Gestures or Speech? Comparing Modality Selection for different Interaction Tasks in a Virtual Environment

Kathrin Janowski (kathrin.manuela.janowski@student.uni-augsburg.de)
Felix Kistler (kistler@informatik.uni-augsburg.de)
Elisabeth André (andre@informatik.uni-augsburg.de)
Human Centered Multimedia, Augsburg University, Universitätsstr. 6a
D-86159 Augsburg, Germany

Abstract

In this paper, we investigate whether users prefer speech or gesture input for four distinct interaction tasks commonly found in virtual environments: *navigation*, *selection*, *dialogue*, and *object manipulation*. For this purpose, we implemented an interactive storytelling scenario in which the users could always choose between gesture and speech commands for each interaction. Both input modalities were processed in real-time using a low-cost depth sensor and microphone. We conducted a study in order to identify the modality preferences for each task. We got strong results for the navigational task, for which gestural interaction seemed to be more suitable, and for the dialogue task which was in favour of speech. For the object manipulation and selection tasks we did not observe a clear preference for one of the modalities, but we found indications for why some participants chose speech and others preferred gestures by analysing the participants' ratings of their experience with the interaction.

Keywords: gestures; speech; modality selection; full body interaction; recognition; virtual environment; navigation; selection; dialogue; manipulation

Introduction

In recent years, novel input technologies which offer to make human-computer interaction closer to human-human interaction have become available to the average user. For example, the three major video game consoles all provide motion recognition while speech recognition is offered on smart phones and personal computers. As speech and body motions serve as the main interaction modalities in the real world, it seems quite logical to use them for immersive interaction in virtual worlds as well. However, they need to be harmonically integrated with the virtual setting and intuitive for the user. In consequence, most current consumer products only use speech or gesture functionality to enhance a specific type of interaction, whereas they still rely on traditional input devices for other types or automate parts of the interaction, e.g. in the racing game "Kinect Joy Ride"[1], the player uses hand motions for steering to the left and right, but the car accelerates automatically, and in the role-playing game "Mass Effect 3"[2], the Kinect microphone is used for speech commands, but the rest of the input happens with a game pad instead of using gestural interaction provided by the Kinect depth sensor.

In this paper, we describe a system in which we solely use gesture and speech interaction. In the corresponding study, the user can always choose which of these two modalities to use for the currently available interactions.

[1] http://xbox.com/kinectjoyride
[2] http://masseffect.bioware.com

Figure 1: User interacting within our study setup.

Related Work

Gesture and speech interaction in virtual environments has already been extensively investigated in the research community for many years, but most research tends to focus on a specific type of interaction. Interaction types for arbitrary virtual environments include *navigation* that serves to reach interaction possibilities, often followed by *selection* that determines the currently relevant entities before *dialogue* or *manipulation* actions are used to change the world state. For navigational tasks, several applications use walking gestures, such as the VisTA-walk system by Kadobayashi, Nishimoto, and Mase (1998) or the system described in (LaViola, Feliz, Keefe, & Zeleznik, 2001). The control schemes considered for VisTA-walk include a joystick-like mapping which lets the user indicate their desired movement direction by physically stepping away from a neutral position. LaViola et al. (2001) use leaning gestures for indicating the direction when travelling short or medium distances, but let the user choose their target directly by walking on a map projected onto the floor for longer distances. The use of speech for navigation is much rarer in literature, but those who apply it also tend to specify the target itself (e.g. "go to location xy") as in (Cohen et al., 1999). Corradini and Cohen (2002) investigate speech and gesture inputs of users during the "Myst III - Exile" game with a Wizard-of-Oz setup. They discover that users tend to combine both modalities and that gestures for manipulating objects mostly follow the objects' affordances. The latter is also confirmed by Kistler, Sollfrank, Bee, and André (2011), who observe that users prefer full body gestures matching the narrative action to pointing at randomly positioned buttons with the action displayed on them as text. For selecting ob-

jects, van der Sluis and Krahmer (2007) examine how pointing gestures and verbal descriptions are combined to single out a particular option. Depending on various difficulty factors, participants focus on one channel while the less suitable one contributes a more general, imprecise expression. Cavazza et al. (2004) focus on speech input for dialogues using multi-keyword spotting to allow for flexible and natural phrasing. They add conversational gestures to reduce ambiguity, but consider speech "the only practical mode of communication" since spoken words are crucial to the narrative and gestures themselves are ambiguous without that context.

In contrast to most of the related work, we present an application that includes all four mentioned interaction tasks. In addition, our system actually applies real-time recognition of inputs as opposed to a Wizard-of-Oz setup. Instead of investigating different implementations of one modality or examining multimodal usage, our goal is to determine the primary interaction modality for each of those tasks. For this purpose, we conducted a study as described in the next section.

User Study

System and Setup

Our system displays a virtual world in a first person perspective on a 50 inch screen using the Horde3D GameEngine[3] as depicted in Figure 1. Each action in our system is linked to both a gesture and a speech command which can be used interchangeably. Gesture recognition is implemented using the "Full Body Interaction Framework" (FUBI)[4] described in (Kistler, Endrass, Damian, Dang, & André, 2012) in combination with a Microsoft Kinect depth sensor[5] placed centred below the screen, and using the OpenNI framework and NiTE middleware[6] for user tracking. Speech is processed with the Microsoft Speech Platform [7] for multi-keyword spotting on the audio stream of a wireless headset's microphone.

Our scenario consists of actions belonging to the four different tasks *navigation*, *selection*, *dialogue*, and *manipulation*. Users need to navigate to various selectable entities, and then perform dialogue and manipulation actions on them before moving on. In total, one has to perform about 17 actions per task to complete the scenario. For performing an action, the users can always choose between speech or gesture input and our primary hypothesis is that the two modalities are not equally suitable for every task. The implemented inputs are explained in the following.

Our application uses a navigation vocabulary for basic movements (i.e. move left/right/forward/backward) and rotations (i.e. turn left/right/up/down) which is considered closer to reality and more flexible than indicating a target directly, which would also overlap with the selection task. Gesture input for movements is based on a walking metaphor similar

[3] http://hcm-lab.de/projects/GameEngine
[4] http://hcm-lab.de/fubi.html
[5] http://kinectforwindows.org
[6] http://openni.org
[7] http://msdn.microsoft.com/library/hh361572.aspx

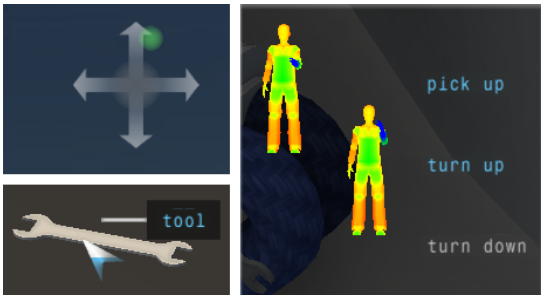

Figure 2: *Upper left:* Movement icon. *Lower left:* Object selection. *Right:* Available actions for the object on the left.

to the joystick control scheme in (Kadobayashi et al., 1998), e.g. the user has to step forward for starting a movement to the front. Similarly, the rotations directly use the torso orientation, e.g. the users actually have to turn left for starting a rotation to the left and they have to lean backwards for tilting their viewing angle upwards, which also resembles the rotation commands described in (LaViola et al., 2001). Feedback for the movement is provided by an icon (see Figure 2 on the upper left) that shows the user's physical position relative to a neutral zone defined as a 40cm x 40cm square about two meters in front of the screen. As for speech, navigation commands consist of naming their type and direction, e.g. saying "turn left" for turning left, or "forward" for moving forward. A label below the movement icon displays the recognized navigation command for feedback. For this task, our hypothesis is that gestures would be preferred to spoken commands as they are closer to natural navigation.

Interactive objects and characters in our scenario are marked with labels which are coloured blue instead of white when they are reachable. Pointing gestures are used to move a cursor across the screen and the user has to hold it above an entity for 0.5 seconds for selection (dwell-based selection), during which the cursor fills up with colour as shown in Figure 2 on the lower left. This is similar to the "button mode" described in (Kistler et al., 2011). The same selection is performed by speaking the entity's name as shown on its label, which is kept unambiguous in our scenario. Either command results in the display of available interactions (manipulation or dialogue) for this entity, presented in the style of a context menu. As both modalities seem equally natural for selection, we do not have a clear hypothesis for this task.

For virtual characters, the context menu displays the sentences which can currently be spoken to them. 15 unique phrases are available throughout the scenario, each of which contains one or several semantically important keywords (coloured in blue) which need to be said in the given order for speech input. The remaining words (coloured in white) are optional and can be changed or omitted by the user. This approach resembles the one described by Cavazza et al. (2004). For applying gestural interaction to the dialogue task, we are again using the pointing gestures as in the selection task. Therefore, the desired sentence is chosen by moving the cur-

sor to a button-like target next to it. The first reason for this decision is that conversational gestures are often ambiguous if used without accompanying speech as stated by Cavazza et al. (2004). Furthermore, not every topic has a straightforward gesture representation - for example, the scenario's very first question of "Where am I?" would be hard to express with a single gesture. As speech seems to be a very obvious choice for dialogue, we hypothesize it to be preferred for this task.

Interactive objects can be manipulated by gestures which resemble real-world actions as suggested by Corradini and Cohen (2002), e.g. raising the knees is used to step onto a bed, and moving the hand like pulling a lever is used for actually doing this. Animated human figures display the motions that are expected from the user as depicted in Figure 2 on the right-hand side. These animations are automatically generated from the same XML gesture definitions used by the FUBI framework for gesture recognition. Based on the given speed limits, state durations and transition times, movement paths for the joints of a virtual character are defined and later applied using inverse kinematics. The speech alternative mainly consist of the action's verb, but occasionally, a second parameter such as a tool or direction is added for clarification, e.g. "turn up" is used for turning a spanner upwards. All currently available speech commands are listed in blue next to the animated figures for the corresponding gestures, whereas actions which may become available later are greyed out. Overall, 14 different keywords and 18 different gestures were included for the manipulation task. The hypothesis for this task is that gestures would be preferred, as they are closer to object manipulation in real life.

Participants and Procedure

Twelve participants (eleven male, one female) were recruited at our university campus. Their age ranged from 24 to 35 years (M = 29.5), all were right-handed, and either native speakers or fluent in German. Seven had rarely used speech input before (0-10 times) whereas five were rather experienced with it (used >10 times or regularly). All were familiar with motion-based interaction (used >10 times or regularly).

They were first introduced to the various controls and could practice them in a simpler virtual setting. Therein, the users were motivated to test both modalities for all four tasks. This introduction took about five to ten minutes. Afterwards, they played the main scenario which lasted about 20 minutes, and they were free to choose either modality for any interaction they encountered. After completing the scenario, the participants filled in a questionnaire which asked for their preferred modality and their opinion on both input options for each task. The latter was done by rating the following statements on a five point Likert scale ranging from 1 (completely disagree) to 5 (completely agree): "It was difficult to recognize or remember the commands for the desired action", "the commands for these actions felt natural", "it was tiring to give the commands" and "the recognition worked reliably". In addition, recognized commands were automatically logged along with the chosen modality and the task they belonged to.

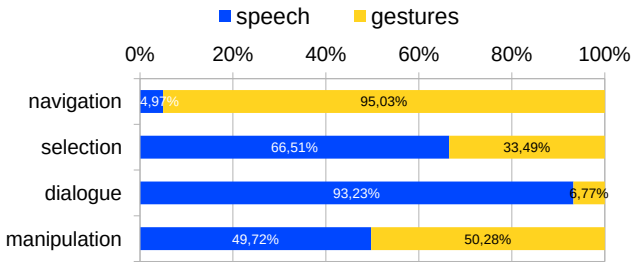

Figure 3: Average modality usage per task.

Results

Figure 3 depicts the average modality usage for the four interaction tasks as logged during the study. Our primary hypothesis that modalities would not be equally suitable for each task was confirmed by a Friedman's ANOVA (used as parts of the data were non-normally distributed) which showed that participants used different ratios of gesture and speech inputs for them ($\chi^2(3) = 30.18$, $p < 0.001$). In particular, Wilcoxon signed-rank tests (with a significance level of 0.0125 for Bonferroni correction) showed that a significantly higher percentage of gestures was used for navigation than for the three other tasks ($T = 0$, $p < 0.0125$, r = -0.62) and a significantly higher percentage of speech was used for dialogue compared to the other tasks ($T_{manipulation} = 1$, $T_{selection} = 0$, $p < 0.0125$, $r_{manipulation}$ = -0.61, $r_{selection}$ = -0.54).

For each of the tasks, Wilcoxon tests were used to compare the average usage and user ratings between speech and gestures. For the dialogue task, participants used significantly more speech utterances than gestures ($T = 0$, $p < 0.01$, r = -0.90). Speech was further rated as significantly less difficult to learn ($M = 1.08$, $SD = 0.29$) than gestures ($M = 2.25$, $SD = 1.22$; $T = 0$, $p < 0.01$, r = -0.74), it was considered to be more natural ($M = 4.92$, $SD = 0.29$) than gestures ($M = 2.83$, $SD = 0.94$; $T = 0$, $p < 0.01$, r = -0.86), less tiring ($M = 1.17$, $SD = 0.39$) than gestures ($M = 2.58$, $SD = 0.90$; $T = 0$, $p < 0.01$, r = -0.83), and more reliable ($M = 4.83$, $SD = 0.39$) than gestures ($M = 3.67$, $SD = 0.78$; $T = 0$, $p < 0.01$, r = -0.79). For the navigation task, we got a significantly higher usage of gestures than speech ($T = 0$, $p < 0.001$, r = -0.99) and a lower difficulty rating for gestures ($M = 1.42$, $SD = 0.67$) than for speech ($M = 2.50$, $SD = 1.24$; $T = 10.5$, $p < 0.05$, $r = 0.59$). We found no significantly different modality usages for the manipulation task, but a significantly better user rating for speech that was rated as less difficult to learn ($M = 1.25$, $SD = 0.45$) than gestures ($M = 2.67$, $SD = 1.07$; $T = 0$, $p < 0.01$, $r = 0.78$), less tiring ($M = 1.33$, $SD = 0.49$) than gestures ($M = 2.25$, $SD = 1.14$; $T = 0$, $p < 0.05$, $r = 0.70$), and more reliable ($M = 4.92$, $SD = 0.29$) than gestures ($M = 3.83$, $SD = 0.94$; $T = 0$, $p < 0.01$, $r = 0.75$).

The stated modality preferences are again in favour of gestures in the navigation task (11 preferred gestures, 1 preferred speech) and of speech in the dialogue task (preferred by all 12). Further, they indicate a tendential preference for speech in the selection task (7 preferred speech, 1 gestures, 4 were

undecided), but an equal distribution for manipulation (5 preferred speech, 5 preferred gestures, 2 were undecided).

Discussion

For navigation, our hypothesis in favour of gesture input was confirmed by its higher usage and stated preference, as well as the fact that gestures were rated as easier to learn. This is in line with Kadobayashi et al. (1998) who considered walking gestures to be more intuitive for navigation than using a mouse. However, there might be different results when using a navigation approach with direct target selection. For the selection task, we found no significant differences, only the stated preferences indicate a tendency for speech. One reason might be distinctions between the selection targets, as three participants mentioned that they liked to reach for an object with their hands whereas two preferred addressing characters by speech. Different sizes and placements of the objects might have further influenced the modality choice, as some objects were more difficult to point at than others, similarly observed by van der Sluis and Krahmer (2007). The hypothesis that speech would be preferred for dialogue as derived from (Cavazza et al., 2004) was clearly confirmed. All participants named it as their preferred modality, it was used most of the time with nine participants even using it for every single sentence, and the user ratings were very positive with all items close to the extremes. Apart from this clear result, it has to be mentioned that there exist dialogue utterances that can be naturally represented by gestures, e.g. nodding for "yes" or a greeting gesture for "hello", but this is not the case for arbitrary sentences. We assumed a preference of gestures for object manipulation, but this hypothesis could not be confirmed as both modalities were used with almost equal preference and the user ratings were even in favour of speech. A similar variety of modalities has been observed by Corradini and Cohen (2002), who additionally reported that users preferred to use both in a multimodal way. Hints for another explanation were observed during our study, as the users seemed to follow two different behaviour types. Speech users seemed to be more focused on progressing, often calling the actions as soon as they appeared on screen instead of first watching the gesture animations to figure out how to perform them. On the other hand, gesture users seemed to perform the task in a consciously more natural way and some also exhibited role-playing behaviour such as worrying about being heard by the virtual characters. Therefore, interaction designers should investigate the preferences of their target group and decide between a more natural and engaging object manipulation using gestures or a faster one using short speech commands.

Conclusion and Future Work

In this paper, we examined which modality users prefer for four main interaction tasks in a virtual environment. We conducted a study on a system in which we successfully implemented all four interaction tasks with real-time recognition for both speech and body gesture input using low-cost technology. It was confirmed that a gestural walking metaphor suits navigational tasks better while speech was chosen for dialogues. For selection and manipulation, no clear preference was obtained, but we observed possible reasons for the different modality choices between the users.

As a next step, we plan to include multimodal fusion of speech and gesture input. This could enhance all current interaction tasks, e.g. navigation with directly indicating a target could be integrated with pointing at it and saying "move to this location". In particular, selection and manipulation should be examined further, as we did not find primary modalities for them. Apart from adding multimodality, another possibility would be to omit the unambiguous labels, so that users have to select entities by describing their properties instead of simply naming them. Further, virtual representations of the user's hands could be used for manipulating objects in a more direct and immersive way.

Acknowledgments

This work was partially funded by the European Commission within the 7th Framework Program under grant agreement eCute (FP7-ICT-257666).

References

Cavazza, M., Charles, F., Mead, S., Martin, O., Marichal, X., & Nandi, A. (2004). Multimodal acting in mixed reality interactive storytelling. *Multimedia, IEEE*, *11*(3), 30 - 39.

Cohen, P., McGee, D., Oviatt, S., Wu, L., Clow, J., King, R., et al. (1999). Multimodal interaction for 2D and 3D environments. *Computer Graphics and Applications, IEEE*, *19*(4), 10 -13.

Corradini, A., & Cohen, P. (2002). On the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence. In *Proceedings of the international CLASS workshop on natural, intelligent and effective interaction in multimodal dialogue systems.*

Kadobayashi, R., Nishimoto, K., & Mase, K. (1998). Design and evaluation of gesture interface of an immersive walkthrough application for exploring cyberspace. In *Proceedings of the third IEEE international conference on automatic face and gesture recognition* (p. 534 -539).

Kistler, F., Endrass, B., Damian, I., Dang, C., & André, E. (2012). Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces*, *6*, 39-47.

Kistler, F., Sollfrank, D., Bee, N., & André, E. (2011). Full body gestures enhancing a game book for interactive story telling. In *Interactive storytelling* (Vol. 7069, p. 207-218). Springer Berlin / Heidelberg.

LaViola, J. J., Jr., Feliz, D. A., Keefe, D. F., & Zeleznik, R. C. (2001). Hands-free multi-scale navigation in virtual environments. In *Proceedings of the 2001 symposium on interactive 3d graphics* (pp. 9–15). New York, USA: ACM.

van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, *44*(3), 145-174.