Speech Recognition Game Interface to Increase Intimacy with Characters

Saki Anzai, Tokio Ogawa, Junichi Hoshino

HAL Id: hal-04144413
https://inria.hal.science/hal-04144413

Submitted on 28 Jun 2023

Speech Recognition Game Interface to Increase Intimacy with Characters

Saki Anzai[1], Tokio Ogawa[1], and Junichi Hoshino[1]

[1] University of Tsukuba, Ibaraki, Japan
tayloref3127@gmail.com, ogawa.tokio@esys.tsukuba.ac.jp,
jhoshino@esys.tsukuba.ac.jp

Abstract. Although the number of games using speech recognition has increased in recent years, psychological effects such as intimacy with screen characters when using voice dialogue have not been clarified. We particularly focused on games that allow users to talk with characters relatively freely, and thought that the intimacy with the interactive partner is related to the enjoyment of the conversation and the game as a whole. In this paper, we selected a game interface that employs voice dialogue in combination with keyboard operation used in conventional games and compared the differences in the overall enjoyment and immersion in the game due to the differences in intimacy with the characters. As a result, we were able to show that the intimacy with the characters and the enjoyment and immersion of the game increased. Finally, a discussion on the type of interface that actually influenced the degree of intimacy and enjoyment while playing the game is given.

Keywords: Game, Speech Recognition, Dialogue, Intimacy.

1   Introduction

In recent years, the number of games using speech recognition has increased. There are various forms of gaming, such as a strategy game that uses voice to give instructions to screen characters to fight or an adventure game that captures the characters within the confines of a dungeon while chanting spells. Many related studies have been conducted, and the contents are various, such as the influence of the introduction of speech recognition on playing games and the analysis of the form of speech input in games. However, psychological effects such as intimacy with the screen characters when using voice dialogue have not been clarified. Especially in games where dialogue with the character is an important factor, we suppose that the intimacy with the character of the dialogue partner is closely related to the fun and immersive feeling of the game itself. If the player has a sense of intimacy with the screen character like a friend, the conversation will be more enjoyable and the overall enjoyment of the game will be enhanced. As a result, we focused on speech recognition and intimacy with the screen characters. The term "intimacy" described here refers to that which is gained in the process of interacting with a partner in a mutual relationship. It is different from "familiarity," which is derived from first impressions of another person or similarity to oneself, or "attachment," which is felt toward a pet. In a previous study, Ikeda stat-

ed that the more intimate the relationship, the more we establish a sense of unity through an emotionally mutual relationship with another person [5]. Emotionally mutual relationships include "shared volition," in which we work together toward the same goal, and "shared emotions," in which we share our feelings with each other. Finally, we arrive at the "mutual relationship" of being together with an irreplaceable friend. In this paper, our primary focus is the development of intimacy in mutual relationships between human identification with animated screen character interaction within the context of the gaming experience.

Bickmore (2005) created a conversational agent with memory retention and gesture functions and tried to operate it as an exercise promotion assistant [9]. As a result, users developed a favorable impression of the agent and increased their desire to continue using the agent. The results also show that a favorable relationship with the interaction partner has a significant impact on the outcomes of various tasks. This shows that a favorable relationship with the interaction partner has a positive impact on engagement, but it does not address fun and immersion. In addition, it is not clear what kind of detailed conversations are useful in games or how to increase the level of intimacy.

To compare the difference in input method of dialogue in the game, we realized a multi-modal game interface that combines voice dialogue with the operation by the keyboard used in the conventional game and compared the influence on the intimacy with the character of the game with the conventional game.

2 Related Works

2.1 Voice Dialogue Interaction

Allison (2018) analyzes how voice input is used in games and summarized the results [1]. He investigated the inputs that affect the state of the game, and analyzed them in six major categories: "story composition," "dialogue structure," "selection," "navigation," "control," and "performance." He subdivided them further and defined 25 design patterns. Carter (2015) studied the relationship between the concept of player-avatar identity and game success in the introduction of voice interaction [2]. By effectively using voice, it is possible to think of one's own character and oneself in the game in an overlapping manner, and it is considered that the immersive experience in the game will be improved. Allison (2019) analyzed the impact of speech recognition on playing games in terms of social experience [3]. Emotions such as embarrassment and incongruity often occur in gamelan by voice operation, but the player's anxiety can be alleviated by clearly communicating the purpose of the voice command. It is also stated that by using voice, people can feel a sense of unity with their own character in the game.

2.2 A Subsection Sample

Intimacy is being studied in the field of social psychology. Nishida (1992) shows that it is easy to have a favorable impression such as friendliness to an interlocutor who seems to be aware of a specific conversation strategy [4]. Conversation strategies

include "acceptance strategies" and "interactive dialogue strategies." For example, it includes situations such as actively listening to the other party's remarks and feeling that they are considering their own utterances or trying to make the other party understand their own remarks.

According to Ikeda (2013), there is a strong relationship between sharing style and intimacy between friends [5]. It can be classified into six sharing styles, namely, "sharing relationships," "sharing places," "sharing feelings," "sharing intentions," "sharing goods," and "sharing sensibilities." For example, "sharing feelings" includes items such as "sharing each other's pain" and "encouraging each other." It is said that these shared styles will become more pronounced among friends as intimacy increases. There are many studies that can be used as a reference for conversations with screen characters, but it does not touch on the content of the utterances on the player side, and it is not clear whether these rules work within the context of voice dialogue in the game. In addition, no research has been conducted on what kind of conversation specifically affects the increase in intimacy.

3   System Overview

We designed game interface used to compare with conventional games, to have smooth voice dialogue with characters, and to measure the increase in intimacy experienced during the gaming event. The following four system requirements are listed:

1) The system must have an in-game environment similar to that of a conventional game.

2) The system can accurately recognize the player's voice.

3) The system can interact with the characters in the game.

4) The system can have conversations that encourage increased intimacy.
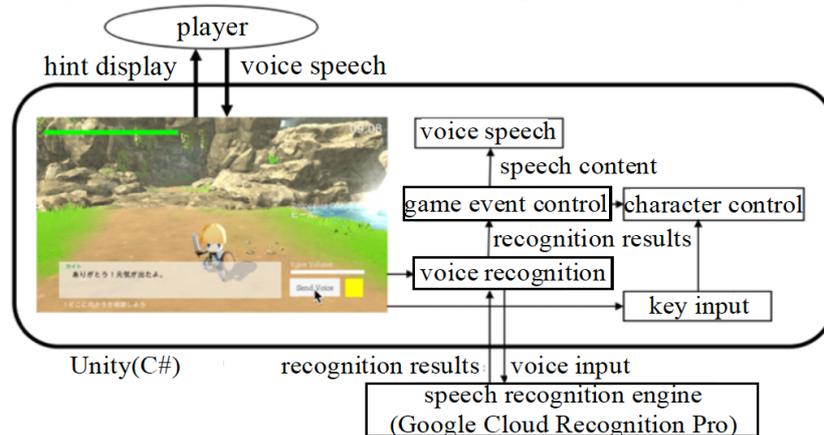
Fig. 1. System configuration diagram

In this research, in order to verify the change in intimacy between the player and the character by voice dialogue, we created an adventure game that can be played in about 10 to 15 min, which makes it easy to give the player a purpose (Fig. 1).

We used Unity to create the game and Google Cloud Speech Recognition for speech recognition. In conventional works that use speech recognition, there are many formats in which instructions are sent to the character on the screen from a bird's-eye view, but in this system, in order to compose a conversation that enhances intimacy, we placed the player and the screen character on equal footing in the same field. This not only allows the player to unilaterally help the character but also enables two-way communication between the player and the character. It is expected that intimacy can be increased by aiming for the goal while creating a form of mutual help in which the character responds when consulted by the player, assisting the character by turning around and answering when calling by name, etc. Also, in this game, not only are conversations related to the progress of the game, but also simple chats and conversations such as "Thank you" or "Hold on" are possible (Table 1). A guide is displayed so that the player does not get lost in the game's progress, but basically the player follows along and converses in the player's own words. Chit-chat is not in the guide, so it is something that the player must think about and say on their own. The character picks up keywords from the player's speech and responds. By following the guide, you can avoid most situations where the character is unable to respond.

Table 1. Example of speech conversations

	Player's Question	Character's response
Discussing our next move	What do I have to do?	Let's try to kill one of the enemies in front of us! Let's go for it!
	Where are we going?	I don't know. Let's go through the tunnel first.
	Where's the ocean?	It's on your right. Let's go there!
Chatting and encouraging each other	Thank you.	You're welcome!
	Let's go.	Let's go!
	Awesome.	Thank you!
	What's your favorite drink?	Milk! I'm gonna grow taller!

4 Composition of Game Content

4.1 Scenario

In this adventure game, the player collaborates with the screen characters to solve problems and aim for the goal. At that time, in addition to operations such as move-

Fig. 2. Evaluation game map and event layout

ment and attack by keyboard, which is used in conventional games, dialogue by speech recognition is used. The game is cleared by the player talking to the screen character by voice, assisting by recovery magic and sometimes solving the problem while consulting about the next action (Fig. 2).

4.2   Characters

The main character (dialogue partner) is a boy named Kite (Fig. 3) The player is a girl Win who travels with Kite (Fig. 4). The game screen is Win's field of view, and she interacts with Kite and assists with his name call and recovery (Fig. 5).

Fig. 3. A boy named Kite. The player can have voice conversations with Kite.

Fig. 4. The player is a girl named Win. The relationship between Win and Kite is that of a companion on a journey together.

Fig. 5. The player is in a position to chase after Kite. Since the game screen is Win's view, only Kite appears on the screen.

4.3 Game Screen Configuration

The game screen consists of the response of the main character, Kite, the display of speech recognition result, the physical strength gauge and the time limit (Fig. 6). Since it is difficult to detect the end of a player's utterance with this system, we decided to have the player press the Send Voice button when the utterance ends. The game is over when Kite's physical strength gauge on the upper left of the screen reaches 0, but this time it is set so that the physical strength does not decrease beyond a certain level due to the experiment. Additionally, Win does not appear in the screen because the game screen is from Win's perspective.
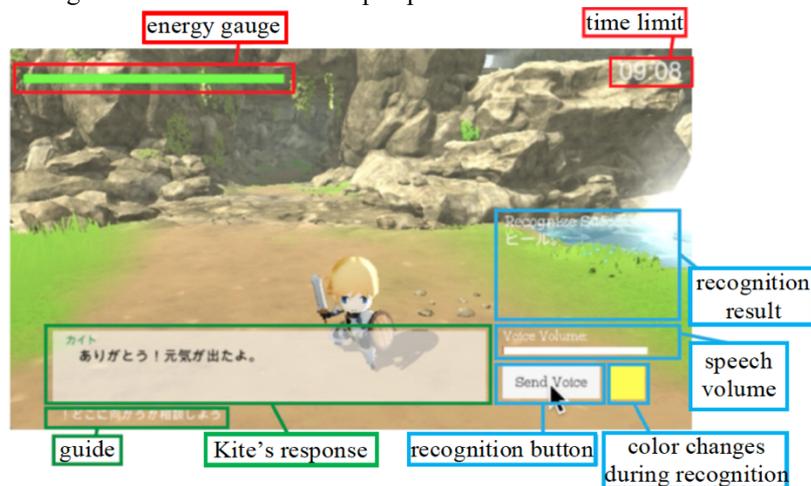

Fig. 6. Game screen configuration

4.4 Event

The first straight road after starting the game is a tutorial. It is possible to learn how to attack and rotate the line of sight and to learn how to cast recovery magic by talking to Kite. When you defeat the two enemies along the way, a woman wearing a brown hood appears and is presented with the task, "I want you to defeat the monsters that gather at the end of the road." If you defeat five enemies and talk to them again, you will be presented with the second task, "I want you to find my sister." If you find the sister wearing a red hood and talk to her, all the tasks will be completed, and the game will end.

4.5 Improvement of the interface through preliminary experiments

We conducted a preliminary experiment with the help of two university students. In the pre-experiment, we prepared a room for the experiment and recorded them playing the game. After playing the game, the students were asked to write freely about their impressions of the conversation with Kite, the game operation, and other comments. As a result, the following comments were raised: "Sometimes I did not know what to say," "I did not know when to say something," "The lag until the voice is

recognized may make the user feel uneasy," and "I feel a little embarrassed to talk when I am being watched." Based on these opinions, we added guides and tutorials to make it easier for users to know what to say, improved the UI so that users can check whether speech recognition is in progress, and exported the game so that users can download it at home to prevent distractions from the surrounding environment interfering with playing the game.

Additionally, we improved the game so that users can look around because they could not rotate their eyes at the time of the preliminary experiment. We also made other adjustments, including fixing minor bugs.

5   User Study

5.1   Experiment 1

An experiment was conducted to verify whether there is a change in the degree of intimacy with a character in a game when speech recognition is used.

Experimental Method

We had a total of 14 participants, 10 male university students and 4 female university students.

This experiment will be referred to as Experiment 1 below. We have prepared two types of games, a text version (hereinafter text version, Fig. 7) in which conversations assuming a conventional game are performed in a choice manner, and a game using voice (hereinafter voice version) produced this time, and each person downloads and plays it. The reason why we did not use a game with free-typing text as a comparison target is that many conventional conversation games use choices for conversation, and we wanted to compare our results with those.

Fig. 7. Game screen of the text version

Half of them, that is, seven people, experienced it in the order of text version → audio version, and the other half experienced it in the reverse order. After each experience, they answered the questionnaire.

The questionnaire adopted a 7-level Likert scale. The smaller number is "not applicable," and the answer is "applicable" when approaching 7. It consists of 10 items about intimacy and 8 items about the game system (Table 2). For question numbers 1-4 to 1-10, refer to previous research on intimacy and sharing styles, and ask questions about sharing of places, sharing of intentions, sharing of feelings, and sharing of relationships, excluding sharing of characteristics (Ikeda, 2013) [5]. Regarding the expression of the question, we referred to seven items about empathy of Kort (2007) [6]. For each of the eight items regarding the gaming system, the reasons for evaluation were voluntarily solicited. Therefore, we solicited comprehensive comments on intimacy and game system in a free-form format. We asked them to answer the first impression of Kite, the impression after playing, and the conversation that left an impression.

Table 2. Questionnaire result table

	question	number of subjects who rated each higher than the other	
		voice	text
1-1	I felt the friendliness with Kite	12	1
1-2	I felt closer than when I started	10	2
1-3	It was fun to have a dialogue	12	1
1-4	I felt that I was working together for the same activity	7	1
1-5	I felt like I was heading towards the same goal	5	1
1-6	I felt Kite and I were encouraging each other	10	2
1-7	When Kite was pleased, I was pleased	8	1
1-8	I played while thinking about the feelings of Kite	9	0
1-9	I shared emotions with Kite	11	1
1-10	I trusted Kite	7	2
2-1	Fun	7	2
2-2	I got hooked on the game	7	2
2-3	I forgot the time while playing	7	3
2-4	It was rewarding	5	2
2-5	There was a sense of accomplishment	9	0
2-6	Ease of operation	1	5
2-7	Easy to understand	1	7
2-8	Novelty and progressiveness of the game	12	0
2-9	Speech recognition accuracy (voice version only)		

Results and Discussion

This is the result of the questionnaire conducted in this experiment (Table 2). In the audio version and the text version, one score was given to each question with the higher score given by each individual, and the total score was shown. In general, the voice version scored higher, while question numbers 2-7 and 2-8 were higher in the text version.

The following figure is a box-and-whisker diagram of the results of Experiment 1 (Fig. 8). We can visually see that the scores of the audio version are higher than those of the text version for most of the questions.
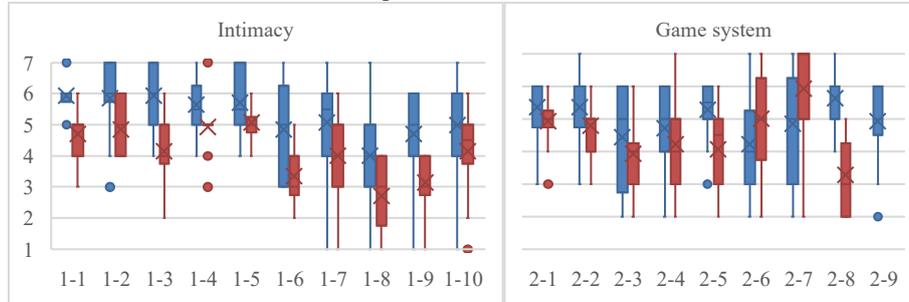
Fig. 8. Box-and-whisker diagram showing the results of Experiment 1

We have considered the results in comparison with the Wilcoxon Signed-Rank test (Table 3) . Questions that showed a significant difference depending on the p value are shown in light gray. As an exception, for question 1-2, whose result depends on whether or not the game was cleared, there is a difference in superiority when the test was performed by excluding the data of one subject who could not clear only the audio version. The voice version has a higher score (shown at the bottom of Table 3).

Table 3. Result of t-test analysis(p value)

question	average value		P value	question	average value		P value
	Voice	Text			Voice	Text	
1-1	5.929	4.714	0.003	2-1	5.357	4.929	0.199
1-2	5.857	4.857	※0.086	2-2	5.357	4.786	0.07
1-3	5.929	4.143	0.006	2-3	4.429	3.929	0.214
1-4	5.643	4.929	0.026	2-4	4.714	4.214	0.121
1-5	5.714	5.071	0.071	2-5	5.286	4.071	0.007
1-6	4.857	3.357	0.011	2-6	4.214	5	0.031
1-7	5.071	4	0.016	2-7	4.857	5.929	0.028
1-8	4	2.714	0.006	2-8	5.643	3.286	0.002
1-9	4.714	3.143	0.002	2-9	4.929		
1-10	5	4.143	0.046				

※0.016

As a result, the voice-based input method made the screen characters more familiar than did the text-based method, and the intimacy increased throughout the gameplay. Also, the dialogue has improved the fun during the gaming experience. In case of the text version, there are negative opinions such as "I felt that the degree of familiarity with Kite and the degree of sharing of emotions decreased because I could talk smoothly and without thinking because I could talk with just a button."

Hence, there was a tendency for the feeling of cooperation and encouraging each other to be strengthened to complete the same task and thinking more about Kite's feelings and share feelings such as fun. We can see that the process of cooperation in the game generated a "mutual relationship" as friends. However, there was no significant difference on the question: "I felt I was moving towards the same goal." Since the final goal is not clearly shown and one of the causes is that it was not incorporated into the conversation, it is possible to devise a scenario with conversational content as an improvement plan. Mentioned in the paragraph below, the introduction

of voice recognition may have affected this result by making the operation more difficult or confusing.

Regarding the gaming system, there was no significant difference between the voice version and the text version in questions 2-1 to 2-4. According to the free-text comments, there was an opinion that the voice version seemed to be really talking to the character and was more interesting than the text version, which had many opinions that the tempo was good, but it was monotonous and felt like work. On the other hand, there was an opinion that the time lag in speech recognition hindered smooth conversation. Regarding speech recognition, the delay may increase depending on the network environment, and it was determined that improvement is necessary. There is a suggestion for improvement: either have the game played in a place with as good of a network environment as possible or switch to a voice recognition system that does not require a network environment. In addition, it is difficult to understand the important parts of the operation just by reading the instruction manual before playing, and to improve the method is to devise the UI and operation explanation. Specifically, for example, it is difficult to know when to speak because they do not know if they are in the middle of speech recognition or not, and sometimes, it is hard to come up with something to say. Based on the results of preliminary experiments, we implemented a user interface where the color of the displayed figure changes during speech recognition, but that was not enough. Since some users did not know what the figure represented, it was necessary to clearly state in text that recognition was in progress.

There are individual differences in the level of understanding of the operations described above, time lag and accuracy of voice recognition, so there are people who give a score as low as 1, as shown in the box-beard diagram.

Regarding the sense of accomplishment and the novelty of the game, the voice version scored higher, showing a difference in superiority. There were opinions such as "I felt new because there are not many games that use voice recognition to progress the story," "I felt like I was in the world with Kite," and "I enjoyed chanting magic with my voice."

5.2   Experiment 2

We improved the game and conducted additional experiments.

Changes in enjoyment through UI improvements

To confirm that improving the UI improves the enjoyment and immersion of the game itself, we made the following improvements to the audio version of the game based on the free-text comments in the questionnaire from Experiment 1 and conducted additional experiments.

- ・Produced a video of operation explanation.
- ・The operation method was always displayed on the game screen during play.
- ・Displayed the text "Recognizing" during speech recognition.
- ・Partially added a guide.

The first two improvements will prevent the game from starting without the user understanding how to operate it. It also makes it easier to know when to speak by displaying the text during recognition. The addition of a guide solves the problem of not knowing what to say.

In addition, the guide that we added as an improvement after the preliminary experiment turned out to be insufficient. In response to the comment that participants were confused about how to talk to the female avatars along the way, we thought of some additions to the guide. For example, we added "Use any words such as 'Hello'" to the phrase "Talk to the woman." Also, when talking to Kite in the scene where he is searching for the female sister, I added a guide that says, "Ask her in detail what she heard about her sister's 'characteristics' and 'where she went'."

We had a total of eight participants: seven male university students and one female university student. Some of them cooperated with us in Experiment 1. As in Experiment 1, we compared the audio version with the text version. In the text version, we also implemented a video to explain the operation and a constant display of the operation method on the screen to match the conditions. As for the time lag of speech recognition mentioned in Section 5.3, we asked the participants to experience the system in a good network environment as much as possible.

After experiencing the game, the participants answered the same questionnaire as in Experiment 1. The following figure shows the box-and-whisker diagram of Experiment 2 (Fig. 9). As in Experiment 1, the score of the audio version is higher than that of the text version for most of the questions. Since both Experiments 1 and 2 evaluated the audio and text versions relative to each other, it is impossible to make a direct comparison between Fig. 8 and Fig. 9.
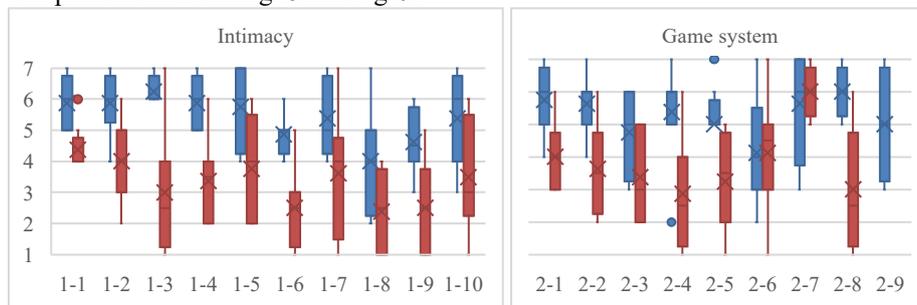
Fig. 9. Box-and-whisker diagram showing the results of Experiment 2

The following table shows the results of the t-test (Table 4). As in the previous experiment, questions that showed significant differences in p-values are shown in light gray, and the higher mean values for the audio and text versions are underlined. The questions with higher scores in the audio version remained the same, while questions 1-5 and 2-1 to 2-4, which were not significantly different in the previous experiment, were significantly different in the current experiment. Questions 2-6 and 2-7, which were significantly higher in the text version in the previous experiment, were not significantly different in this experiment. The results show that the improvement of the UI reduced the complexity of the game operations. The overall enjoyment and immersion of the game has been inherently improved by no longer being distracted by the complexity of the controls. It can be inferred that the subjects were able to con-

centrate more on the conversation with the characters than in Experiment 1. In the free-text comments, the subjects mentioned that they felt the improvement was effective: "the text of the guide was polite" and "the difficulty level was appropriate."

Table 4. Additional experiment questionnaire result table

ques-tion	average value		P value	ques-tion	average value		P value
	Voice	Text			Voice	Text	
1-1	5.875	4.375	0.026	2-1	5.75	4	0.01
1-2	5.875	4	0.017	2-2	5.625	3.625	0.017
1-3	6.25	3	0.018	2-3	4.75	3.375	0.026
1-4	5.875	3.375	0.011	2-4	5.375	2.875	0.017
1-5	5.75	3.75	0.042	2-5	5	3.25	0.026
1-6	4.875	2.5	0.017	2-6	4.125	4.125	0.914
1-7	5.375	3.625	0.016	2-7	5.625	6	0.496
1-8	4	2.375	0.027	2-8	6	3	0.011
1-9	4.625	2.5	0.027	2-9	5		
1-10	5.375	3.5	0.043				

Discussion

In the free-text comments made during the additional experiments, many of them referred to "being able to talk freely," such as "it was interesting to be able to speak freely and have Kite respond differently" and "it was interesting to think about what to say and try to see what kind of response would come." In contrast, in the case of the text version, there was a negative evaluation that "the attachment to Kite was diminished due to the inability to have conversations other than the choices," indicating that free speech is effective in increasing intimacy.

Additionally, while most of the conventional games using speech recognition have a system in which you give instructions to the characters on the screen from a bird's-eye view as a third party, this game allows you to talk with the characters on an equal footing as a character who exists in the scene. These social relationship between the player and the character to increase the intimacy with the screen character allowed the player to become more emotionally involved with the character and to feel the shared fun and adventure. A related comment was "The relationship between the player and the character as partners in the game world is interesting. I like the fact that there is always an incentive to continue the conversation because the game world always brings up issues that need to be solved." Of the four items listed as dialogue elements, the item "discussing the next action" seems to have had a particularly positive impact on the user experience. We found that the position of the user's avatar and the setting of the relationship between the avatar and the character are also deeply related to the degree of intimacy and the enjoyment of the game.

6 Conclusion

In this paper, we compared the effect of the game interface with voice dialogue in developing intimacy with the screen character in relation to the conventional game, and it was shown that we could enjoy the dialogue in the game more and intimacy

also increased by introducing voice input. The enjoyment and immersion of the entire gaming experience could also be improved by removing the difficulty and incomprehensiveness of operation due to the introduction of speech recognition. In the future, we will examine whether the results of this study can be applied when the content of the conversation and the story of the game are enriched and lengthened. In addition to the content of the conversations, we will also be able to recognize the intonation of the player's voice, give the character emotions and memories, and generate corresponding conversations.

References

1. Allison, F., Carter, F., Gibbs, M., Smith, W.: Design Patterns for Voice Interaction in Games. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play(CHI PLAY'18), Melbourne, VIC, Australia, pp. 5–17 (2018).
2. Carter, M., Allison, F., Down, J., Gibbs, M.: Player Identity Dissonance and Voice Interaction in Games. In Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, the 2015 Annual Symposium, pp. 265–269 (2015).
3. Allison, F., Newn, J., Swith, W., Carter, M., Gibbs, M.: Frame Analysis of Voice Interaction Gameplay. Proceedings of the 2019 CHI conference on Human Factors in Computing Systems(CHI '19), Glasgow, Scotland, UK, no. 393, pp. 1–14 (2019).
4. Nishida, K.: Effects of conversation act on conversation strategy and interpersonal cognition. The Japanese Journal of Psychology, vol. 63, no. 5, pp. 319–325 (1992).
5. Ikeda Y., Hayama, D., Kosaka, Y., Satoh, Y.: Relationship between Intimacy and Styles of Sharing Among University Friends. Japan Society of Youth and Adolescent Psychology, vol. 24, no. 2, pp. 111–124 (2013).
6. Kort, D. Y., IJsselsteijn, A. W., Poels, K.: Digital Games as Social Presence Technology: Development of the Social Presence in Gaming Questionnaire (SPGQ). Proceedings of the 10th Annual International Workshop on Presence, pp. 195–203 (2007).
7. Lim, S. J., Baba, A.: The development of measurement of empathy in digital games. Digital Games Research Association JAPAN Proceedings of 2016 Summer Conference (2016).
8. Brockmyer, H. J., Fox, M. C., Curtiss, A. K., McBroom, E., Burkhart, M. K., Pidruzny, N. J.: The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. Journal of Experimental Social Psychology, vol. 45, no. 4, pp. 624–634 (2009).
9. Bickmore, W. T., Picard, W. R.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer-Human Interaction, vol. 12, no. 2, pp.293-327 (2005).