OverVoice: Voice Controlled Games Engagement Study

Chaitanya Mattey
University of San Francisco
San Francisco, CA 94117, USA
cmattey@dons.usfca.edu

Prof., dr. Beste Yuksel
University of San Francisco
San Francisco, CA 94117, USA
byuksel@usfca.edu

Abstract
Player immersion is an extremely sought-after goal in video game design. Using voice as a supplement to conventional game input can add that extra layer of immersion. Adding voice inputs can also make games accessible to visually/physically impaired people. Of course, much would depend on the nature of the game, and the expected surrounding environment of the player. Voice is an important part of our daily interactions and binding certain actions in a video game to voice input would be an interesting study. As a proof-of-concept a system where participants played a flash game controlled with keyboard and voice inputs separately was piloted, their facial expressions were recorded, and the results are shown in this paper.

With the dawn of every new phase of technology, from the advent of Machine Learning agents to the rise of Virtual Reality, Video Games have proven to be the ideal testing ground for these technologies. Players seek an immersive experience and Voice Input in addition to conventional inputs can add that extra layer of immersion. In this study we

Author Keywords
voice control; human computer interaction; machine learning; facial expression; game design

Introduction
Video game design is an extremely involved task. The designers need to work on the plot of the games, the characters design and development, game worlds, non-player characters, deciding game genres and many more aspects have to coherently come together in order for a successful game launch.

Interaction with the game is one of the most important aspects, and developers have to change the game world based on the players' actions. The ways in which the player can interact with the game have been stagnant in the last few decades. The most common ways have been using a keyboard and a mouse. With the advent of Virtual Reality, VR headsets have become another way of interacting with the game world, the input in this case

Figure 1: A still from the game Lifeline, during the initial stage where the game asks you to pronounce certain words, and confirms if it's able to understand you or not, before jumping into gameplay

.

being the spatial orientation of the player. Some research has also tried playing games by processing written natural language reinforcement learning [1].

What if we could add another layer of interaction with the game, using voice recognition technology. Doing so would make the game more accessible and at the same time add another layer of immersion to the game. This is a challenging problem because developers would have to accommodate a variety of responses from the players and respond with appropriate actions. In addition to the plethora of inputs, designers would also have to account for the variations in the accents, dialects for players in different cultures.

Speech recognition technology is witnessing a tremendous boom. With digital assistants like Siri, Cortana and Alexa being used more often every day, it is safe to say that voice commands are here to stay. Some modern-day algorithms are even known to almost beat the Turing Test such as the newly announced Google Duplex, which successfully booked a Salon appointment for its user, by calling and conversing with the Salon owner. With this level of sophistication already in existence, it is about time that voice starts playing a bigger role in video games.

In our paper we elaborate on our path to conduct a preliminary study of the impact of using voice commands on player engagement levels. And later we discuss certain findings, strengths and drawbacks of our approach.

Related Work

Voice controlled games have slowly been on the rise. One of the oldest game, now regarded as a cult classic is 'Lifeline' released in 2003 by SCEI and Konami [1]. The defining feature of the game was that it was entirely controlled using the players voice through a microphone. The main character Rio was capable of understanding almost 500 verbal commands. The list of commands included 'stop', 'hurry', 'dodge', and 'smoke a Cigar'. It asked the player to pronounce certain words before gameplay, in order to make sure, words were understandable as shown in Figure 1. In spite of the revolutionary gameplay, the game met with average reviews, partly because of the often buggy interpretation of the voice commands, as seen in [2]. Not to mention that the game was released way back in 2003 and speech recognition has come a long way since then.

More recently there has been an advent of games with some form of speech input on mobile devices such as 'Chicken Scream' [3]. Most of these games take the voice intensity into account and don't do much in the way of actually understanding the content and context of the uttered sounds as a means of immersive gameplay.

Method

The study comprised of 3 systems working with each other: A free flash game hosted on a website [4], a facial recognition system to measure player emotions as they play the game, and the voice recognition system that translated user voice input to game input. The facial recognition system was changed after the pilot studies as discussed later. The control condition test is carried out first followed by the experimental condition. We briefly describe the face recognition system and the voice recognition system used during the study followed by the actual study phases.

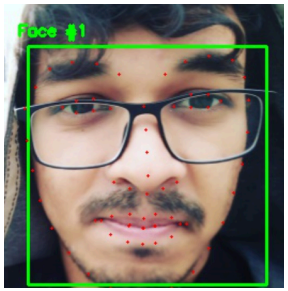

Figure 3: OpenCV's dlib library based on [5] can be used to detect landmark points on a face which can be used to identify facial expressions

.

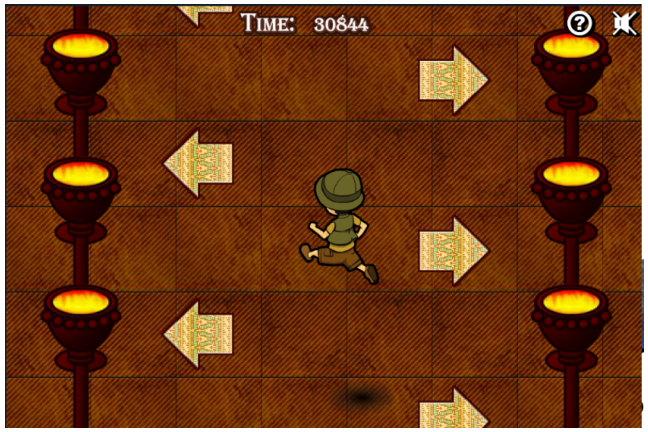

Figure 2: A still from the Jumping Arrows game used during the study

We employed the Affdex SDK provided by Affectiva in a Unity environment to carry out the facial recognition. The characteristics monitored included joy, sadness, engagement, and disgust. This method was preceded by an OpenCV model for facial emotion detection built in-house by calculating changes in location of facial landmarks in real-time(Figure 3), but was later substituted with the Unity method, since Affdex gives separate float values for the levels of different characteristics. Obtaining continuous engagement levels would have required extensive experimenting of different expressions which is left for a different exercise. During the facial emotion recognition using Unity, the monitored expressions were stored in a csv file for further analyses.

The voice recognition model was built using an Artificial Neural Network. The Sequential model was built on Keras with a TensorFlow backend and consisted of a single hidden layer. The model is trained on around 60 audio clips of 'Left', 'Right' and 'Quiet' input. The final input to the model is the user voice commands either asking the character to move 'Left' or 'Right', that are classified by the model, which then triggers keypresses in order to play the game.

The study was conducted in two consecutive phases. The first phase served as the control condition wherein the players played the game for 1 minute using conventional keyboard input, while their facial expressions are monitored and written to a file. For the second phase, the players are made to play the same game for 1 minute, but this time using voice commands, and their facial expressions recorded similar to phase 1.

Results

The study was conducted on a total of 5 participants. Analysis was conducted on the recorded csv files using pandas python library in a Jupyter notebook. The various expressions recorded are plotted as a continuous time graph. One thing to note is that, in some cases the time recorded for voice input are longer than 1 minute as participants took longer to get accustomed to the voice controls.

From the outset as seen in Figure 4, it can be observed that playing the game with voice controls provided a much more active gaming experience. The disgust levels were omitted from the graphs as they were too often misleading and coincided with high joy levels which was highly counter-intuitive.

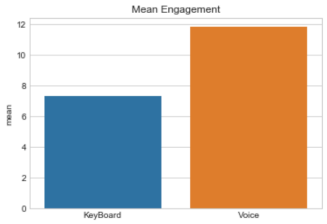

Figure 5: The Mean engagement values for all the participants was significantly higher while using Voice commands, p=2.005e-14 (Paired T-test)

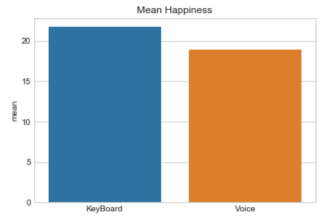

Figure 6: The Mean happiness values were higher in the case of playing with Keyboard, p=1e-3 (Paired T-test)

.

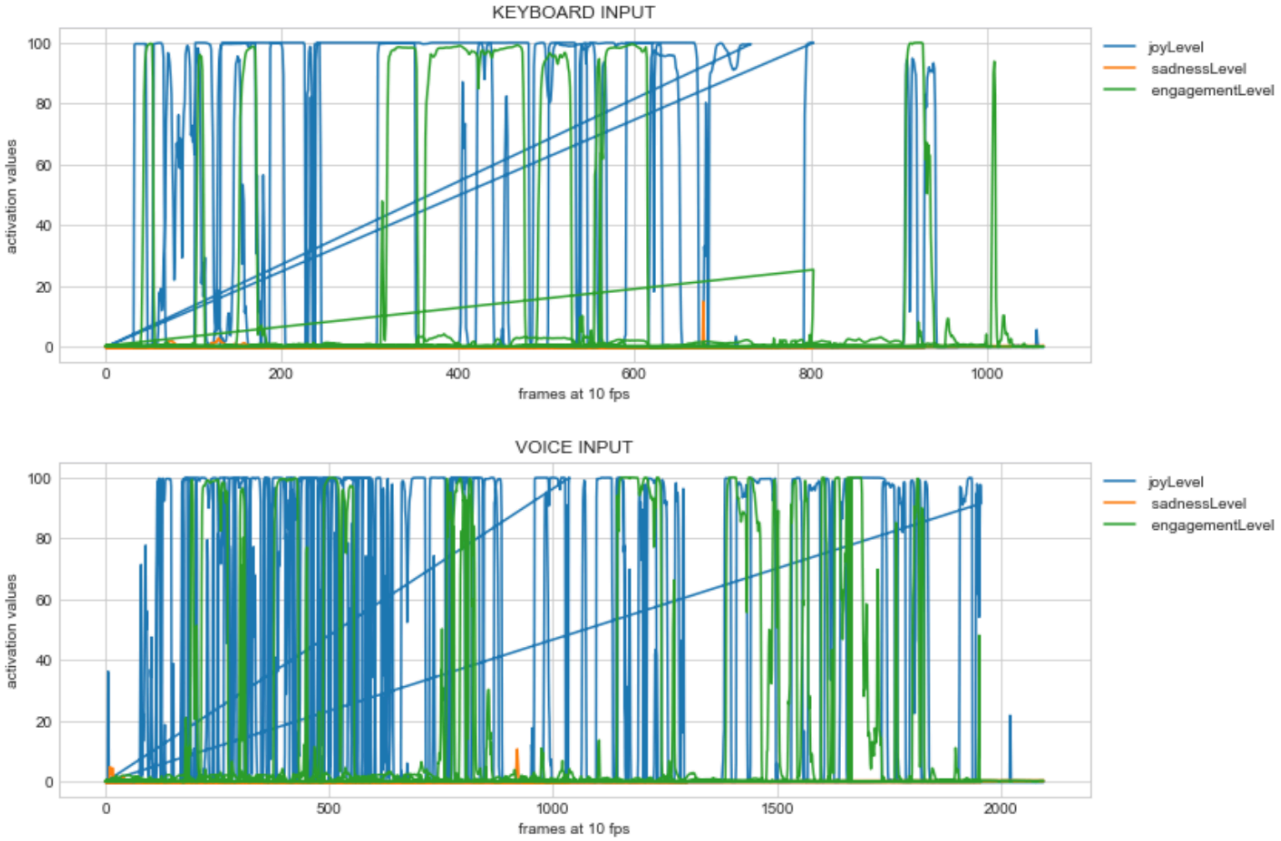

Figure 4: Continuous Time Series graph for the two types of input, Keyboard and Voice, for all participants combined

Discussion

We can observe from the results that significantly high levels of player engagement exist while using voice commands as compared to using the keyboard, this is in line with our assumption that using voice commands should add another layer of sophistication and immersion for the players leading to a more engaging gaming experience.

It is interesting to note that happiness levels show a decrease moving from keyboard to voice inputs, Figure 6. This can be credited to becoming accustomed to the

game, and it no longer holds a sense of surprise as the player has just played the game during phase 1.

Conclusion

The creation of games with supplementary voice controls is certainly a budding and potentially lucrative business idea for the video game industry. The study shows that there is certainly an interesting and engaging aspect to having a voice-controlled game. As we move forward in this area, we can expect to speech technology from big companies such as Google, Apple and Amazon to flow into this sector as well, leading to increased research and surge in vocal aspects of playing video games.

References

1. Kaplan, Russell, Christopher Sauer, and Alexander Sosa. "Beating atari with natural language guided reinforcement learning." *arXiv preprint arXiv:1704.05539* (2017).

2. Lifeline video game Wiki page:

   https://en.wikipedia.org/wiki/Lifeline_(video_game)

3. https://clips.twitch.tv/BoldCarefulKimchiYee

4. Chicken Scream, Google Play page:

   https://play.google.com/store/apps/details?id=com.perfecttapgames.chickenscream&hl=en_US

5. http://www.coolmath-games.com/0-jumpingarrows

6. Kazemi, Vahid, and Sullivan Josephine. "One millisecond face alignment with an ensemble of regression trees." 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014. IEEE Computer Society, 2014