Exploring Option-Selection in Voice-Driven Interfaces

Team Name: Human Computer Destruction

Nicholas Goh
University of Toronto
Toronto, Canada

Richard Ngo
University of Toronto
Toronto, Canada

Stuart Simpson
University of Toronto
Toronto, Canada

Vera Sipicki
University of Toronto
Toronto, Canada

ABSTRACT

In recent years, voice interaction systems have grown popular due to improvements in the accuracy of speech recognition. In an environment with multiple sound streams, a common error is the complete failure to detect the initial activation phrase and therefore fail to detect voice commands. In this paper, we propose a head gesture interaction system for selecting menu items that combines a Bluetooth connected wearable with a voice user interface (VUI). We believe that this head gesture system is an intuitive, easy to use alternative to VUIs. The experimental system detects four distinct, directional (up, down, left, right) head gestures and based on the direction, designates a menu item as selected. Two trials are reported: one with the gesture-based system, and the other with an existing VUI. Each trial was studied in an experiment conducted with 12 participants. Users found head motions to be intuitive and were interested in a production quality device. However, technical issues were raised concerning the sensitivity of the head placements and detection of movement.

Author Keywords

Voice User Interface, Natural Language Processing,
Gesture-based Interface.

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]

INTRODUCTION

With the sudden rise in availability and popularity of smartphones, the existence of digital assistants such as Apple's Siri and Google Assistant are becoming commonplace in our day to day lives. Often VUI interfaces falter either because of simple background noise or simply because understanding the subtle nuances of language in longer, more complex sentences, is currently a highly difficult task for a computer. Present limitations, in part, may currently place VUIs into more of a novelty category rather than one as an essential tool in our daily lives. These current usability limitations result in VUIs often switching from the voice based interface to an on screen menu based interface, as the tasks become more complex. In comparison to voice only interfaces, mode switching may not be as convenient or natural; it involves frequent shifts in the position of the user's cell phone and attention, to varying audio based and visual based modes of interaction.

Mode switching solves certain communication issues but it does not provide an interface that is strictly voice-only. Voice-only interfaces began with touch-tone telephone banking systems and are likely to prosper as home automation devices like Amazon Alexa grow in popularity. It is essential that we explore solutions for menu navigation that do not add additional display before these flaws permanently limit the design space of VUIs.

RELATED WORKS

Throughout the years, a lot of work research has been conducted with a focus on the relative usability of VUIs. In 2001, the use of touch-tone telephone interfaces began to grow rapidly as more and more financial institutions adopted the technology as a way of streamlining and optimizing interactions with call centers. Although it was increasing in popularity, there were no clear measure(s) of what benefit(s) the new interfaces brought since, some users expressed frustration while using them. In a study done by Suhm and Peterson, an experiment was conducted to evaluate commercial touch-tone and speech-enabled telephone VUIs using the single measure of task-completion time. The results found that for each call, an average of 9.8 seconds was saved if the customer was first subjected to a series of touch-tone VUIs to help expedite the call [7]. This work was later built upon in a follow-up study by Suhm et al. in which they evaluated the compared efficiency of using short touch-tone menus against longer, more detailed touch-tone menus [8].

In spite of the acknowledged utility of voice user interfaces, they were only primarily used as tool for large companies and banks. Consequently, user-friendliness of a VUI was never properly or fully researched until more recent years when consumer electronics manufacturers began adopting VUIs. Portet et al. completed a study in 2013 where they investigated usability of VUIs amongst the elderly, and found many useful conclusions when it comes to pain points and fears of using VUIs. One of the most interesting findings was that 95% of their participants said they would continue using a VUI even if it misinterpreted orders which contradicted a lot of what people believed at the time about it's usage barriers[5]. Instead, the biggest barriers to adoption were intuitiveness and ease of use, rather than interpretation error rates.

These studies were important for future researchers such as ourselves that will be investigating the usability and alternatives to VUIs as they create a framework by which we should be evaluating our own interfaces to make them comparable.

DESIGN CONCEPT

The purpose of our gesture-based interface is to add a more intuitive option for option-selection to voice-only interfaces without introducing an additional major component such as a display. The details of our apparatus can be found in the apparatus section, with a picture of the display mounted being found in **Figure 2**. This interface would be incorporated into an everyday calling experience to allow greater control when interacting with tools such as telephone banking systems. Ideally, support for this functionality would be built directly into the proprietary telephone application of modern smartphones, although it is possible for it to be adapted as an external tool that binds gestures to keypresses. Although it is possible to track diagonal directionality, we decided for this phase of the implementation to make a simple prototype, and as a result, the directional mappings do not support diagonals (see **Figure 1**).

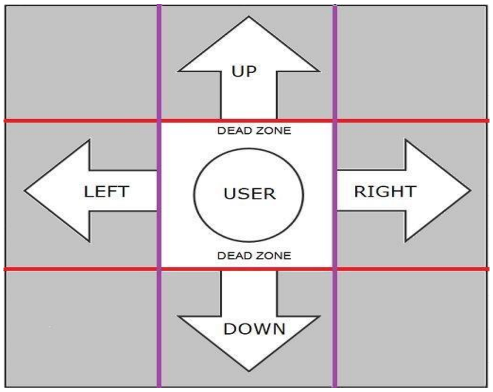

Figure 1: Depiction of the directional mappings used in the gesture-based interface. Left and right take priority over up and down in the corner boxes.

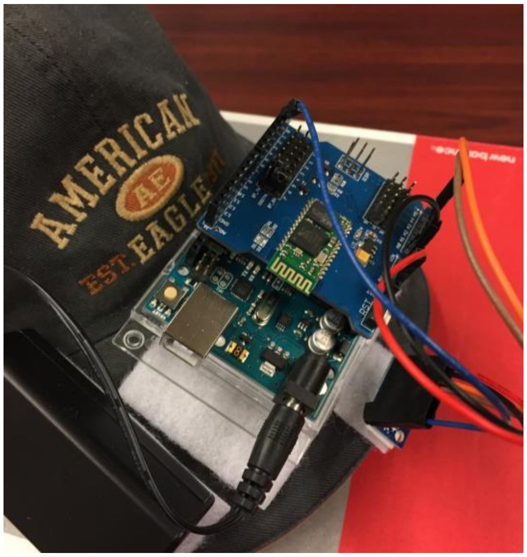

Figure 2 - Apparatus mounted to a hat

Arduino UNO (green), attached bluetooth shield (blue) and MPU 6050 Sensor (peak of hat) are all visible, as well as the nine volt battery pack (left).

METHODOLOGY

In this section, we discuss the details of our experiment, the apparatus required to conduct our experiment, and our participants. We will also discuss how we collected data for the study, including details about any surveys or questionnaires given.

Experimental Design

Our experiment was a 2x3 within-subjects design with our independent variable being the interface type that was used: (gesture-interface versus voice-interface). The two dependent variables that we were interested in were: task-completion time as well as subjective usability feedback.

PROCEDURE

After signing the consent forms, the first task participants were asked to complete was a pre-trial survey to help us better understand the background of our participants. This survey was immediately followed by a brief explanation of the two systems that participants would be interacting with. We then allowed them a period of time to ask clarifying questions, and to familiarize themselves with both of the interfaces before we began timing tasks.

Once the participants had become comfortable with our apparatus, each participant was asked to complete a series of tasks using both the gesture based interface and VUI that were described earlier. This meant that our experiment would be of a within-participants design. In the first set of tasks, participants were asked to input a set of directional actions, and the the time taken to complete each set of inputs was recorded. Using the gesture based interface, participants input the directions using the model described in **Figure 1**, while with the VUI, participants were asked to dictate the set of commands.

In the second series of tasks, participants were given similar goals, with the difference being that rather than directions being input, participants were instead inputting components of a food order. In the specific series of tests we conducted, left was instead replaced by “Add a Pizza”, up was replaced with “Add a side”, right was replaced with “Add a dessert” and down was replaced with “Add a drink”. This was again repeated using the VUI as participants were asked to place an identical order by vocalizing their orders.

If a participant was having trouble completing a task, he or she was allowed a temporary rest break as necessary, and was then encouraged to repeat the task. Errors were noted by the experimenter in a log, but were corrected by the participant, before continuing.

After testing was complete, participants were then provided with a second survey to collect information on participants’ perceived usability of the systems. This survey used Likert scales to identify both impressions and pain points of the system. Finally, participants were provided with a debriefing session.

Participants

Participants recruited for our study were selected using a convenience sampling of the University of Toronto campus. In total, our participant group consisted of 12 voluntary participants, of which 11 were male and 1 was female, and the age range for our participants was from 21 to 30 with a mean age of 26.4. With the exception of one participant, every other participant had experience with VUIs prior to the study, although a majority stated that they had not used one in the last 30 days.

Apparatus

The head gesture system uses an Arduino Uno interfaced serially with a MPU6050 six-axis gyroscope + accelerometer sensor. The Arduino itself also serially connects with a laptop computer via the standard bluetooth protocol and continually sends updated yaw, pitch and roll measurements in euler angles. The laptop computer’s software interprets the received raw data and converts it into one of five gestures based upon the categories of **figure 1**.

When the Python program on the laptop begins, it measures the current head alignment of the user, which then becomes the local dead zone. When the user then moves their head, the directions of up, down, left and right are measured against this baseline. The computer outputs synthesized speech letting the user know that they’ve completed this action, or have moved from a gesture back to the center. For our experiment, we considered a gesture to be a movement from the center box, (or “dead zone”) to one of the four directions. A “command” must begin at the center position. The user has to move from an active gesture “area” to the center before the next gesture can be performed.

For the VUI used as a comparison, two options were explored, the first being the Python SpeechRecognition library that would leverage online services such as Wit.ai and IBM Watson to synthesize text from a given audio clip. Initially, this approach was used and incorporated into a testing script. However as the pilot study was conducted, it was discovered that the campus network would not support an appropriate response time for our experiment, and instead we used an online dictation tool (http://www.dictation.io).

Measures

The two dependent variables that we examine in our study are task-completion time and subjective usability feedback gathered from our participants. Task-completion time is an important metric to monitor as it signifies first and foremost if our system will be a viable replacement for existing interfaces, and secondly it signifies how efficient users can be when interacting with our interface. Usability feedback is also important to note as it will help us better identify particular ‘pain’ points or redeeming factors of our interface that we might miss during our experimental observations.

Data Collection

To record task-completion time, a researcher operated a standard stopwatch. To collect usability feedback, a post-experiment survey was administered to each participant with many questions using a Likert scale to evaluate perceived usability, ease of use and more of the gesture-based interface. In addition, each experimenter was given identical observation sheets with which they could make notes on each participant’s experience and record timings.

RESULTS

In this section we will discuss the raw findings that were gathered from our study, and outline the statistical significance that we found along with methods used to reach these conclusions.

Results from the pre-experiment questionnaire have been broadly covered in the above participants section. It is important to note that of participants that had used VUIs before, Google Now and Siri were clearly the most commonly used interfaces.

It was observed that from the two trials, that average task-completion time between the gesture-based interface (~5.753(s),~5.677(s)) and the voice-based interface, (~3.421(s), ~3.353(s)) each differed by roughly two seconds each, with voice-based interfaces being the faster method of interaction. Between each trial, it was noticed that average task-completion time was minimally different with the second trial being slightly faster, although this could be attributed to increasing familiarity and comfort with each interface. This meant that our gesture-based interface took longer to complete the same tasks than the voice-based interface.

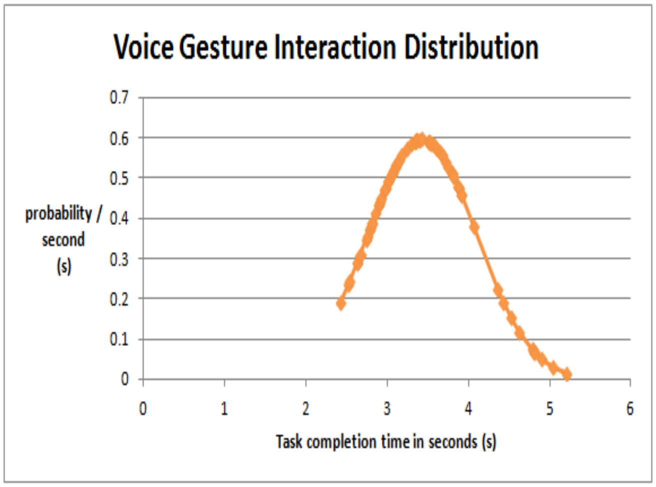

Figure 3: Distribution graph of the voice interaction system (avg ~ 3.43 seconds, std. deviation of ~0.67)

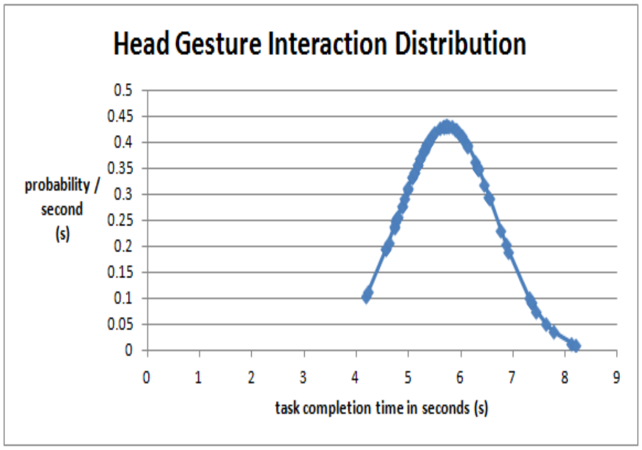

Figure 4: Distribution graph of the head gesture system (avg ~ 5.73 seconds, std. deviation of ~0.92)

A Wilcoxon signed-rank test was used to analyze the findings of the experiment since it was most suitable for analyzing a usability experiment of the within subjects design that consisted of only two experimental conditions.

As can be seen in Figure 3 and Figure 4, both the head gesture system and voice-based task-completion times closely follow a normal distribution. To investigate changes in the completion times from one condition to another, we will compare the two sets of times from the same participants. Our assumptions for using the Wilcoxon Rank Based Test are:

1. The dependent variable is continuous (interval) measured in seconds (s).
2. There are two categorical independent variables (head gesture and voice-based).
3. The distribution of differences between pairs is symmetrical in shape and approximately normal.

The signed-rank test was performed first on raw data from head gesture versus voice-based gesture (Z = - 7.374, p = 0.000) and then performed on the averages of each trial in each independent variable (Z = -2.201, p = 0.0277).

The null-hypothesis of this experiment is that there is no measurable time period difference between participants using the head gesture system and voice interface system, i.e. the median difference between pairs of observations is zero.

Assuming a significance level of 0.05, it is noted that (p < 0.05) in both Wilcoxon signed rank tests performed. As such, the p-values were denoted as not significant and the null hypothesis was rejected. Therefore the findings of the experiment were significant.

From our post-trial survey, we found that despite the apparent flaws in our interface, participants generally responded positively about the usability and accessibility of the interface . From our results, we found that a majority of our participants strongly agreed with the idea that the gesture-based system was intuitive (80.1%), easy to use (80.1%) and accessible. The levels of agreement for the head gesture-based interface were on par with the levels of agreement for similar traits (87.5% intuitive, 87.5% easy to use) when asked about the voice-interface. This leads us to believe that participants found the overall usability of our gesture-based interface to be similar, (or nearly but not quite as good) to that of the voice-based interface. From the feedback about pain points in the gesture-based interface, it was found that the most common cause of distress for participants was the sensitivity of the gesture-based interface which had been observed during the experiment as participants sometimes did not fully input a command when they felt that it should have.

DISCUSSION

Overall, from our results we found that while our current implementation is inferior to a VUI, participants were still responding positively to the experience and were optimistic for future iterations. These findings reflected the work done by Portet et al. as it showed that users were not put off by possible input errors or sensitivity issues and were still content as long as they were capable of understanding and using the interface. During debriefing with participants it was noted that several of the participants mentioned that the time difference between either interfaces felt minimal to them.

A possible explanation for some of the timing issues could likely fall within confounding factors in our experiment. One factor that may have been an issue in our experiment is that in our attempts to create a controlled environment, we created an environment ideal for VUI, skewing results and task-completion times towards the VUI. Because we used a controlled room where noise was kept to a minimum and provided participants with an external microphone for voice inputs, this created an unnatural and altogether unlikely situation to be using a VUI in. This means that audio inputs were likely processed faster and with greater accuracy than what might be expected in a real-world scenario.

Another confounding factor to consider was that of the primacy-recency effect[4]. In our experimental design, we did not accommodate for alternating the first and last interfaces that participants used with the first always being the gesture-based interface and the last always being the VUI. Because the tasks in each trial were identical for each interface, it may be possible that participants were still learning during the trials conducted with the gesture-based interface, and were more familiar with the trials when they completed tasks with the VUI. Finally, because we had a sample size of 12 participants all taken from the same location, it is hard for us to generalize these results beyond our initial participant group.

With regards to the usability feedback, we believe that our findings are very strong and valuable for future study. As a result of the survey asking participants to directly compare the usability of the gesture-based system with the VUI used, we can directly compare what participants perceived to be the difficulty differences between each interface.

CONCLUSION

Despite some of the shortcomings of our initial interface design as well as confounds in our experimental design, we believe the results we gathered will be significant for future endeavors. In this study, we have identified that there is a clear desire for additional interface options as alternatives to voice-based interfaces. We conclude that although our current implementation is not an interface that participants wish to be using in the future, participants would be interested in exploring an improved iteration of interfaces similar to our design.

ACKNOWLEDGMENTS

We would like to thank Professor Holman and Varun Perumal for their assistance in this project.

REFERENCES

[1] Asthana, S., Singh, P., and Singh, A. Exploring adverse effects of adaptive voice menu. CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13, (2013).

[2] Evreinova, T.V., Evreinov, G., and Raisamo, R. Integrating discrete events and continuous head movements for video-based interaction techniques. Behaviour & Information Technology 30, 6 (2011), 739–746.

[3] Jia, P., Hu, H.H., Lu, T., and Yuan, K. Head gesture recognition for hands-free control of an intelligent wheelchair. Industrial Robot: An International Journal 34, 1 (2007), 60–68.

[4] Morrison, M. (2015, March 13). Primacy and Recency Effects in Learning. Retrieved April 10, 2017, from https://rapidbi.com/primacy-and-recency-effects-in-learning/

[5] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. 2013. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Personal Ubiquitous Comput. 17, 1 (January 2013), 127-144. DOI=http://dx.doi.org.myaccess.library.utoronto.ca/10.1007/s00779-011-0470-5

[6] Reimer, B., Mehler, B., Dobres, J., et al. Effects of an 'Expert Mode' Voice Command System on Task Performance, Glance Behavior & Driver Physiology. Proceedings of the 6th

International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14, (2014).

[7] Bernhard Suhm, Barbara Freeman, and David Getty. 2001. Curing the menu blues in touch-tone voice interfaces. In CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01). ACM, New York, NY, USA, 131-132. DOI=http://dx.doi.org/10.1145/634067.634147

[8] Bernhard Suhm and Pat Peterson. 2001. Evaluating commercial touch-tone and speech-enabled telephone voice user interfaces using a single measure. In CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01). ACM, New York, NY, USA, 129-130. DOI=http://dx.doi.org/10.1145/634067.634146