2	"The speech interface simplified tasks significantly since I didn't have to press buttons some of which were difficult to reach".	avoids button pressing, simplifies tasks	task simplification, less physical
Qp3	"Great, even though I have used VR before it was hard for me to use the panel but the speech felt more natural".	natural feeling, panel interaction difficulties	natural interaction, usability
Qp4	"At first it was challenging, but after a few tasks, I got a good grasp and it felt like a compact, effective tool".	initial difficulty, improved over time	learning curve, adaptability
Qp5	"Easy to learn and work. It can be really useful for professionals".	easy to learn, professional usefulness	learning curve, adaptability
Qp6	"Easier to learn and adapt to. Made the tasks easier. I didn't have to press the buttons and some of the buttons are not easy to press".	avoids button pressing, panel interaction difficulties	learning curve and adaptability
Qp7	"The VR speech assistant was able to understand my questions at least 90% so that's a plus".	accurate understanding of questions	speech recognition accuracy
Qp8	"Overall good, may improve speech recognition. I was able to visualize but still struggled to find handles, there were a lot of buttons in the panel displayed".	difficult to find buttons, visualization clarity	finding controls and commands
Qp9	"It's very responsive and its ability to understand the instruction in many ways. It's handy in many ways".	responsive, flexible instructions	reduced latency, ease of use
Qp10	"I could stay focused on the model without needing to stop and press buttons, which is crucial in a workflow setting".	workflow continuity	workflow efficiency
Qp11	"I think it is easier to stick with the flow using speech while working. I felt effortless and it might be very interesting for the dentists to play around with efficiency. I feel this speech interface might be an artificial assistant".	task flow continuity, artificial assistant potential	natural interaction, task Flow

There were negative aspects of the speech interface as well, mostly due to inaccuracy for speech recognition (Qn1, Qn2, Qn3) and overall accuracy (Qn4) leading to bad user experience. The primary reasons behind this were varying accents and pitch of the voice, the results being misinterpretation of full or half sentences, leading to wrong results. Many commented on the overall task limitation in the speech interface (Qn4, Qn6, Qn7), stating that a major aspect of the whole setup was manual work, like placing the dental implant. This was also in line with the expectation for completely hands-free interaction with speech, as in Qn6, where the expectation was that manual placement of the dental implant be done by speech, and as in Qn5, where there was a suggestion that no hand controller button be used for speech activation, thus automating the whole process.

Table 7. Negative remarks from participants, by codes and themes.

ID	Quotes	Codes	Themes
Qn1	"My accent was not understood clearly. Sometimes I had to speak slowly so that it understands entirely".	struggles with accent	speech recognition accuracy
Qn2	"It interpreted few words wrong maybe due to lower voice and accent".	misinterpreted commands	speech recognition accuracy
Qn3	"It was difficult for the system to take in long sentences".	struggles with long commands	speech recognition, accuracy
Qn4	"Some commands are not interpreted correctly, limitation in the amount of tasks. The system responded identically to 'Hide implants' and 'Show implants,' showing a need for command differentiation".	incorrect commands, limited tasks	overall accuracy, task limitation
Qn5	"Automating the speech activation, similar to Siri or Alexa, could reduce the need for physical input, making the interface more hands-free".	automation suggestion, reduced physical input	automation, hands-free design
Qn6	"The speech was good in total, but in some scenarios, manual intervention was required. So for completing the task by using speech interface alone was not achieved".	manual intervention needed	task limitation
Qn7	"Including more command variations and adding common questions would make it feel more intuitive".	suggests expanded commands, common questions	task limitation, command variation

5. Discussion

The described speech interface designed for changing tools and asking questions in medical VR applications demonstrated comparable performance to a traditional button panel interface, in regard to task completion metrics. The participants successfully completed an equal number of tasks with both interfaces, showing that a speech interface is a viable alternative to button-based methods. This aligns with prior research on single-word voice commands by [34], where medical experts rated speech modalities as satisfactory, useful, natural, and accurate. Comparable results were observed in metrics like Ease of Control, Comfort, and Natural and Intuitive Use, further validating the speech-based system's usability and its potential for broad application in dynamic medical environments. Systems that incorporate conversational fidelity and intent recognition have shown positive results in medical scenarios, with speech being an effective medium in VR for medical training and diagnosis [15,16,35]. This study proves that similar technology could be extended to being an assistant in surgical scenarios, providing realism through natural language, adaptivity, ease of control, comfort, finding control, and less physical effort, through objective and subjective results. The use of speech not only improves interaction in many key areas but also enables independent work. These results highlight the potential of speech-based systems to improve user interactions in medical VR, making them more effective and user-friendly in completing tasks. Another reason for the relatively good scores across various usability and cognitive load metrics was reduced latency, with the results being almost in real time, which was not observed in prior studies [39,40]. This system also enables users or developers to make easy configurations, where a number of commands, feedback, and answers can be personalized. Reduced physical load and effort were also reported by similar studies [17,35] that dealt with physical elements in VR, like the keyboard, manual data entry, etc. A novel finding in this research was the improved ease of finding controls with speech interfaces, supported by the "Finding Controls" metric and open-ended responses, addressing a gap not explored in previous studies. This functional parity suggests that speech interfaces could serve as viable alternatives to button panels in similar contexts, particularly in terms of usability.

The participants reported slightly higher cognitive load with the speech interface, though the difference was statistically insignificant. The qualitative results suggest that this increased cognitive load was primarily due to mental demand, which contrasts with past studies [15,17] where speech interfaces were associated with reduced cognitive effort. The higher mental demand in this study stemmed from small inaccuracies in command recognition, which was commented on in the qualitative results. These inaccuracies were universal across the participants and became more noticeable when the users spent extended periods inside the VR environment, even though error-handling mechanisms were available. For example, during a task requiring the user to issue 10 speech commands, the first 5 commands were typically recognized accurately, creating a smooth interaction flow. However, when a misrecognition occurred (e.g., at the sixth command, even if it was identical to a previously successful command), it disrupted the interaction flow. The users then needed to undo the incorrect action and reissue the correct command. It was observed that such disruptions, although infrequent, negatively impacted user experience by breaking immersion and increasing frustration. If multiple consecutive recognition errors occurred, frustration levels increased further, contributing to lower usability ratings and higher reported cognitive load.

An anticipated aspect of the speech interface was the ability to automate physical activities within VR (Qn6, Qn8, Qn9 in Table 7), such as picking up dental implants and placing them in jaws, as well as answering questions like "is the placement accurate or not" through a "Question Answering" feature. The participants also recommended adding conversational elements to understand, such as polite expressions ("Thank you", "Please"), to make interactions more natural and human-like, akin to modern voice assistants like Alexa or ChatGPT.

Future Work

In further studies, a broader emphasis could be given to training examples in order to utilize diverse accents and dialects. Even small inaccuracies may lead to lower user experience in real-time interaction, which was observed and noted in the experiments. Aspects such as accents, medical verbatim, and dialects supporting these could be taken into consideration while training the speech-to-text model.

It was seen that the participants were often confused about which controls were automated under the speech interface. With the emergence of and familiarity with LLM-based services, such as ChatGPT and home automation speech assistants, it was anticipated that the speech assistant in this study would support a diverse range of actions. However, even after its introduction and demonstration of the use case, the interface did not fully meet these expectations, highlighting areas for further development to align with advancements in similar technologies. This could be achieved by adding more training examples and leveraging LLMs to recognize the actions in a more intelligent manner. Lightweight LLMs could be applied with reduced latency. For answering in questions, advanced techniques like RAG (retrieval-augmented generation) could be used to generate more robust answers. There were also expectations, like picking of the dental implant, looking at implant positioning, and showing the accuracy of placement points. The anticipation from the speech interface was for it to function as a more intelligent system, not only controlling objects within medical VR scenarios but also extracting and providing relevant information from the interface to enhance decision making and interactivity. Furthermore, incorporating conversational elements such as greetings and medically appropriate phrases could enhance the naturalness and interactivity of the system.

A multimodal approach combining speech and button interfaces on screen could reduce cognitive load and improve performance by mitigating some negative aspects of speech, like task limitations. Although the present study focused on comparing the speech interface to a button-based interface, integrating the two could leverage the strengths of both, minimizing cognitive demand. A similar conclusion was drawn by [34], who suggested that combining modalities enhances usability and overall interaction efficiency. Other modalities, such as gaze tracking, would further enable reducing cognitive load, as the target of an action could be determined from the user's gaze direction. Gestures could also be an effective addition. For example, inaccuracies could be easily undone with a wave of the hand.

6. Conclusions

As immersive technologies are emerging in modern medicine, this study highlights the value of speech-based interaction in virtual reality (VR) environments. By integrating natural language processing into a surgical planning scenario, we have demonstrated that voice-driven commands can serve as a practical and intuitive alternative to traditional VR controls. The participants were able to complete tasks with similar effectiveness, while also experiencing benefits such as reduced physical effort, improved comfort, ease of finding controls, and more natural engagement with the system. At the same time, the study raised key challenges that must be addressed in speech recognition accuracy, accent variation, and user expectations shaped by mainstream voice assistants. These findings emphasize the need for speech systems that are not only technically sound but also sensitive to the nuances of real-world use, especially in high-stake medical contexts.

This research showed that speech interfaces are more than just convenience features. They can meaningfully support user cognition, streamline workflows, and lower the barriers to interacting with complex medical tools in VR. As language technologies evolve, there is strong potential to build voice-enabled virtual assistants that offer context-aware support, enhance training, and ultimately improve patient outcomes. As medical training and planning increasingly adopt immersive technologies, speech-based systems offer a scalable path toward more accessible, user-friendly, and intelligent VR experiences. This work lays the groundwork for the development of multimodal, AI-enhanced virtual assistants that can augment human capabilities in complex healthcare environments.

Author Contributions: Conceptualization, M.N., J.K. and R.R.; methodology, M.N.; software, M.N.; validation, M.N.; formal analysis, M.N.; investigation, M.N.; resources, R.R.; data curation, M.N.; writing—original draft preparation, M.N.; writing—review and editing, M.N., J.K. and R.R.; visualization, M.N.; supervision, J.K. and R.R.; project administration, R.R.; funding acquisition, R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research Council of Finland grant number 345448 and Business Finland grant number 80/31/2023.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki. Ethical review and approval were waived for this study, due to the nature of the experimental tasks and the participant population.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data collected during the study are available upon request.

Acknowledgments: The authors wish to thank members of TAUCHI Research Center for their support in the practical experiment arrangements. We would also like to thank Planmeca Oy for providing software and the 3D skull models.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AR	augmented reality
BERT	bidirectional encoder representations from transformers
DNN	deep neural networks
LLM	large language model
LUIS	Azure language understanding
NLP	natural language processing
NLU	natural language understanding
QA	quality assurance
STT	speech-to-text
TLX	task load index
TTS	text-to-speech
VR	virtual reality
VSP	virtual standardized patient
XR	extended reality

References

1. Al-Khalifah, A.; McCrindle, R.; Sharkey, P.; Alexandrov, V. Using virtual reality for medical diagnosis, training and education. *Int. J. Disabil. Hum. Dev.* **2006**, *5*, 187–194. [CrossRef]
2. Reitinger, B.; Bornik, A.; Beichel, R.; Schmalstieg, D. Liver surgery planning using virtual reality. *IEEE Comput. Graph. Appl.* **2006**, *26*, 36–47. [CrossRef]
3. Bhat, S.H.; Hareesh, K.; Kamath, A.T.; Kudva, A.; Vineetha, R.; Nair, A. A Framework to Enhance the Experience of CBCT Data in Real-Time Using Immersive Virtual Reality: Impacting Dental Pre-Surgical Planning. *IEEE Access* **2024**, *12*, 45442–45455. [CrossRef]
4. Kangas, J.; Järnstedt, J.; Ronkainen, K.; Mäkelä, J.; Mehtonen, H.; Huuskonen, P.; Raisamo, R. Towards the Emergence of the Medical Metaverse: A Pilot Study on Shared Virtual Reality for Orthognathic–Surgical Planning. *Appl. Sci.* **2024**, *14*, 1038. [CrossRef]
5. Chen, W.; Kuniewicz, M.; Aminu, A.J.; Karaesmen, I.; Duong, N.; Proniewska, K.; van Dam, P.; Iles, T.L.; Hołda, M.K.; Walocha, J.; et al. High-resolution 3D visualization of human hearts with emphases on the cardiac conduction system components—a new platform for medical education, mix/virtual reality, computational simulation. *Front. Med.* **2025**, *12*, 1507005. [CrossRef]
6. Kumar, S.; Fred, A.L.; Ajay Kumar, H.; Miriam, L.J.; Jane, I.C.; Padmanabhan, P.; Gulyás, B. Role of Augmented Reality and Virtual Reality in Medical Imaging. In *Introduction to Extended Reality (XR) Technologies*; Scrivener Publishing LLC.: Beverly, MA, USA, 2025; pp. 157–171.
7. Stucki, J.; Dastgir, R.; Baur, D.A.; Quereshy, F.A. The use of virtual reality and augmented reality in oral and maxillofacial surgery: A narrative review. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2024**, *137*, 12–18. [CrossRef]
8. Zajtchuk, R.; Satava, R.M. Medical applications of virtual reality. *Commun. ACM* **1997**, *40*, 63–64. [CrossRef]
9. Tene, T.; Vique López, D.F.; Valverde Aguirre, P.E.; Orna Puente, L.M.; Vacacela Gomez, C. Virtual reality and augmented reality in medical education: An umbrella review. *Front. Digit. Health* **2024**, *6*, 1365345. [CrossRef]
10. King, F.; Jayender, J.; Bhagavatula, S.K.; Shyn, P.B.; Pieper, S.; Kapur, T.; Lasso, A.; Fichtinger, G. An immersive virtual reality environment for diagnostic imaging. *J. Med. Robot. Res.* **2016**, *1*, 1640003. [CrossRef]
11. Bueckle, A.; Buehling, K.; Shih, P.C.; Börner, K. 3D virtual reality vs. 2D desktop registration user interface comparison. *PLoS ONE* **2021**, *16*, e0258103. [CrossRef]
12. Park, S.; Suh, G.; Kim, S.H.; Yang, H.J.; Lee, G.; Kim, S. Effect of Auto-Erased Sketch Cue in Multiuser Surgical Planning Virtual Reality Collaboration System. *IEEE Access* **2023**, *11*, 123565–123576. [CrossRef]
13. Cockburn, A.; McKenzie, B. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Minneapolis, MN, USA, 20–25 April 2002; pp. 203–210.

14. Fernandez, J.A.V.; Lee, J.J.; Vacca, S.A.S.; Magana, A.; Pesam, R.; Benes, B.; Popescu, V. Hands-Free VR. *arXiv* **2024**, arXiv:2402.15083. [CrossRef]
15. Ng, H.W.; Koh, A.; Foong, A.; Ong, J.; Tan, J.H.; Khoo, E.T.; Liu, G. Real-time spoken language understanding for orthopedic clinical training in virtual reality. In *Proceedings of the International Conference on Artificial Intelligence in Education*; Springer: Cham, Switzerland, 2022; pp. 640–646.
16. Maicher, K.; Stiff, A.; Scholl, M.; White, M.; Fosler-Lussier, E.; Schuler, W.; Serai, P.; Sunder, V.; Forrestal, H.; Mendella, L.; et al. Artificial intelligence in virtual standardized patients: Combining natural language understanding and rule based dialogue management to improve conversational fidelity. *Med. Teach.* **2022**, *45*, 279–285. [CrossRef]
17. Prange, A.; Barz, M.; Sonntag, D. Speech-based medical decision support in vr using a deep neural network. In Proceedings of the IJCAI, Melbourne, Australia 19–25 August 2017; pp. 5241–5242.
18. McGrath, J.L.; Taekman, J.M.; Dev, P.; Danforth, D.R.; Mohan, D.; Kman, N.; Crichlow, A.; Bond, W.F.; Riker, S.; Lemheney, A.; et al. Using virtual reality simulation environments to assess competence for emergency medicine learners. *Acad. Emerg. Med.* **2018**, *25*, 186–195. [CrossRef]
19. Dobbala, M.K.; Lingolu, M.S.S. Conversational AI and Chatbots: Enhancing User Experience on Websites. *Am. J. Comput. Sci. Technol.* **2024**, *11*, 62–70. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
21. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the naacL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Volume 1.
22. Trivedi, A.; Pant, N.; Shah, P.; Sonik, S.; Agrawal, S. Speech to text and text to speech recognition systems-Areview. *IOSR J. Comput. Eng.* **2018**, *20*, 36–43.
23. Abdul-Kader, S.A.; Woods, J. Survey on chatbot design techniques in speech conversation systems. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*. [CrossRef]
24. Kumari, S.; Naikwadi, Z.; Akole, A.; Darshankar, P. Enhancing college chat bot assistant with the help of richer human computer interaction and speech recognition. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020; pp. 427–433.
25. Inupakutika, D.; Nadim, M.; Gunnam, G.R.; Kaghyan, S.; Akopian, D.; Chalela, P.; Ramirez, A.G. Integration of NLP and speech-to-text applications with chatbots. *Electron. Imaging* **2021**, *33*, 1–6. [CrossRef]
26. Lopes, D.S.; Jorge, J.A. Extending medical interfaces towards virtual reality and augmented reality. *Ann. Med.* **2019**, *51*, 29. [CrossRef]
27. Yang, J.; Li, E.; Wu, L.; Liao, W. Application of VR and 3D printing in liver reconstruction. *Ann. Transl. Med.* **2022**, *10*, 915. [CrossRef]
28. Huettl, F.; Saalfeld, P.; Hansen, C.; Preim, B.; Poplawski, A.; Kneist, W.; Lang, H.; Huber, T. Virtual reality and 3D printing improve preoperative visualization of 3D liver reconstructions—Results from a preclinical comparison of presentation modalities and user's preference. *Ann. Transl. Med.* **2021**, *9*, 1074. [CrossRef] [PubMed]
29. Nunes, K.L.; Jegede, V.; Mann, D.S.; Llerena, P.; Wu, R.; Estephan, L.; Kumar, A.; Siddiqui, S.; Banoub, R.; Keith, S.W.; et al. A Randomized Pilot Trial of Virtual Reality Surgical Planning for Head and Neck Oncologic Resection. *Laryngoscope* **2024**, *135*, 1090–1097. [CrossRef] [PubMed]
30. Isikay, I.; Cekic, E.; Baylarov, B.; Tunc, O.; Hanalioglu, S. Narrative review of patient-specific 3D visualization and reality technologies in skull base neurosurgery: Enhancements in surgical training, planning, and navigation. *Front. Surg.* **2024**, *11*, 1427844. [CrossRef] [PubMed]
31. Li, Z.; Kiiveri, M.; Rantala, J.; Raisamo, R. Evaluation of haptic virtual reality user interfaces for medical marking on 3D models. *Int. J. -Hum.-Comput. Stud.* **2021**, *147*, 102561. [CrossRef]
32. Rantamaa, H.R.; Kangas, J.; Kumar, S.K.; Mehtonen, H.; Järnstedt, J.; Raisamo, R. Comparison of a vr stylus with a controller, hand tracking, and a mouse for object manipulation and medical marking tasks in virtual reality. *Appl. Sci.* **2023**, *13*, 2251. [CrossRef]
33. Alamilla, M.A.; Barnouin, C.; Moreau, R.; Zara, F.; Jaillet, F.; Redarce, H.T.; Coury, F. A Virtual Reality and haptic simulator for ultrasound-guided needle insertion. *IEEE Trans. Med. Robot. Bionics* **2022**, *4*, 634–645. [CrossRef]
34. Rantamaa, H.R.; Kangas, J.; Jordan, M.; Mehtonen, H.; Mäkelä, J.; Ronkainen, K.; Turunen, M.; Sundqvist, O.; Syrjä, I.; Järnstedt, J.; et al. Evaluation of voice commands for mode change in virtual reality implant planning procedure. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 1981–1989. [CrossRef]
35. Yang, J.; Chan, M.; Uribe-Quevedo, A.; Kapralos, B.; Jaimes, N.; Dubrowski, A. Prototyping virtual reality interactions in medical simulation employing speech recognition. In Proceedings of the 2020 22nd Symposium on Virtual and Augmented Reality (SVR), Porto de Galinhas, Brazil, 7–10 November 2020; pp. 351–355.

36. O'Hara, K.; Gonzalez, G.; Sellen, A.; Penney, G.; Varnavas, A.; Mentis, H.; Criminisi, A.; Corish, R.; Rouncefield, M.; Dastur, N.; et al. Touchless interaction in surgery. *Commun. ACM* **2014**, *57*, 70–77. [CrossRef]
37. Li, Z.; Akkil, D.; Raisamo, R. Gaze-based kinaesthetic interaction for virtual reality. *Interact. Comput.* **2020**, *32*, 17–32. [CrossRef]
38. Rakkolainen, I.; Farooq, A.; Kangas, J.; Hakulinen, J.; Rantala, J.; Turunen, M.; Raisamo, R. Technologies for multimodal interaction in extended reality—A scoping review. *Multimodal Technol. Interact.* **2021**, *5*, 81. [CrossRef]
39. Hombeck, J.; Voigt, H.; Lawonn, K. Voice user interfaces for effortless navigation in medical virtual reality environments. *Comput. Graph.* **2024**, *124*, 104069. [CrossRef]
40. Chen, L.; Cai, Y.; Wang, R.; Ding, S.; Tang, Y.; Hansen, P.; Sun, L. Supporting text entry in virtual reality with large language models. In Proceedings of the 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR), rlando, FL, USA, 16–21 March 2024; pp. 524–534.
41. Chen, J.; Liu, Z.; Huang, X.; Wu, C.; Liu, Q.; Jiang, G.; Pu, Y.; Lei, Y.; Chen, X.; Wang, X.; et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* **2024**, *27*, 42. [CrossRef]
42. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–38. [CrossRef]
43. Liu, Z.; Heer, J. The effects of interactive latency on exploratory visual analysis. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2122–2131. [CrossRef]
44. Gallent-Iglesias, D.; Serantes-Raposo, S.; Botana, I.L.R.; González-Vázquez, S.; Fernandez-Graña, P.M. IVAMED: Intelligent Virtual Assistant for Medical Diagnosis. In Proceedings of the SEPLN (Projects and Demonstrations), Jaén, Spain, 27–29 September 2023; pp. 87–92.
45. Trivedi, K.S. Fundamentals of Natural Language Processing. In *Microsoft Azure AI Fundamentals Certification Companion: Guide to Prepare for the AI-900 Exam*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 119–180.