All Birds Must Fly: The Experience of Multimodal Hands-free Gaming with Gaze and Nonverbal Voice Synchronization

Ramin Hedeshy[*]
University of Stuttgart
Stuttgart, Germany
ramin.hedeshy@ipvs.uni-stuttgart.de

Chandan Kumar[*]
Fraunhofer Institute for Industrial Engineering IAO
Stuttgart, Germany
chandan.kumar@iao.fraunhofer.de

Mike Lauer[*]
University of Stuttgart
Stuttgart, Germany
st155614@stud.uni-stuttgart.de

Steffen Staab
University of Stuttgart
Stuttgart, Germany
University of Southampton
Southampton, UK
s.r.staab@soton.ac.uk

ABSTRACT

Eye tracking has evolved as a promising hands-free interaction mechanism to support people with disabilities. However, its adoption as a control mechanism in the gaming environment is constrained due to erroneous recognition of user intention and commands. Previous studies have suggested combining eye gaze with other modalities like voice input for improved interaction experience. However, speech recognition latency and accuracy is a major bottleneck, and the use of dictated verbal commands can disrupt the flow in gaming environment. Furthermore, several people with physical disabilities also suffer from speech impairments to utter precise verbal voice commands. In this work, we introduce nonverbal voice interaction (NVVI) to synchronize with gaze for an intuitive hands-free gaming experience. We propose gaze and NVVI (e.g., humming) for a spatio-temporal interaction applicable to several modern gaming apps, and developed 'All Birds Must Fly' as a representative app. In the experiment, we first compared the gameplay experience of gaze and NVVI (GV) with the conventional mouse and keyboard (MK) in a study with 15 non-disabled participants. The participants could effectively control the game environment with GV (expectedly a bit slower than MK). More importantly, they found GV more engaging, fun, and enjoyable. In a second study with 10 participants, we successfully validated the feasibility of GV with a target user group of people with disabilities.

CCS CONCEPTS

• **Human-centered computing → Pointing**; *Accessibility technologies*; *Interaction design.*

KEYWORDS

game interaction; eye tracking; nonverbal voice inputs; humming

ACM Reference Format:
Ramin Hedeshy, Chandan Kumar, Mike Lauer, and Steffen Staab. 2022. All Birds Must Fly: The Experience of Multimodal Hands-free Gaming with Gaze and Nonverbal Voice Synchronization. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3536221.3556593

1 INTRODUCTION

Computer games as a source of entertainment have become a part of our lives. We use games to enhance education and to practice various types of cognitive learning strategies [32]. For instance, we can exploit games to reinforce the creativity of learners [16]. Games can involve players in forming playing strategies to solve problems, thus enabling players to practice persistent problem solving [18]. Games can also help players to develop organizational and systemic thinking skills [20].

There have been new interaction mechanisms and modalities introduced to make games more interactive, engaging, and fun (e.g., gaze, gestures, swipe, etc.). However predominately the basic control always relies on the hand (or finger) movement, while the other mechanism aims to enrich the experience of players. In this work, we focus on complete hands-free control, to bring a novel gameplay experience for end users. Moreover, the goal is to support inclusive interaction that enables users with limited motor control for gameplay experience, i.e., hands-free input methods are essential for those who suffer from impairment (e.g., quadriplegic or Parkinson) and have difficulty in operating their hands. To support hands-free interactions, eye gaze [26] and vocal input [43] are the two most natural candidates.

Voice has been used as an input modality in various user interfaces. Speech recognition is nowadays used in games and other applications. While speech commands are powerful, they also come with drawbacks. In fast-paced games, the delay between issuing a voice command and the reaction of the game is a noticeable disadvantage [29, 34]. Moreover, performing verbal speech command is not possible for people suffering from speech disorders. An alternative to recognizing spoken commands is nonverbal commands,

[*] Authors contributed equally to this research. The author's list is sorted alphabetically by last name.

ICMI '22, November 7–11, 2022, Bengaluru, India
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556593

also characterized as nonlexical [40], e.g., humming, or whistling [19], using sound volume, sound pitch, and other features of voice for immediate, real-time control. Nonverbal voice interactions do not require expensive recognition processes and can be performed faster, reveal less semantic content to third parties, and some are even possible with a closed mouth e.g., buzzing or humming [35].

The use of eye gaze as an input modality in games has been also researched with questions like how could eye trackers support gameplay experience. In this regard, most of the previous works have used eye tracking to understand and evaluate user attention to improve the gameplay experience [36]. There are also several approaches which actively use gaze to adapt the game environment, such as to provide automatic orientation [5, 33], or social cues of player views [22]. Eye tracking as a control mechanism to move and object or avatar have also been tested in some experimental studies like first-person shooter games [13, 15], chess [44], puzzles [6, 42], fight simulators [28]. However, it has been found very difficult for end users to observe, perceive, and control only using gaze. Hence most of the modern applications use gaze in a multimodal setting which often diminishes the significance of gaze as hands-free interaction mechanism.

In this work, we investigate how gaze and voice can effectively be harmonized to support each other to provide complete hands-free control while offering an engaging gameplay experience. In general, spatial and temporal interaction is imperative in gameplay, and their coordination is a crucial aspect. Hence, we utilize gaze for its natural orientation characteristics, i.e., to give direction, and the temporal characteristic of nonverbal voice (such as continuous humming), i.e., to provide movement. We developed a 2D game environment "all bird must fly" to demonstrate and evaluate the synchronization of these modalities. To the best of our knowledge, this is the first research to combine gaze with nonverbal voice for game interaction.

We have conducted two studies, a study with 15 participants to assess the proposed multimodal method's efficacy and engagement compared to mouse-keyboard control as a baseline. In the second study, we investigated whether gaze and nonverbal voice-controlled game interaction is a viable alternative for people with motor impairment through a feasibility test with the target group.

In the following, we summarise the relevant background information and review the state of the art concerning gaze and voice in gaming in Section 2. The design and implementation of the game are presented in Section 3. Section 4, describes the experimental design of the user evaluation, and Section 5 highlights the results obtained. The discussion is represented in section 5. Finally, in Section 6, conclusions are drawn and related future work is discussed.

2 RELATED WORK

2.1 Eye Tracking in Computer Games

Eye gaze has been evaluated as a means of human-computer interaction for decades [14, 41]. Since then, it has been exploited as an input modality in games and a variety of other applications. For example, research proposed the use of gaze to control a player's orientation as they explore the virtual world [5, 33]. Leyba and Malcolm [23] evaluated the eye gaze performance as an aiming device in computer games. Several studies examined the performance of eye tracking in first-person shooter games [13, 15]. A minimalistic 3D flying game that only required steering was examined in a study called "Gaming with Gaze and Losing with a Smile" [28]. In spite of being harder to play the game by gaze, gaze interaction scored significantly higher than mouse regarding entertainment and engagement. Gowases *et al.* [6] developed a puzzle game playable using either mouse or gaze dwell-time. They also reported a higher subjective immersion during gaze-based problem-solving trials. Overall, although many of the aforementioned research acknowledge either similar or lower game performance when players use gaze control in a game compared to conventional control methods, studies reported a higher game engagement when the interface was controlled by gaze. This indicates that regardless of the performance, people are interested in alternative forms of computer interactions. Isokoski *et al.* [12] provided a comprehensive overview of gaze-controlled games as well as the implications and challenges of using gaze input in games. They stated that most games cannot currently be played efficiently using eye tracker input alone and that solutions for gaze-based velocity control and trigger operation are needed.

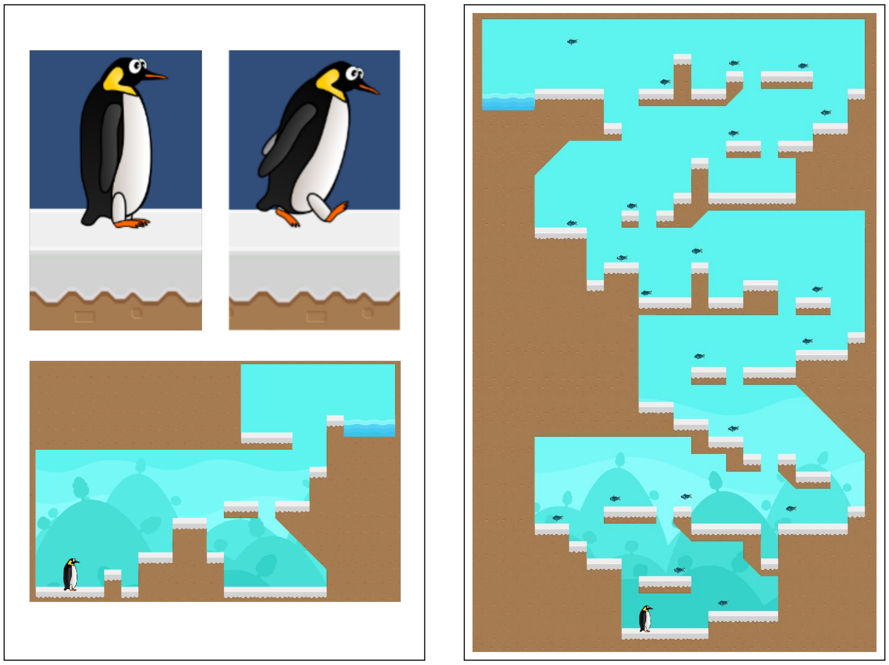
2.2 Voice in Computer Games

Research on voice control of digital games has been undertaken since at least the 1970s [1]. Allison *et al.* provides an extensive overview of video games that incorporate voice recognition in [1, 2]. Here we survey only a few that stand out to be more related to this study. There are several studies focused on speech rehabilitation exercises [4, 21, 24, 27], and improving engagement [31]. Similar to this study, other research has explored the implementation of voice input as an alternative game control scheme, to enable access for players with motor impairments or other disabilities that prevent them from using physical controls [8, 34].

Several studies have investigated controlling games through non-speech features of voice, such as the volume [39] and pitch [7, 10, 34] of vocal input. In a study comparing both types of voice commands using the game Tetris, nonverbal voice commands were much more efficient and accurate than traditional speech recognition [34]. Harada *et al.* [8] conducted a quantitative experiment to determine the performance characteristics of non-speech vocalization for discrete input generation in comparison to existing speech and keyboard input methods. The results from their study validated that non-speech voice input can offer significantly faster discrete input compared to a speech-based input method by as much as 50% [8]. The mobile game "Scream Go Hero" from *Ketchapp* [17] uses quiet and loud noises as inputs in a jump-and-run style game. The app reached over ten million downloads in the *Google Play* store which indicates the popularity of such an interaction technique. However, the use of unimodal voice input limits the complexity of the game. Therefore, these approaches lend themselves to relatively simple game mechanics such as one-dimensional movement.

2.3 Eye Gaze and Voice in Computer Games

There are only several studies proposing a multimodal use of gaze and speech in gaming. Wilcox *et al.* [42] created a third-person adventure puzzle game that could be controlled by both gaze and speech and by gaze alone. They used a focus group to measure the



(a) Game Character and Training Level

(b) Main Level

Figure 1: Example images from the All Birds Must Fly game.

effects of the interface control method but they did not conduct a user study. Another gaze-voice controlled game is "Rabbit Run" [29], which is evaluated against a keyboard-mouse combination. Despite their worse game performance in gaze-voice control, their participants reported a higher level of immersion for gaze-voice control compared to keyboard-mouse control. Uludagli and Acarturk [37] conducted a comparative study of gaze-voice and touchscreen interface control. They found that the participants exhibited higher game performance when they used the touchscreen control and that participants who used the gaze-voice control method were more psychologically absorbed. Van der Kamp *et al.* [38] investigated the performance of gaze and voice for controlling the cursor in a computer drawing software. Their participants reported the feeling of having less control, speed, and precision compared to control by mouse and keyboard. We have not found any study that combines eye gaze and nonverbal voice interactions in computer games. To the best of our knowledge, the only study which proposes a multimodal use of gaze and nonverbal voice control is applicable to a completely different interaction scenario of gaze-based text entry [9].

In summary, the related work has identified and studied gaze and voice as a promising addition to the gameplay experience. The main limitation of gaze is its inability to act as explicit control in the game environment, while the verbal commands of voice input are limited due to the speech recognition latency which affects the real-time performance. So far, it is unclear how these two modalities can be effectively combined to provide complete hand-free control. Therefore, in this work, we investigate how eye gaze as an orientation mechanism and a robust version of voice input (nonverbal commands) are harmonized for effective game control, which can be feasible and applicable for modern gaming apps requiring spatio-temporal interactions.

3 ALL BIRDS MUST FLY: GAME DESIGN

The primary goal of this work is to explore if gaze and nonverbal voice can provide effective game control with an immersive and fun experience, while being a viable alternative to conventional input like mouse and keyboard, i.e., to support people with motor impairment. Therefore, we chose the most common 2D game environment, with the underlying phenomena of objects or player avatars being controlled with a goal to succeed. The most common interaction style is spatio-temporal, which combines both direction and distance attributes e.g., the simple but popular *Jump and Run* design. There are many different games based on this principle, one of which is the famous *SuperMario* [1].

The game premise was chosen as "All Birds Must Fly", representing a penguin which is a bird that technically cannot fly, however, to reach their objectives they comprehend and capitalize on their other abilities, i.e., long jumps. We believe that this inherently reflects the motivation and courage of several people with disabilities, in enhancing their special abilities to compete and achieve their objectives. In the proposed game, a penguin (player avatar) is at the bottom of a pit, and its objective is to reach the top by walking and jumping through several hurdles. Figure 1 presents the

[1] https://supermario-game.com/

game character, the training map, and an overall view of the game interface.

During the gaze-voice-controlled gameplay, the player is able to jump or walk to the (either right or left) by simply looking at the intended landing target and making a humming or buzzing sound. The control of the game interface by mouse and keyboard, as an alternative to the gaze-voice control of the game, involves pressing the space bar to control the jump or walk actions of the penguin and targeting the landing spot by mouse. To earn extra points, the players could collect fish distributed throughout the map. The final game design, including the difficulty levels, challenges, rewards, and interaction smoothing was the outcome of a user-centered design, i.e. the feedback from participants in several pilot studies.

Gaze was used as its conventional mode of tracking users' visual focus and emulating automatic movement like a mouse. However, integrating nonverbal voice in gameplay is a novel phenomenon. Allison *et al.* [2] provided 25 design patterns for voice interactions in games. As we use gaze for navigation, input was required for triggering the jump actions. *Volume Control* as the simplest form of voice interaction would have been the easiest choice. However, as stated by Allison, it is a difficult task for players to maintain precise control over their voice volume. We opted for the *Pitch Control* pattern as it brings greater flexibility. Players can use quieter voice inputs to minimize fatigue and social embarrassment. However, we simply consider a continuous humming pitch to trigger the jump and do not match the pitch changes to any special actions as seen in the related works e.g., [34] that use this design pattern. Non-verbal gestures are recognized as short melodic patterns of defined pitch profile and length. Usually, a set of tone gestures for individual commands or operations is defined to cover different operations. However, since we also utilize gaze, we only needed one tone gesture. A tone with continuous pitch has been assigned to command moving toward users' gaze. We used humming as it is easy and comfortable to produce. In contrast to voice-based input, (i) humming can be detected easily, (ii) introduces no privacy issues, and (iii) can even be performed by many whose speech is impaired. Besides, Unlike speech, humming involves no continuously gliding pitch movement but consists of rising and falling pitch steps. We have therefore decided to use humming for the nonverbal control in this study. For measuring the pitch of humming, we have used autocorrelation, as described in [30].

4 EXPERIMENT

We conducted two studies. Firstly, a user study was performed to evaluate the performance of the gaze plus nonverbal voice (GV) compared to the mouse and keyboard (MK) performance. Then, we conducted a feasibility study with people with motor and partly also speech impairments to investigate the practical usability for the target group.

4.1 Study 1: Comparison of Input Modalities.

4.1.1 Participants. Six male and nine female participants from the campus voluntarily participated in the first study, with a mean age of M = 24.4 years (standard deviation [SD] = 2.82). Five of the participants stated being experienced in video games, while the other ten stated having low to medium experience. None of the participants had previously used eye tracking or voice-activated interactions.

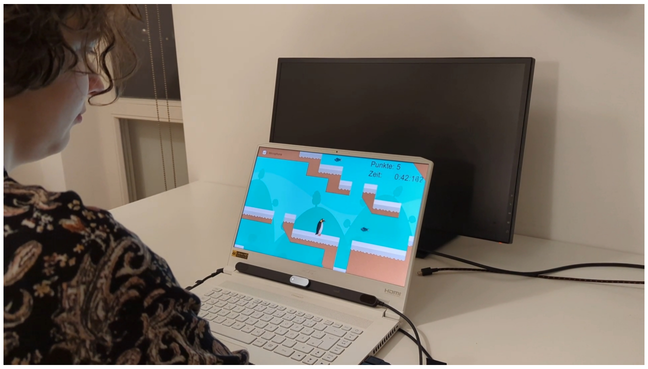


Figure 2: Experimental setup: A participant performing the experiment on a laptop computer equipped with an eye tracker.

4.1.2 Procedure. The first study took place at the university lab. After the calibration of the eye tracker, they were presented with a video tutorial to get familiar with the game environment and the interaction controls. Before the two main sessions, participants also took their time in the training level to familiarise themselves with the controls, (Figure 1). The aim of the training session was also to decrease the chance of having a training effect during the main sessions. Moreover, to offset the learning effects, the two input methods were counterbalanced with participants divided into two groups. After the tutorial session, the participants played the game twice with one of the control methods (either by gaze-voice control or by mouse-keyboard control). Finally, the questionnaire form was immediately filled in for the corresponding method. After filling the questionnaire, participants were shown another tutorial explaining the other control method, and repeated the procedure for the other method. The whole experimental session took approximately 45 minutes.

4.1.3 Apparatus. We used the Eye Tracker 4C from Tobii and attached it to a laptop with a 15 inch screen. A calibration was carried out for each participant to ensure that the collected data would be reliable. The laptop's built-in microphone array was used for audio capture. The microphone sensitivity was adjusted to filter as much background noise as possible while maintaining accurate voice detection. The recognition threshold was adjusted accordingly for each participant. Gaze was recorded with a tracking frequency of 90 Hz. No chin rest was used. The eye tracker was placed at the lower edge of the screen. See Figure 2 for a picture of the setup. The eye-tracker tracking-box dimensions as reported by the manufacturer were 16" × 12" / 40 cm × 30 cm at a distance of the head of 29.5" / 75 cm. A mouse was also connected to the laptop to allow for mouse input required when using the MK control method.

The game was developed using the Unity game engine. The game interface was similar using the both interaction methods differing only in used control method.

As for the questionnaire, We have extracted questions from well-structured game experience questionnaires [3, 11], and the

Measure	MK (Study 1)	GV (Study 1)	Related t-test (MK vs GV)	GV (Study 2)
Jump Success Rate	M=96.41 SD=3.09	M=94.1 SD=3.31	t=1.6251 p=.1153	M=95.19 SD=2.12
Time to reach finish line (s)	M=91.5 SD=4.05	M=163.44 SD=56.1	t=-4.0203 p<.001	M=354.13 SD=141.04
Idle time between actions (s)	M=0.460 SD=0.2	M=0.94 SD=0.2	t=-6.4817 p<.001	M=1.62 SD=0.42
Proportion of motion in total time	M=0.65 SD=0.1	M=0.49 SD=0	t=5.2131 p<.001	M=0.38 SD=0.1
Walking percentage	M=0.21 SD=0.1	M=0.25 SD=0.1	t=-1.1668 p=.2531	M=0.18 SD=0.1
Duration of temporal action (s)	M=0.64 SD=0.08	M=0.72 SD=0.1	t=-2.1455 p=.0407	M=0.75 SD=0.21

Table 1: Means over two sessions for using *MK* compared to *GV*. The last gray column shows the results from the second study with people with motor and speech impairments.

related work [29] to assess the proposed interaction technique. We followed the GEQ guidelines following the cited literature along the dimensions of enjoyment, immersion, challenge, etc. We selected the appropriate statements, e.g., "I was deeply concentrated in the game", "I lost track of time", "I felt exhausted", etc. For the control element, we presented some precise questions related to the environment, e.g., "the avatar (penguin) moved exactly the way I wanted". We asked questions to ascertain the input feasibility, e.g., "Did you run out of breadth?", and overall comparison of the modalities "how would you rate the gameplay experience with mouse and keyboard". The questionnaire is represented in appendix part A.

4.1.4 Result. The results and the analyses of study 1 is presented in the parts below, for game performance and for questionnaire results separately. We plotted a histogram and also tested the collected data for normal distribution with a one-sample Kolmogorov-Smirnov test. Paired t-tests were used to see if there were statistically significant differences between the quantitative performance measures in the two methods. The results of the statistical analysis can be seen in Table 1.

We defined the success rate of the players as the inverse ratio of the number of failures (i.e. the number of unsuccessful jumps) to the total number of jumps. Although a higher success rate was achieved using *MK* on average, the differences were not statistically significant, ($MK \approx 96\%$, $GV \approx 94\%$, $p < .1153$). The Cohen's d effect size for success rate is 0.72. Figure 3 illustrates the mean jump success rate and its standard deviation for each session and method.

Participants were 44% faster using *MK* (1:32 min) compared to *GV* (2:43 min), which results in a large Cohen's d effect size of 1.80. The learning effect is not significantly evident with the mean completion time decreasing from 1:41 to 1:22 using *MK*, ($p = .2547$). However, it is noticeable using *GV* from 3:06 to 2:18, ($p = .0506$). This is expected for *MK* as participants are familiar with this input. However, as for *GV*, it indicates that the method is intuitive and does not require a long training time. The mean completion times for each input method over the two sessions are shown in Figure 4.

Participants collected all the 21 distributed fish on the map except one participant who missed collecting two fish when using the *MK*. However, this failure did not cause a significant confound in the timing analysis.

We collected the time participants were idle between their movements to discover the extra time players need for resting between their actions using *GV*. The average idle time using *MK* is about half a second and approximately a second using *GV*, ($p < .001$). There is no evident of significant improvement between the sessions ($p_{MK} = .7020$, $p_{GV} = .0897$).

Relatively, the proportion of motion in the total time is significantly larger using *MK* compared to the hands-free variant, ($MK \approx 65\%$, $GV \approx 49\%$, $p < .001$), which means that players were on average more active when playing with mouse and keyboard.

The mean percentage of walking movements was approximately the same across all modalities. With 21% for *MK* and 25% for *GV*. Participants often mentioned that walking using *GV* was easier, although the difference is not significant.

During the experiment, We observed that the participants made longer temporal inputs with voice control than with the keyboard. This is also reflected in the data: the average time of action, i.e. the duration for which a key was pressed, lasted 650ms with the keyboard and 724ms with voice control ($p = .0407$).

4.1.5 Subjective feedback. We solicited participants' feedback in two categories: *Experience*, and *Control*. Questionnaire responses were on a five-point scale. Figure 5 shows the mean scores given by

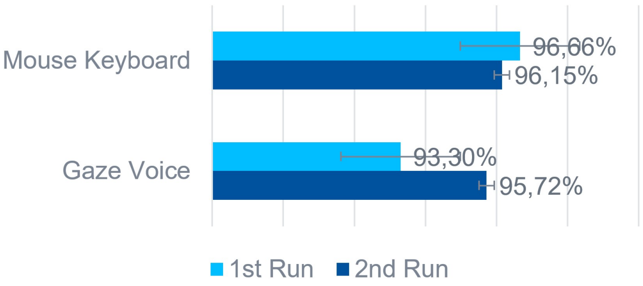


Figure 3: Jump success rate for each method and session in Study 1, plotted as bars. Error margins indicate the standard deviation.

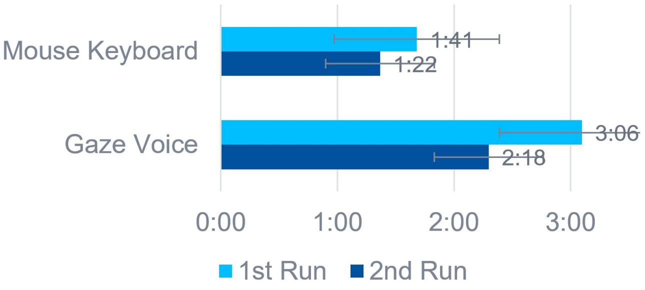


Figure 4: Mean completion time for each method and session in Study 1, plotted as bars. Error margins indicate the standard deviation.

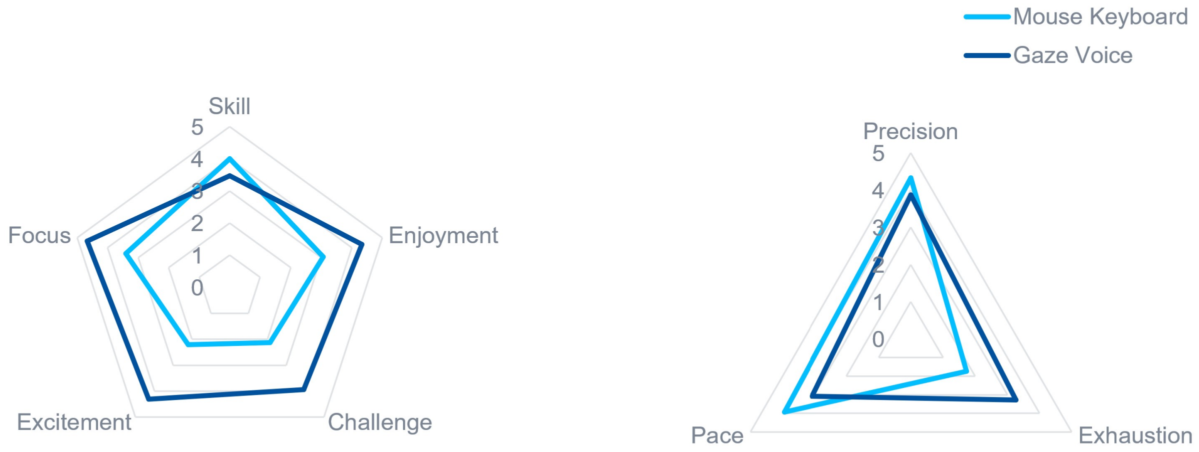


Figure 5: Subjective response (1 to 5) by method and questionnaire item in study 1.

the players for each group and for the two game interface control method.

When asked how well the participants thought they had played, the conventional *MK* ($M = 4$, $SD = 0.75$) performed slightly better ($t = 1.47$, $p = .15$) compared to *GV* ($M = 3.47$, $SD = 1.19$). However, the players found *GV* control methods significantly more enjoyable ($t = -4.29$, $p < .001$). As it is shown in the radar chart, *GV* is also considered to be more challenging $t = -4.71$, $p < .001$), and exciting ($t = -6.96$ $p < .001$). Participants claimed that they needed more concentration while playing the game using the *GV*($M = 4.67$, $SD = 0.49$) than the *MK*($M = 3.4$, $SD = 0.98$).

There was no significant difference ($t = 1.20$, $p = .24$) between the control types in the category of feeling the time (5). The time spent playing the game passed at a similar average rate for *MK* ($M = 3.67$, $SD = 0.97$) and *GV* ($M = 3.2$, $SD = 1.15$), but responses for the hands-free variant were more widely distributed.

Moreover, we also asked the participants about their overall input type preference. Ten participants selected *GV* as their preferred method, and five opted for *MK*.

4.2 Study 2: Practical Usability Evaluation.

4.2.1 Participants. We reached out to institutions such as special schools and associations for people with disabilities and found ten participants. Five male and five female participated, with a mean age of M = 19.5 years, vary between 12 to 57 years old (standard deviation [SD] = 13.93). One participant had previously used an eye tracker but not as a means of game control input. two of the participants had the Duchenne Muscular Dystrophy disorder, two had difficulties due to Cerebral Palsy and two reported having Spastic Tetraparesis. Besides, three of them reported suffering form speech disorders. However, it did not hinder them from performing NVVI and participating in the experiment. Nonetheless, four participants could not play the game due to eye twitches problem in addition to a participant who needed a ventilator for breathing.

4.2.2 Procedure. The second study with people with disabilities took place at the participants' homes, as it was more convenient for some participants due to the pandemic, and their disabilities in particular. The participants only tried the *GV* method and answered the questions about the hands-free variant as they were not able to play the game using the mouse-keyboard control.

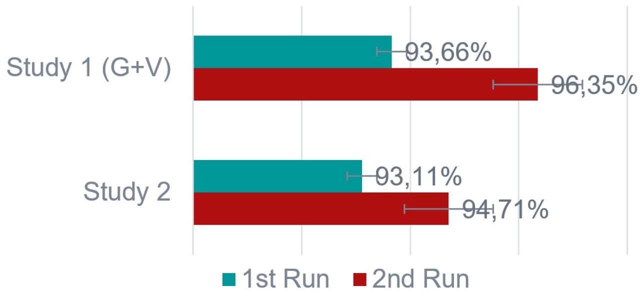


Figure 6: Jump success rate for each method and session in Study 2, plotted as bars. Error margins indicate the standard deviation.

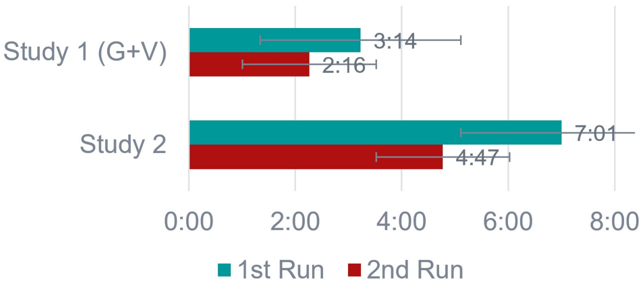


Figure 7: Mean completion time for each method and session in Study 2, plotted as bars. Error margins indicate the standard deviation.

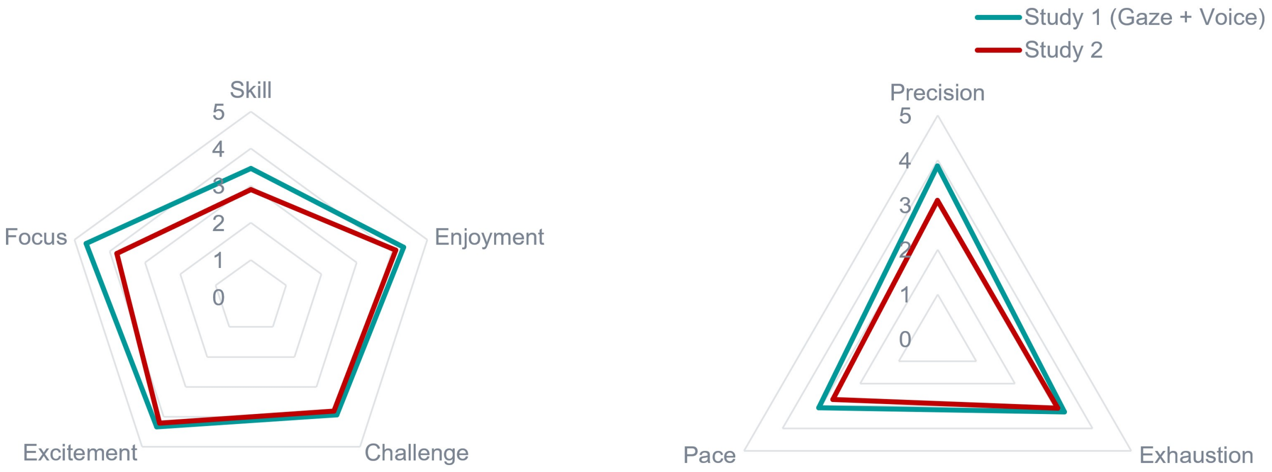


Figure 8: Subjective response (1 to 5) compared to Study 1.

4.2.3 *Apparatus.* The same apparatus and the same technique were used as in Study 1. The questionnaire was almost the same as that of the first study. The only difference is that the questions about comparison with mouse and keyboard were omitted.

4.2.4 *Result.* Overall, the average time required by participants to complete the game is 5 : 54 *min*, about two minutes slower than the non-disabled participants of Study 1. Although the participants improved on average from 7 : 01 *min* in the first session to 4 : 47 *min* in the second session, the improvement was not statistically significant ($p = .2085$). The mean completion time achieved by participants with motor impairments in Study 2 compared to the mean completion time achieved by the non-disabled participants in Study 1 using the hands-free proposed method are shown in Figure 7.

The participants also had a slightly lower mean success rate than the non-disabled participant in Study 1. The mean success rate for *GV* method in Study 2 compared to the corresponding values in Study 1 over the two sessions are shown in Figure 6.

The proportion of motion in the total time in Study 2 is about 38% compared to 48% in Study 1 ($p = 0.012$). The participants in Study 2 rested about 1.6 seconds, whereas those using *GV* in Study 1 rested about 1.0 seconds before issuing the next command.

The players in both studies were very similar in the way they moved. The proportion of walking to all movements in the first and the second study is 19% and 18%, respectively ($p = .7791$). Similarly, the mean duration of temporal actions in both studies are very similar, 743*ms* in Study 1, and 755*ms* in Study 2, ($p = .9047$).

Two participants failed to collect all the distributed fish on the map, each missing one behind, yet they both had a slower completion time compared to the mean.

4.2.5 *Subjective feedback.* Figure 8 shows the mean scores given by the participants with disabilities for each group. To compare the results with the first study, the participants' feedback from the first study is also presented on the radar chart.

On average, the participant rated their overall performance 2.9 from 5 ($SD = 0.74$). Similar to the first study everyone unanimously found the game enjoyable ($t = 0.73$, $p = .4748$) with an average score of 4.1 ($SD = 0.75$). Furthermore, all participants were of the same opinion that the game is quite a challenge with the *GV* control system ($M = 4$, $SD = 1$) but exciting ($M = 4.2, SD = 0.79$).

As for control precision, the average response is about 3.1 ($SD = 0.57$), which also corresponds to our observation. The lower score can be attributed to several factors based on individual impairments. For instance, two participants suffering from squinting and twitching of the eyes were not able to accurately use the eye tracker.

Moreover, we have asked the participants for comments about the effort required by the game controls technique. The mean score in study 1 and study 2 is 3.27 ($SD = 1.1$) and 3.1 ($SD = 1.45$), respectively. This indicates that despite the physical limitations, participants found gaze and continuous voice control a feasible hands-free interaction method.

5 DISCUSSION

The results of the quantitative analysis show a better performance and higher accuracy for *MK* compared to *GV*, besides the fact that the participants are much more familiar with *MK* and never used eye tracking or nonverbal voice commands preceding this study. The results of the subjective feedback, on the contrary, are in favor of *GV*. Two-third of the participants in the first study preferred the multimodal hands-free version and described it as much more fun. The participants found *GV* immersive and more enjoyable than *MK*. Furthermore, in the study with people with disabilities, all participants found the game enjoyable. A participant commented on the *GV* as being imaginable in other applications and found such a computer interaction method useful in their life. It is interesting to note that eye tracking and nonverbal interactions are both novel inputs for most users. Also, it was surprising for us to see that majority of the participants with disabilities have never used even a basic eye tracker that can be their only way to communicate and interact. Therefore, We believe that, with more practice, users will achieve higher performance than those reported in this paper.

Although the participants were overall positive with *GV*, yet, sometimes they struggled to synchronise their gaze and voice inputs.

One participant stated, "it is difficult not to look around and at the penguin while doing the jump". Another issue was that many participants had initially tried, both consciously and unconsciously, to control the penguin with head movements. Head movements do not break the gaze tracking when using the new eye trackers. Nevertheless, head movement will still have an impact on the data if the subject moves too fast. Another complication with the real-world deployment of the proposed approach would be the detection of voice inputs with external audio events, e.g., talking, coughing, sneezing, or ambient noise. No sound effects or music were used in this game to eliminate the effect of false positive recognition. Further work could classify humming and distinguish it from noise to make the approach more robust.

Different designs and modalities as well as different target groups make the comparison of studies difficult. Nevertheless, similar to the prior works, we have compared our method against the traditional use of a mouse and keyboard to provide a baseline for comparison. Rabbit run [29] reported an issue with voice recognition being slow. The speed of their participants was on average 3.3 times slower playing using *GV*. In our study, mean completion time using nonverbal *GV* is 44% slower than *MK*. This indicates that nonverbal voice inputs can be a quicker alternative to verbal voice commands as they can be recognized easier and performed faster by players.

The comparison of verbal and non-verbal inputs is well documented in the related work. Sporka *et al.* [34] compared the performance of non-speech input and speech recognition for real-time game control. Their result shows that non-speech input excelled in both time and accuracy, their participants were on average 2.5 times faster. Speech recognition is slower because the recognition system has to wait for the end of utterance detection. We have to also wait for some silence. The paper from amazon [25], suggests reducing the lower threshold below 400ms negatively affects the performance due to an increase in early end-point rate. In our work, synchronizing a simultaneous gaze and speech command would not have been feasible since the gameplay requires instant responsiveness. For continuous nonverbal voice recognition, we do not need to wait for the end of an utterance. We break the voice stream into small window frames parsing them continuously. We also do not need to wait for silence. Consequently, there is almost no delay. A brief test revealed a mean recognition time of approximately 40ms, which is by far less than the reasonable 0.2 seconds response time.

6 CONCLUSIONS AND FUTURE WORK

This work presents a multimodal hands-free video game interaction method based on eye tracking and nonverbal voice inputs. To the best of our knowledge, this is the first study that combines eye tracking with nonverbal voice inputs for game interactions. Besides, this is the first study performed evaluating mouse and keyboard versus gaze and NVVI in addition to a feasibility test with people with speech and motor impairments. The participants finished the game slower using mouse and keyboard. However, the qualitative responses and explicit feedback indicate a clear preference for gaze and NVVI as an exciting, engaging, and fun game interaction method.

For the experiment, we designed a 2D game with the aim of evaluating the interaction method. However, it will be more natural to use gaze input in first-person games since the player shares the same view as the character. In this demo game we only used humming to trigger the jump action. However, other kinds of nonverbal voices e.g, whistling can be defined by the user for triggering other actions. A further study could differentiate the length, pitch, and tone of the player's voice to enable more complex interactions. This interaction technique in general can be used as a handsfree alternative in interfaces operatable by single-touch interactions e.g., general interactions such as swiping, scrolling, map exploration, etc. We envision further work to utilize such interaction method based on gaze and nonverbal voice inputs in more complex games and other exciting applications in the domain of communication, and virtual or augmented reality.

ACKNOWLEDGMENTS

We would like to thank our colleague Raphael Menges (University of Stuttgart) for his technical advises. We would also like to thank all the participants of the experiment and the administration team of the Analytic Computing department of Stuttgart University for their support. This research was partially funded by BMBF,[2] a funding program of the federal ministry of education and research under grant no. 16DHBKI041.

REFERENCES

[1] Fraser Allison, Marcus Carter, and Martin Gibbs. 2020. Word play: a history of voice interaction in digital games. *Games and Culture* 15, 2 (2020), 91–113.

[2] Fraser Allison, Marcus Carter, Martin Gibbs, and Wally Smith. 2018. Design Patterns for Voice Interaction in Games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 5–17. https://doi.org/10.1145/3242671.3242712

[3] Jeanne H Brockmyer, Christine M Fox, Kathleen A Curtiss, Evan McBroom, Kimberly M Burkhart, and Jacquelyn N Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of experimental social psychology* 45, 4 (2009), 624–634.

[4] Sandra Cano, Victor Peñe nory, César A. Collazos, Habib M. Fardoun, and Daniyal M. Alghazzawi. 2015. Training with Phonak: Serious Game as Support in Auditory – Verbal Therapy for Children with Cochlear Implants. In *Proceedings of the 3rd 2015 Workshop on ICTs for Improving Patients Rehabilitation Research Techniques* (Lisbon, Portugal) *(REHAB '15)*. Association for Computing Machinery, New York, NY, USA, 22–25. https://doi.org/10.1145/2838944.2838950

[5] James Gips, Philip DiMattia, Francis X Curran, and Peter Olivieri. 1996. Using eagleeyes—an electrodes based device for controlling the computer with your eyes—to help people with special needs. In *Proceedings of the 5th International conference on Computers helping people with special needs. Part I.* 77–83.

[6] Teresia Gowases, Roman Bednarik, and Markku Tukiainen. 2008. Gaze vs. mouse in games: The effects on user experience.

[7] Perttu Hämäläinen, Teemu Mäki-Patola, Ville Pulkki, and Matti Airas. 2004. Musical computer games played by singing. In *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04), Naples*.

[8] Susumu Harada, Jacob O. Wobbrock, and James A. Landay. 2011. Voice Games: Investigation Into the Use of Non-speech Voice Input for Making Computer Games More Accessible. In *Human-Computer Interaction – INTERACT 2011*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 11–29.

[9] Ramin Hedeshy, Chandan Kumar, Raphael Menges, and Steffen Staab. 2021. Hummer: Text Entry by Gaze and Hum. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 741, 11 pages. https://doi.org/10.1145/3411764.3445501

[10] Takeo Igarashi and John F. Hughes. 2001. Voice as Sound: Using Non-Verbal Voice Input for Interactive Control. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (Orlando, Florida) *(UIST*

[2] https://www.bmbf.de/

'01). Association for Computing Machinery, New York, NY, USA, 155–156. https://doi.org/10.1145/502348.502372

[11] Wijnand A IJsselsteijn, Yvonne AW de Kort, and Karolien Poels. 2013. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven* (2013), 3–9.

[12] Poika Isokoski, Markus Joos, Oleg Spakov, and Benoît Martin. 2009. Gaze controlled games. *Universal Access in the Information Society* 8, 4 (2009), 323–337.

[13] Poika Isokoski and Benot Martin. 2006. Eye tracker input in first person shooter games. In *Proceedings of the 2nd Conference on Communication by Gaze Interaction: Communication by Gaze Interaction-COGAIN 2006: Gazing into the Future.*

[14] Robert J. K. Jacob. 1990. What You Look at is What You Get: Eye Movement-Based Interaction Techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[15] Erika Jönsson. 2005. If looks could kill–an evaluation of eye tracking in computer games. *Unpublished Master's Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden* (2005).

[16] Yasmin B Kafai. 2005. The classroom as" living laboratory": Design-based research for understanding, comparing, and evaluating learning science through design. *Educational Technology* (2005), 28–34.

[17] Kechapp. 2020. *Scream Go Hero*. http://www.ketchappgames.com Last visited: 17.02.2022.

[18] Kristian Kiili. 2007. Foundation for problem-based gaming. *British journal of educational technology* 38, 3 (2007), 394–404.

[19] Suk-Jun Kim. 2018. *Humming*. Bloomsbury Publishing USA.

[20] Eric Klopfer, Scot Osterweil, Katie Salen, et al. 2009. Moving learning games forward. *Cambridge, MA: The Education Arcade* (2009).

[21] Tian Lan, Sandesh Aryal, Beena Ahmed, Kirrie Ballard, and Ricardo Gutierrez-Osuna. 2014. Flappy Voice: An Interactive Game for Childhood Apraxia of Speech Therapy. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play* (Toronto, Ontario, Canada) *(CHI PLAY '14)*. Association for Computing Machinery, New York, NY, USA, 429–430. https://doi.org/10.1145/2658537.2661305

[22] Michael Lankes, Matej Rajtár, Oleg Denisov, and Bernhard Maurer. 2018. Socialeyes: social gaze in collaborative 3D games. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*. 1–10.

[23] J Leyba and J Malcolm. 2004. Eye tracking as an aiming device in a computer game. *Course work (CPSC 412/612 Eye Tracking Methodology and Applications by A. Duchowski), Clemson University* 14 (2004).

[24] Marta Lopes, João Magalhães, and Sofia Cavaco. 2016. A Voice-Controlled Serious Game for the Sustained Vowel Exercise. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology* (Osaka, Japan) *(ACE '16)*. Association for Computing Machinery, New York, NY, USA, Article 32, 6 pages.

[25] Roland Maas, Ariya Rastrow, Chengyuan Ma, Guitang Lan, Kyle Goehner, Gautam Tiwari, Shaun Joseph, and Björn Hoffmeister. 2018. Combining Acoustic Embeddings and Decoding Features for End-of-Utterance Detection in Real-Time Far-Field Speech Recognition Systems. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5544–5548. https://doi.org/10.1109/ICASSP.2018.8461478

[26] Päivi Majaranta. 2012. Communication and text entry by gaze. In *Gaze interaction and applications of eye tracking: Advances in assistive technologies*. IGI Global, 63–77.

[27] A.A. Navarro-Newball, D. Loaiza, C. Oviedo, A. Castillo, A. Portilla, D. Linares, and G. Álvarez. 2014. Talking to Teo: Video game supported speech therapy. *Entertainment Computing* 5, 4 (2014), 401–412. https://doi.org/10.1016/j.entcom.2014.10.005

[28] Anders Møller Nielsen, Anders Lerchedahl Petersen, and John Paulin Hansen. 2012. Gaming with Gaze and Losing with a Smile. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 365–368.

[29] J O'Donovan, J Ward, S Hodgins, and V Sundstedt. 2009. Rabbit run: Gaze and voice based game interaction. In *Eurographics Ireland Workshop, December.*

[30] Lawrence Rabiner. 1977. On the use of autocorrelation analysis for pitch detection. *IEEE transactions on acoustics, speech, and signal processing* 25, 1 (1977), 24–33.

[31] Najmeh Sadoughi, André Pereira, Rishub Jain, Lolanda Leite, and Jill Fain Lehman. 2017. Creating prosodic synchrony for a robot co-player in a speech-controlled game for children. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI.* 91–99.

[32] Valerie J. Shute and Fengfeng Ke. 2012. *Games, Learning, and Assessment*. Springer New York, New York, NY, 43–58. https://doi.org/10.1007/978-1-4614-3546-4_4

[33] J David Smith and TC Nicholas Graham. 2006. Use of eye movements for video game control. In *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*. 20–es.

[34] Adam J Sporka, Sri H Kurniawan, Murni Mahmud, and Pavel Slavík. 2006. Non-speech input and speech recognition for real-time control of computer games. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 213–220.

[35] Melisa Stevanovic. 2013. Managing participation in interaction: the case of humming. *Text & Talk* 33, 1 (2013), 113–137. https://doi.org/doi:10.1515/text-2013-0006

[36] Alexander Streicher, Sebastian Leidig, and Wolfgang Roller. 2018. Eye-tracking for user attention evaluation in adaptive serious games. In *European Conference on Technology Enhanced Learning*. Springer, 583–586.

[37] Cagkan Uludagli and Cengiz Acarturk. 2018. User interaction in hands-free gaming: a comparative study of gaze-voice and touchscreen interface control. *Turkish Journal of Electrical Engineering & Computer Sciences* 26, 4 (2018), 1967–1976.

[38] Jan van der Kamp and Veronica Sundstedt. 2011. Gaze and Voice Controlled Drawing. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications* (Karlskrona, Sweden) *(NGCA '11)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages. https://doi.org/10.1145/1983302.1983311

[39] Marco Filipe Ganança Vieira, Hao Fu, Chong Hu, Nayoung Kim, and Sudhanshu Aggarwal. 2014. PowerFall: A Voice-Controlled Collaborative Game. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play* (Toronto, Ontario, Canada) *(CHI PLAY '14)*. Association for Computing Machinery, New York, NY, USA, 395–398. https://doi.org/10.1145/2658537.2662993

[40] Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics Cognition* 14 (08 2006), 129–182. https://doi.org/10.1075/pc.14.1.08war

[41] Colin Ware and Harutune H. Mikaelian. 1986. An Evaluation of an Eye Tracker as a Device for Computer Input2. In *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface* (Toronto, Ontario, Canada) *(CHI '87)*. Association for Computing Machinery, New York, NY, USA, 183–188. https://doi.org/10.1145/29933.275627

[42] Tom Wilcox, Mike Evans, Chris Pearce, Nick Pollard, and Veronica Sundstedt. 2008. Gaze and voice based game interaction: the revenge of the killer penguins. *SIGGRAPH Posters* (2008).

[43] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in Speech User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 369–376. https://doi.org/10.1145/223904.223952

[44] Oleg Špakov. 2005. EyeChess: the tutoring game with visual attentive interface. (01 2005).

A QUESTIONNAIRE

The questionnaire consists of three parts: the participant information, the experience, and the control. The experience section is about enjoyment, challenge, and engagement. The control part focuses on evaluating the input modalities. The questions in control section are answered separately for *MK* and *GV* methods.

A.1 Participant information

- How old are you?
- What is your gender?
- What is your current occupation?
- Do you have experience with video games (mouse and keyboard)? *1 (Never played) - 5 (I play very often)*
- Do you have experience with Eye Tracking? *1 (Not at all) - 5 (highly experienced)*
- Do you have experience with voice-controlled games? *1 (Not at all) - 5 (highly experienced)*

A.2 Experience

- How well do you think you played? *1 (Not good at all) - 5 (Very good)*
- How much did you enjoy playing? *1 (Very boring) - 5 (Very entertaining)*
- Was it challenging to play? *1 (Not at all) - 5 (Very challenging)*
- Was it exciting to play? *1 (Not at all) - 5 (Very exciting)*
- Did you focus deeply on the game? *1 (Not at all) - 5 (Very concentrated)*

- Have you lost track of time? *1 (Very slowly) - 5 (Very quickly*

A.3 Control

- Did you run out of breath? *(Not at all, sometimes, often)*
- Did the penguin move exactly the way you wanted it to? *1 (Very inaccurate) - 5 (Very accurate)*
- How would you rate Mouse/Eye Tracking overall? *1 (Very bad) - 5 (Very good)*
- How would you rate Keyboard/Voice overall? *1 (Very bad) - 5 (Very good)*
- Was it exhausting to play? *1 (Not at all) - 5 (Very exhausting)*
- How fast did you play? *1 (Slow) - 5 (Fast)*