A Voice-Controlled Dialogue System for NPC Interaction using Large Language Models

Milan Wevelsiep, Nicholas Thomas Walker, Nicolas Wagner and Stefan Ultes
Natural Language Generation and Dialogue Systems
Otto-Friedrich-University of Bamberg Bamberg, Germany
nicholas.walker@uni-bamberg.de, stefan.ultes@uni-bamberg.de

Abstract

This paper explores the integration of voice-controlled dialogue systems in narrative-driven video games, addressing the limitations of existing approaches. We propose a hybrid interface that allows players to freely paraphrase predefined dialogue options, combining player expressiveness with narrative cohesion. The prototype was developed in Unity, and a large language model was used to map the transcribed voice input to existing dialogue options. The approach was evaluated in a user study (n=14) that compared the hybrid interface to traditional point-and-click methods. Results indicate that the proposed interface enhances the player's degree of joy and perceived freedom while maintaining narrative consistency. The findings provide insights into the design of scalable and engaging voice-controlled systems for interactive storytelling. Future research should focus on reducing latency and refining language model accuracy to further improve user experience and immersion.

1 Introduction

Voice interaction in video games remains a niche yet promising feature, especially as advances in technology offer new possibilities for immersion and interaction of the player. Traditional approaches to voice-controlled dialogues with Non-Playable Characters (NPCs) in games generally fall into two categories: reading out pre-written dialogue lines or free speech input in AI-generated dialogues. The former often limits player expression, while the latter can lack narrative consistency and control. This paper aims to present a novel approach that serves as a middle ground between these two approaches, combining the flexibility of player input with structured narrative cohesion.

The goal of this paper is to explore the implementation of a voice-controlled interface (VCI) that allows players to freely phrase their responses while still choosing from pre-defined dialogue options. By evaluating this hybrid approach, we aim to determine its impact on the player experience, particularly in the context of narrative-driven games. Specifically, we address the following research questions:

1. How does the use of a voice-controlled interface impact the immersion and user experience in a game with a narrative focus?
2. Does the player using this VCI have a sense of freedom given a restricted set of predefined dialogue options?
3. To which degree of accuracy can the player's spoken responses be reliably mapped to a given set of dialogue options?

The key contribution of this work lies in an approach to enable spoken interaction in a narrative-driven game that balances player freedom with narrative consistency. We present findings that highlight the potential of this approach in enhancing immersion and user satisfaction while maintaining cohesive storytelling.

The remainder of the paper is structured as follows: Section 2 presents and discusses other approaches that include voice control into games and discusses how our approach differs. Section 3 contains the core concept of the voice-controlled dialogue system with a description of its realization in Section 4. Sections 5, 6, and 7 present the user study design, the results and their discussion.

2 Related Work

Voice control as a narrative device in video games has gained significant attention for its potential to enhance player immersion (Allison et al., 2020). Natural voice interactions are generally well-received, as they enhance player flow and reduce identity dissonance (Carter et al., 2015). Players often mimic character voices (Allison et al.,

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 29–38
May 27–30, 2025. ©2025 Association for Computational Linguistics

2019; Osking and Doucette, 2019), deepening immersion, though this can be challenging when there are differences in player and character attributes such as gender (Carter et al., 2015). Persistent issues with voice interfaces include unnatural interactions, difficulty recalling commands, slower response times compared to button inputs (Allison et al., 2019), and recognition failures (Zargham et al., 2022). This section reviews notable approaches specifically for voice-controlled dialogues with non-playable characters (NPCs).

One established approach is the use of read-out-loud interfaces, where players speak predefined dialogue lines to interact with NPCs (Osking and Doucette, 2019)(Cuebit, 2018). Here, players cannot freely phrase their voice input but are restricted to the phrasing of the dialogue option they are choosing. This method is reliable and can enhance immersion by encouraging players to embody their characters. For instance, *Flowers for Dan dan* (Osking and Doucette, 2019) used a read-out-loud interface where players verbally selected dialogue options by reading the text of the dialogue option, resulting in higher emotional engagement compared to traditional point-and-click controls. Similarly, the *Dragonborn Speaks Naturally* modification for *Skyrim* (Cuebit, 2018) adopted this approach to create more immersive player-NPC interactions without the need for complex AI systems. The main advantage of read-out-loud interfaces is their practical integration into existing games, as they rely on predefined dialogue options and require minimal changes to the game's dialogue system. However, the restrictive nature of reading out predefined dialogue lines may limit the player's sense of agency, reducing immersion over extended play sessions.

Dynamic dialogue generation represents another approach, where NPC responses are generated in real-time using AI techniques such as natural language processing (NLP) or large language models (LLMs). This approach provides players with greater freedom and more natural interactions by allowing them to speak freely rather than selecting from predefined options. For example, the game Façade (Mateas and Stern, 2003), later modified by Dow et al. (Dow et al., 2007), employed a "Wizard of Oz" technique to simulate natural speech input. Building on top of its underlying AI systems—natural language processing, autonomous character behaviour, and a drama manager—this approach fostered dynamic and immersive conversations. Fraser et al. (Fraser et al., 2018) extended this concept by incorporating sentiment analysis to adapt NPC responses based on player emotions, thereby enhancing engagement. Similarly, *Bot Colony* (Joseph, 2019) and *Vaudeville* (Bumblebee-Studios, 2023) utilized AI-driven dialogue systems to generate NPC responses. While the use of LLMs in dialogues with NPCs, such as those in *Vaudeville*, can create human-like dialogue that enhances player engagement, they also present challenges including hallucinations, inconsistencies, and difficulty maintaining narrative coherence. Fraser et al.'s (Fraser et al., 2018) sentiment-driven approach demonstrated improvements in emotional immersion; however, concerns regarding scalability in larger game environments and negative player reactions to AI-generated dialogue remain (Cox and Ooi, 2024; Akoury et al., 2023).

Building on these existing methods, this paper proposes a middle-ground solution that integrates the strengths of both approaches. By allowing players to use free-form speech while mapping their input to predefined dialogue options, our method seeks to maintain immersion and deliver a natural interactive experience without compromising narrative control. This hybrid approach offers a more scalable and robust solution for voice-controlled dialogues in narrative-driven games by addressing the challenges identified in earlier research.

3 Concept

The core of the proposed approach lies in a middle-ground solution for integrating voice interaction in narrative-focused video games. It combines predefined dialogue options with the player's ability to paraphrase freely. To achieve this, players are given predefined dialogue choices that contain very concisely worded versions of the core messages. However, instead of asking the players to read them out loud, they are encouraged to paraphrase these options in their own words. For example, an option like "Ask for more information" can be expressed as "Could you give me more details?". While in some cases the participants nonetheless opted to read the text as given or nearly so, others were more creative in their formulations. Hence, a pre-defined dialogue flow controls the overall dialogue while users can speak freely and naturally.

The overall architecture of modelling the dialogue and processing new user input is shown in Figure 1. New user input is first processed in the *Understanding* component that utilizes a large lan-

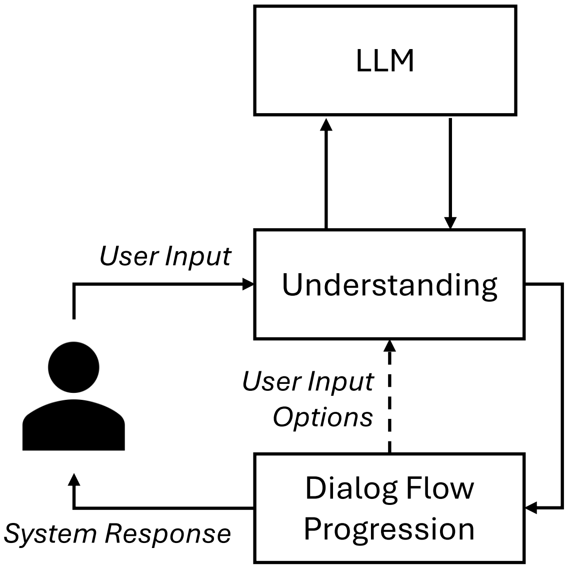


Figure 1: Overall dialogue architecture: an LLM is used to map new user input to one of the user input options defined by the dialogue flow.

guage model to map the user input to one of the possible dialogue options. These options are part of the pre-defined dialogue flow and represented in textual form. The large language model is then prompted to either map the user input to one of the dialogue options given the previous system response, or to map it to *misunderstood*.

Once the user input is mapped to one of the predefined dialogue options, the dialogue progresses to the next node of the pre-defined dialogue flow which defines the system output along with a new set of dialogue options as possible user inputs. The new set of dialogue options is subsequently used together with the following user input in the *Understanding* component.

Thus, this concept draws from read-out-loud interfaces (Osking and Doucette, 2019) and dynamic input methods (Fraser et al., 2018; Bumblebee-Studios, 2023) alike. It allows player freedom and the capability to maintain narrative control. Unlike fully generative NPC responses, which often lack coherence, this approach relies on a structured dialogue graph to ensure consistency while enabling natural voice interaction. Allowing players to phrase their responses freely is expected to enhance immersion and engagement compared to restrictive read-aloud interfaces.

4 Prototype Development

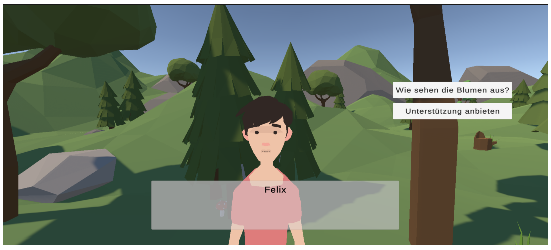


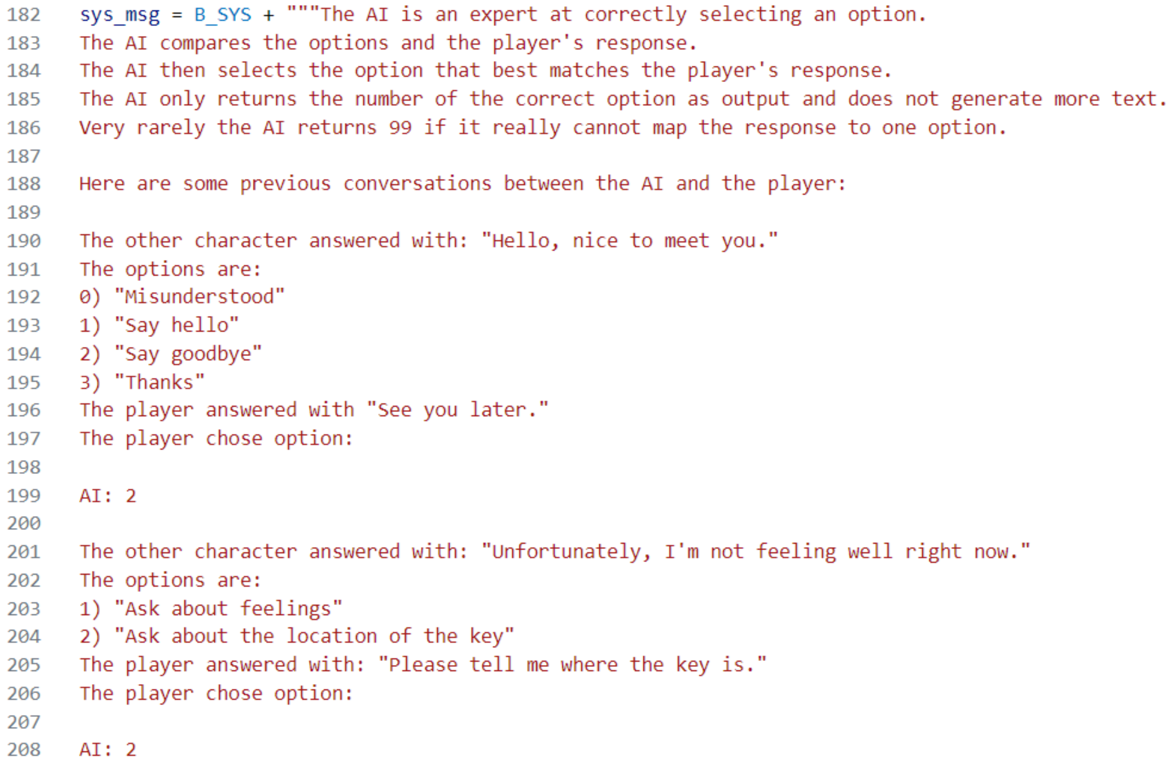
Figure 2: A screenshot of the tutorial-section of the game. The two dialogue options, displayed in German language, translate to "How do the flowers look like?" and "Offer support".

The proposed concept is realized in a prototype implementation of a narrative-driven game. The prototype was built using Unity, chosen for its flexibility and extensive library of assets. Unity handled all game mechanics, visual elements, character interactions, and user interface components. Custom C# scripts managed core game interactions, such as dialogue flow, NPC responses, and player controls. The game environment and characters were created using free Unity Asset Store resources, providing a functional game world for voice interaction testing. A screenshot of the game is shown in Figure 2. In the tutorial shown in the figure, the player is instructed by an NPC to help search for flowers in the forest by selecting one of two options of how to respond.

Voice Interaction: Player speech input was captured in Unity and processed through the Whisper AI service for transcription. Given that the language model performed better with English input, the transcribed German text was translated into English via the Google Cloud Translation API before further processing.

Dialogue Management: A structured dialogue graph, implemented with Unity's internal tools and custom C# scripts, served as the backbone for dialogue flow. Each node in this graph represented a specific narrative point linked to predefined player options. Player input was mapped to these options using the Llama-2 13B language model, hosted on an Nvidia A100 GPU. The model received a prompt that included the transcribed and translated response, the current NPC dialogue, and available dialogue options. The model then returned the option number that best matched the player's intent. Prompt engineering was applied by using *langchain* to improve mapping accuracy and reduce latency. An excerpt of the system message is shown in Figure 3.

User Interface: Developed within Unity, the user interface displayed available dialogue options and



```
sys_msg = B_SYS + """The AI is an expert at correctly selecting an option.
The AI compares the options and the player's response.
The AI then selects the option that best matches the player's response.
The AI only returns the number of the correct option as output and does not generate more text.
Very rarely the AI returns 99 if it really cannot map the response to one option.

Here are some previous conversations between the AI and the player:

The other character answered with: "Hello, nice to meet you."
The options are:
0) "Misunderstood"
1) "Say hello"
2) "Say goodbye"
3) "Thanks"
The player answered with "See you later."
The player chose option:

AI: 2

The other character answered with: "Unfortunately, I'm not feeling well right now."
The options are:
1) "Ask about feelings"
2) "Ask about the location of the key"
The player answered with: "Please tell me where the key is."
The player chose option:

AI: 2
```

Figure 3: An excerpt of the system message used in the few-shot prompting. The system message includes an explanation of the task and examples on how to map player's responses to options.

provided immediate visual feedback. When players used voice input, the UI indicated ongoing processing and highlighted the chosen option after recognition, helping players understand the system's response to their spoken input.

Key Features: The prototype includes several key design elements to enhance player experience and immersion. The game adopts a first-person perspective, allowing players to interact directly with the environment and NPCs to create a more engaging experience. While in a dialogue, the player can choose the dialogue options hands-free, i.e., the player does not need to press a button to start or stop the voice input. After a brief period of silence detection, Unity processes the player's speech and matches it with the available dialogue options. The VCI also provides a mechanism for revising dialogue options. When the system misinterprets a player's input, phrases such as “I didn't mean that” trigger a "Misunderstood" option. In order to avoid mental overload by visualizing this additional option, the revise-option only becomes visible upon selection. The participants in the user study have been made aware of that option in an initial introduction. The prototype also includes a tutorial designed to help players familiarize themselves with the basic controls and mechanics.

Limitations: The game's world and characters are constructed from different resources from the Unity Asset Store. Therefore, the game environment appears visually inconsistent and the NPC's facial expressions and animations are limited. Both aspects lead to a presumably less immersive and believable experience. Response latency in the voice interface presents another issue, with delays sometimes interrupting the natural flow of conversation. Finally, while the automatic German-to-English translation system generally performs well, it occasionally misinterprets nuanced phrases, which can result in faulty mapping of player's speech input to the dialogue options.

5 User Study

The user study aimed to evaluate the proposed voice-controlled interface (VCI) by comparing it with a traditional point-and-click interface (PCI). The study combined usability testing, A/B testing, and surveys to assess the system's impact on user experience, perceived freedom, and system accuracy.

The study involved 14 participants, consisting of an equal number of male and female individuals,

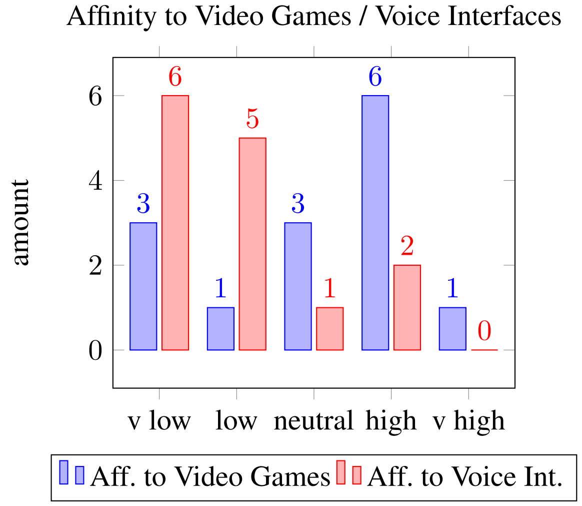


Figure 4: Self-reported affinity towards video games and voice interfaces by the participants.

aged between 23 and 36 years (mean age: 29, standard deviation: 4). Most participants (11) held university degrees, and all were native German speakers. Participants were personally recruited and included a mix of friends, acquaintances, and individuals with no close connection to the researchers. This group represented varied levels of familiarity with gaming and voice interfaces. While participants had moderately high experience with video games, their exposure to voice interfaces was comparatively limited (see Fig. 4). Participants alternated between the two interfaces to counterbalance order effects, with one group using the VCI first and the other starting with the PCI. The procedure included the following phases:

1. **Introduction and Orientation:** Participants were briefed on the study, signed consent forms, and received instructions on gameplay mechanics. A presentation highlighted the use of voice input, including the correction feature for misunderstood inputs.

2. **Tutorial Level:** Participants completed a short tutorial using the VCI to familiarize themselves with the system. Assistance was provided during this phase as needed.

3. **Main Game Playthrough:** Participants played the main game with one interface while the researcher minimized observer effects. The task of the game is to help the NPC Felix to find a missing key. During this task, the dialogue hints that Felix is bothered by something else, and the player has the option to inquire further about this issue or ignore it and focus on finding the key. Each session ended upon reaching one of the game's three possible outcomes. The three endings correspond to low, medium and high levels of empathy as determined by the level of empathy shown to Felix in the player's responses over the course of the dialogue.

4. **Post-Play Questionnaire:** Participants completed a questionnaire assessing the interface they had just used.

5. **Second Playthrough:** Participants replayed the main game with the alternative interface, followed by the same questionnaire.

6. **Final Questionnaire:** A comprehensive questionnaire captured additional metrics like accuracy, enjoyment, and overall preference.

The study was conducted on a laptop equipped with the Unity-based prototype. Voice input was captured using a Logitech webcam microphone, chosen for its accuracy over the laptop's built-in microphone. Participants completed questionnaires on the same laptop. Audio recordings documented verbal interactions, while logs captured system responses, dialogue choices, and observational notes.

Two primary data sources, questionnaires and play-through documentation, informed the study's findings.

Questionnaires: Participants responded to a series of structured questions using seven-point Likert scales. The questionnaires were adapted from existing instruments, namely the SASSI (Hone and Graham, 2000) for assessing the speech interface with regard to the usability aspects, and the GUESS (Vieira et al., 2019) for measuring video game satisfaction and user experience. The adapted questionnaire covered five scales:

- **System Response Accuracy:** Assessed how reliably player inputs were mapped to predefined options (Items I1–I2).

- **Likeability:** Measured user enjoyment and perceived freedom (Items I3–I4).

- **Cognitive Demand and Habitability:** Evaluated ease of use and confidence in issuing voice commands (Items I5–I6).

- **Annoyance and Speed:** Captured frustration and delays during gameplay (Item I7).

- **Immersion:** Examined how natural and engaging the interactions felt (Item I8).
- **Preference and Overall Assessment:** Assessed which interface players preferred (Item I9).

Playthrough Documentation: Logs recorded dialogue choices, LLM prompt-responses, and voice input accuracy. Audio recordings and observational notes provided qualitative insights into user behavior, naturalness of interactions, and system responsiveness.

The study faced several limitations that must be acknowledged. The small and relatively homogeneous sample, consisting of younger participants with higher education, is not representative of the broader gaming population. The limited duration of the study restricted participants' familiarity with the interfaces, potentially limiting the learning curve and long-term usability assessment. Some prototype limitations, such as latency, translation inaccuracies, and limited NPC animations, likely influenced user perceptions of the system. Factors like mood, time of day, and external distractions could also have impacted participant performance and feedback.

6 Results

The results of the user study are presented in this section, focusing on the impact of the voice-controlled interface (VCI) on immersion, user experience, perceived freedom, and accuracy of spoken inputs. A total of 14 participants completed the study, which involved gameplay with both the VCI and a traditional point-and-click interface (PCI), followed by corresponding questionnaires. An overview of the results for the VCI is shown in table 1.

The questionnaire responses were collected on a 7-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (7). For positively phrased statements, higher values indicate a more favorable response, while for negatively phrased statements, the scale was reversed to ensure consistency in interpretation, where higher values always reflect a positive attitude towards the VCI.

Impact on Immersion and User Experience (R1)

Research Question R1 evaluated the overall impact of the VCI on immersion and other aspects of user experience. Items I3, I5, I6, I7, and I8 were analysed:

Table 1: Summary of the Results for the Voice-Controlled Interface

Item	**Median**	**Mean**
I1: Accuracy of Mapping	5.0	4.89
I2: Correction of Misunderstood Input	–	–
I3: Degree of Joy	5.0	4.93
I4: Expressing Freedom	4.0	3.93
I5: Ease of Use	4.0	4.07
I6: Confidence in Using the VCI	3.0	3.86
I7: Annoyance	4.5	4.38
I8: Immersion	3.0	3.68
Overall Assessment and Preference:		
Use in real games	5.5	5.43
Preference if improved	6.0	5.79

Degree of Joy (Item I3): Participants rated enjoyment of the VCI with a median of 5.0 and a mean of 4.93, suggesting a moderately positive experience. When compared directly with the PCI, the VCI scored higher (median 5.5, mean 5.43), indicating enhanced enjoyment through voice interaction.

Ease of Use (Item I5): Ease of use received mixed ratings, with a median of 4.0 and a mean of 4.07. Participants noted higher cognitive demand for the VCI due to the need for paraphrasing. In comparing both interfaces directly with each other, participants reported the VCI as more demanding (median 2.5, mean 3.07). In part, this can be due to higher familiarity with a traditional interface. However, the mental load for putting a paraphrased dialogue option into one's own words most likely further contributed to this.

Confidence in Using the VCI (Item I6): Confidence levels varied, with a median of 3.0 and a mean of 3.86. Participants expressed moderate confidence but reported uncertainty regarding whether their phrasing would be correctly recognized, suggesting a need for improvement.

Annoyance (Item I7): General annoyance was low (median 5.0, mean 5.21), but participants gave a more neutral rating of their attitude towards the VCI response time (median 3.5, mean 3.93). Reducing latency could significantly improve the overall experience.

Immersion (Item I8): The VCI provided slightly better immersion compared to the PCI (median 3.0, mean 3.57), but neither fully replicated natural dialogue. Improvements in natural language processing are needed to enhance immersion further.

Sense of Freedom (R2)

Research Question R2 examined participants' perceived freedom while using the VCI:

Expressing Freedom (Item I4): Participants felt moderately free to express themselves (median 5.0, mean 4.5). While compared to the PCI, the VCI allowed more authentic expression (median 5.0, mean 5.07), the limitations of predefined options occasionally hindered free expression (median 3.0, mean 3.36).

Degree of Accuracy (R3)

Research Question R3 focused on the accuracy of mapping spoken responses to dialogue options:

Accuracy of Mapping (Item I1): Mapping accuracy was rated positively (median 5.0, mean 4.89), with a system accuracy of approximately 90%. Participants often adhered closely to predefined phrasing, positively influencing accuracy.

Correction of Misunderstood Input (Item I2): The correction feature was rarely used due to infrequent mapping errors. However, its hidden nature led to participants often overlooking this functionality, suggesting a need for better visibility and usability.

Overall Assessment and Preference

Participants rated the VCI positively for potential use in real games (median 5.5, mean 5.43). While direct preferences between the VCI and PCI were mixed (median 4.0, mean 4.36), most participants indicated they would use the VCI if its accuracy and speed were improved (median 6.0, mean 5.86).

Additionally, no significant correlation was found between participants' familiarity with video games or voice interfaces and their perception of the VCI. This suggests that the VCI is accessible and engaging for a broad audience, regardless of prior experience, supporting its potential appeal in diverse gaming contexts.

7 Discussion

This section offers a comprehensive discussion of the user study results and final reflections on the voice-controlled interface (VCI) prototype, synthesizing the findings, implications, limitations, and directions for future research.

Table 2: Summary of participant responses to direct comparison questions between the voice-controlled interface (VCI) and the point-and-click interface (PCI). Higher values indicate a greater preference for the VCI.

Item	Median	Mean
I3: Joy (VCI vs. PCI)	5.5	5.43
I4: Expressing Freedom (VCI vs. PCI)	5.0	5.07
I5: Ease of Use (VCI vs. PCI)	2.5	3.07
I3: Boredom (VCI vs. PCI)	6.0	5.50
I8: Immersion (VCI vs. PCI)	4.0	4.36

Interpretation of Results

The user study findings show that the VCI prototype was generally well-received by participants, offering notable advantages in engagement and user experience compared to the conventional point-and-click interface (PCI). Participants expressed a preference for the VCI, indicating its potential to enhance player involvement and enjoyment, despite the presence of technical issues like response latency and speech recognition challenges.

Impact on Immersion and User Experience (R1): Participants found the VCI enjoyable, though delays in processing voice input caused frustration and moderate annoyance. Confidence in using the system was mixed, likely due to the unfamiliarity of combining predefined options with the freedom to paraphrase responses. Improvements in response time and system reliability are essential to enhance immersion and user comfort. Despite these flaws, the VCI had a slight advantage over the PCI in terms of immersion, highlighting its potential for narrative-driven games.

Sense of Freedom (R2): Participants appreciated the ability to paraphrase predefined options, which contributed to a sense of authenticity and self-expression. However, the restricted nature of predefined choices occasionally limited participants' sense of freedom. Future iterations of the VCI could improve flexibility, reducing perceived constraints and enhancing player empowerment.

Accuracy of Mapping (R3): Participants generally found the VCI predictable, though inconsistencies in speech recognition affected how reliably spoken input was mapped to dialogue options. The correction feature for misunderstood inputs was underutilized due to its hidden presentation. Despite these issues, the technical approach—using a language model (LLM) for mapping—shows promise, particularly with improved speech recognition and responsiveness.

Practical and Theoretical Implications

The positive reception of the VCI suggests that voice interaction, particularly in narrative contexts, is an engaging feature for video games. The hybrid approach of combining predefined dialogue options with paraphrasing offers a scalable solution for integrating voice control into games without compromising narrative coherence. Allowing players to "play as themselves" enhances player embodiment, especially in games where player agency is a core feature, such as role-playing games (RPGs).

Addressing technical limitations such as response time and speech recognition accuracy is essential for the commercial adoption of the VCI. Improvements in these areas would significantly enhance player experience, making the interface more reliable and enjoyable. Adding the flexibility to toggle the VCI on and off would give players greater control, catering to diverse preferences.

The study also contributes to understanding how voice interaction can be effectively integrated into video games. Unlike traditional top-down communication, the VCI allows for more natural interactions with non-playable characters (NPCs), fostering immersion by enabling players to project their identity onto the character. To achieve deeper immersion, improvements in system speed, accuracy, and NPC responsiveness are still required.

Limitations

The study faced several limitations that affect the generalizability of the findings. Methodologically, the short duration of the study restricted participants' ability to become familiar with the VCI, limiting insights into long-term usability. The controlled environment may not fully replicate real-world gaming conditions, influencing interactions and feedback. Additionally, the small sample size and participant homogeneity limit the applicability of the findings to a broader gaming audience.

An additional limitation of the study is the influence of the presented options on the players' thinking. The specific wording of the options may influence the way in which participants phrase their statements in the dialogue. Further work which analyzes differences in user input dependent upon how options are presented or if options are displayed at all would likely yield additional insights.

Similarly, specifics within the dialogue options may also be interpreted in specific or more general ways by participants. For instance, an "Ask for more information" dialogue option may be interpreted as pertaining to specific or general information. The specificity or generality of dialogue options may thus constitute an additional factor for participant experiences that would be of interest to subsequent research.

Technical limitations also played a significant role in shaping user experience. Latency issues and inconsistencies in speech recognition disrupted conversation flow and reduced immersion. The fixed time required for voice recognition, combined with delays in transcription and language model processing, significantly affected user satisfaction. Additionally, the lack of expressive character animations and authentic voice output further hindered immersion and the believability of NPC interactions.

A final limitation worth mentioning regarding immersion is that various aspects of language such as sarcasm, irony, or other nuanced aspects of how humans naturally communicate were out of scope for this study. The relative advantages of a voice-controlled interface over a point-and-click interface will likely be most strongly observable in a system that incorporates further subtleties of human expression.

Future Research

Future research should prioritize addressing the technical and methodological limitations identified in this study. Enhancing the speed and accuracy of voice recognition through real-time transcription and more advanced language models could significantly improve the VCI's performance. Incorporating dynamic dialogue generation could also provide more flexible and adaptive player-NPC interactions, addressing the constraints of predefined dialogue options.

Long-term studies are needed to understand the sustained effects of voice interaction on player engagement and immersion. Integrating the VCI into commercial games for extended periods would provide valuable insights into how players adapt to and perceive the system. Additionally, future research should explore the role of voice interaction in fostering emotional connections between players and NPCs, particularly through improved NPC animations, responsive dialogue, and enhanced player agency.

Dynamic Dialogue Generation within Dialogue Graphs: Dynamic dialogue generation is a promising direction for enhancing flexibility in voice interactions. The current prototype relies on predefined dialogue options, limiting adaptability. By integrating dynamic dialogue generation into dialogue graphs, NPC responses can be generated based on the player's phrasing and narrative context, improving natural interaction flow.

This hybrid approach, using language models to generate context-aware responses while maintaining the structure provided by dialogue graphs, could offer a more personalized experience. NPC responses could vary based on player phrasing, past interactions, and storyline context, making conversations more engaging and lifelike. However, challenges such as maintaining emotional authenticity, ensuring lip synchronization, and minimizing latency need to be addressed. Future research should explore lightweight language models capable of efficient operation within game environments.

8 Conclusion

This study contributes to the growing field of voice interaction in video games, particularly NPC interactions. The hybrid VCI approach—combining predefined dialogue options with the ability to paraphrase—has proven to be an engaging feature that enhances a player's sense of freedom and overall user experience. Allowing players to interact naturally, in their own words, creates a more personalized experience that can align well with narrative-driven games.

While the study's findings show the potential of voice-controlled interfaces, the technical and methodological challenges identified must be addressed for long-term success. Improvements in system speed, reliability and more advanced character design are critical for impacting the sense of immersion to a greater extent. With these enhancements, voice interaction could become an integral part of video game dialogue systems, providing a richer and more immersive player experience.

References

Nader Akoury, Qian Yang, and Mohit Iyyer. 2023. A Framework for Exploring Player Perceptions of LLM-Generated Dialogue in Commercial Video Games. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2295–2311, Singapore. Association for Computational Linguistics.

Fraser Allison, Marcus Carter, and Martin Gibbs. 2020. Word Play: A History of Voice Interaction in Digital Games. *Games and Culture*, 15(2):91–113.

Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Glasgow Scotland Uk. ACM.

Bumblebee-Studios. 2023. Vaudeville. `https://bumblebeestudios.itch.io/vaudeville`.

Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 265–269, London United Kingdom. ACM.

Samuel Rhys Cox and Wei Tsang Ooi. 2024. Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie L.-C. Law, Ewa Luger, Morten Goodwin, Sebastian Hobert, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, volume 14524, pages 167–184. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Cuebit. 2018. Dragonborn Speaks Naturally. `https://www.nexusmods.com/skyrimspecialedition/mods/16514?tab=description`.

Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1475–1484, San Jose California USA. ACM.

Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken Conversational AI in Video Games: Emotional Dialogue Management Increases User Engagement. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 179–184, Sydney NSW Australia. ACM.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering*, 6(3&4):287–303.

Eugene Joseph. 2019. From Virtual to Real: A Framework for Verbal Interaction with Robots. In *Proceedings of the Combined Workshop on Spatial Language Understanding*, pages 18–28, Minneapolis, Minnesota. Association for Computational Linguistics.

M. Mateas and A. Stern. 2003. Façade: An Experiment in Building a Fully-Realized Interactive Drama. In *Game Developer's Conference: Game Design Track*.

Hunter Osking and John A. Doucette. 2019. Enhancing Emotional Effectiveness of Virtual-Reality Experiences with Voice Control Interfaces. In Dennis Beck, Anasol Peña-Rios, Todd Ogle, Daphne Economou, Markos Mentzelopoulos, Leonel Morgado, Christian Eckhardt, Johanna Pirker, Roxane Koitz-Hristov, Jonathon Richter, Christian Gütl, and Michael Gardner, editors, *Immersive Learning Research Network*, volume 1044, pages 199–209. Springer International Publishing, Cham. Series Title: Communications in Computer and Information Science.

Estela Aparecida Oliveira Vieira, Aleph Campos Da Silveira, and Ronei Ximenes Martins. 2019. Heuristic Evaluation on Usability of Educational Games: A Systematic Review. *Informatics in Education*, 18(2):427–442.

Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. 2022. "I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, New Orleans LA USA. ACM.