The Vocal Joystick: Evaluation of Voiced-based Cursor Control Techniques

Susumu Harada, James A. Landay
DUB Group
University of Washington
Seattle, WA 98195 USA
{harada,landay}@cs.washington.edu

Jonathan Malkin, Xiao Li, Jeff A. Bilmes
Department of Electrical Engineering
University of Washington
Seattle, WA 98195 USA
{jsm,lixiao,bilmes}@ee.washington.edu

ABSTRACT

Mouse control has become a crucial aspect of many modern day computer interactions. This poses a challenge for individuals with motor impairments or those whose use of hands are restricted due to situational constraints. We present a system called the Vocal Joystick which allows the user to continuously control the mouse cursor by varying vocal parameters such as vowel quality, loudness and pitch. A survey of existing cursor control methods is presented to highlight the key characteristics of the Vocal Joystick. Evaluations were conducted to characterize expert performance capability of the Vocal Joystick, and to compare novice user performance and preference for the Vocal Joystick and two other existing speech based cursor control methods. Our results show that Fitts' law is a good predictor of the speed-accuracy tradeoff for the Vocal Joystick, and suggests that the optimal performance of the Vocal Joystick may be comparable to that of a conventional hand-operated joystick. Novice user evaluations show that the Vocal Joystick can be used by people without extensive training, and that it presents a viable alternative to existing speech-based cursor control methods.

Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User interfaces—*Voice I/O*; K.4.2 [**Computers and Society**]: Social Issues—*Assistive technologies for persons with disabilities*

General Terms

Human Factors, Measurement, Performance

Keywords

Voice-based interface, speech recognition, cursor control, Fitts' law, continuous input

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

1. INTRODUCTION

The mouse has been one of the most successful and pervasive computer input devices since the emergence of the graphical user interface in the 1960s. Along with the keyboard, it has shaped the manner in which people interact with computers. Today's common computer operating systems, such as Windows, Mac OS, and many variants of Linux, all provide WIMP (Windows, Icons, Menus and Pointing devices) style interfaces. The success of the mouse as an input device and the GUI as the predominant computer interface can be attributed primarily to the intuitive and simple mapping between the required manipulation of the device and the resulting effect presented directly on the graphical interface.

Despite the success of the mouse-centric computer interfaces of today, the underlying assumption of the availability of a pointing device in the system design process has kept such interfaces from being easily accessible by people for whom the use of a mouse is not an option. Some of the reasons that preclude mouse usage may include various types of motor impairments (e.g., arthritis, muscular dystrophy, spinal cord injuries, and amputation) as well as situational impairments (e.g., mobile environments and hands occupied for other tasks) [16].

In this paper, we will focus primarily on computer interaction that involves the use of some form of pointing device such as the mouse or a joystick, and how these interfaces can be made accessible to those without the use of their hands. Although it is true that a number of interactions that typically involve pointing and clicking with a mouse can be substituted by a sequence of keyboard input events, there still remain a substantial number of tasks that require (or benefit from) a pointing device. Examples of such tasks include selection of unnamed objects or arbitrary points on a screen, and continuous and dynamic path following for applications such as games and drawing. Even in cases when targets can be acquired through other means, such as tabbing through the set of targets, pointing-based selection may significantly enhance the speed and enjoyment of the interaction.

There currently exists a number of mouse alternative devices that provide pointing capabilities to individuals who cannot use a conventional mouse, such as eye trackers, head trackers, and mouth operated joysticks. However, many such devices are costly or have limited level of controllability in terms of speed and degrees of freedom afforded. Among such solutions, speech based systems offer promise due to their lower cost and the fact that they typically do not require any elaborate hardware setup.

We present a system called the Vocal Joystick as a potential solution to address the challenges faced by the current mouse alternative systems. The system can recognize both verbal and non-verbal vocalizations by the user along with other continuous vocal characteristics such as pitch and loudness, and maps them to interface control parameters such as mouse cursor movement. The key benefit of the system is that it offers immediate processing of continuous vocal input, meaning that the user's subtle variations in intent are processed instantaneously and continuously, and are reflected immediately in the interface without delay. Although the overarching goal of the Vocal Joystick project is to develop low-cost, easy to use and efficient computer input method for supporting multitudes of interface manipulations such as robotic arm control or power wheelchair operation, in this paper we focus on the voice-based cursor control capabilities of the system.

The questions we seek to answer in this paper are:

- Can cursor control based on continuous vocal parameters be modeled by Fitts' law of human motor performance (thereby allowing it to be placed on the map of other motor-controlled devices for comparison)?
- How does expert performance of the Vocal Joystick compare to the mouse?
- How does the Vocal Joystick compare to existing speech-based cursor control methods?
- Can people use the Vocal Joystick effectively?

We conducted two sets of evaluations to answer the above questions. The contributions of this paper can be summarized as follows:

- Characterization of voice-based cursor control using Fitts' law model
  As far as we know, this is the first work that validates and quantifies voice-based input under the theoretical framework provided by Fitts' model of human motor performance and compares it to other such devices.
- Presentation of a novel voice-based system capable of continuous and effective cursor control

In the following sections, we will first present an overview of the Vocal Joystick system. Next, we provide an overview of existing mouse input alternatives and other speech-based cursor control systems to highlight the benefits that the Vocal Joystick offers. We will then present the results from two evaluations of the Vocal Joystick system. The paper closes with an overview of other application areas in which we will be applying the Vocal Joystick system in the future.

2. THE VOCAL JOYSTICK SYSTEM

In this section we provide a brief overview of the Vocal Joystick system and highlight its features. A more detailed technical description of the architecture and recognition algorithms can be found in [1] and [13].

2.1 System overview

The Vocal Joystick system is written in C++ and runs on standard personal computers (currently supported on Windows and Linux operating systems) and requires no special hardware other than a microphone and a sound card, which is standard on most modern computers.

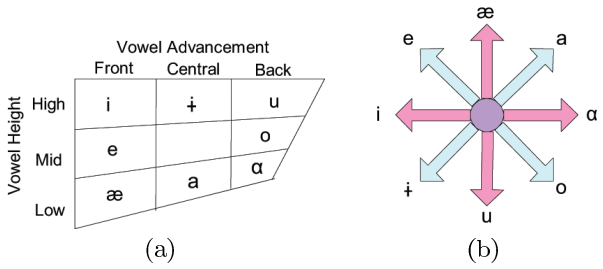

Figure 1: (a) Vowel sounds (shown using International Phonetic Alphabet symbols) as a function of their dominant articulatory configurations. (b) Mapping of vowel sound to direction in the 8-way mode Vocal Joystick. In 4-way mode, only the vowels along the horizontal and vertical axes are used.

The key feature of the Vocal Joystick is that unlike traditional automatic speech recognition (ASR) systems, which recognize sequence of discrete speech sounds, it processes continuous vocal characteristics every audio frame (10ms) and transforms them into various control parameters such as cursor movement. This results in a highly responsive interaction where a change in vocal characteristics initiated by the user is reflected immediately upon the interface.

The vocal characteristics that we currently extract include pitch, power, and vowel quality. The vowel quality is classified by a multi-layer perceptron adapted to each user which is then mapped directly to the 2-D vowel space (Figure 1(a)). Because this vowel space can be traversed continuously by smoothly modifying the articulatory configuration, it provides a natural mapping to continuous 2-D directions (Figure 1(b)).

In order to make the system more accessible to novice users who may not be familiar with all the vowel sounds, the Vocal Joystick can be operated in either a 4-way mode or an 8-way mode (as well as a continuous adaptive-filtering mode, which was not used for this paper). In 4-way mode, only the four vowels along the vertical and horizontal axes in Figure 1(b) are used. This increases the tolerance to slight deviations from the expected vowel sounds, albeit at the expense of sacrificing the number of directions that one can move in.

Aside from continuous vocal characteristics, the system can also recognize discrete sounds, currently short sounds starting with a consonant. This is used to perform actions such as a mouse click. Since very short sounds can be used (currently the consonant "k" is used to issue a mouse click), the system can respond to the command with minimal processing delay.

Although we will be focusing on the cursor control functionality of the Vocal Joystick for this paper, the core of the Vocal Joystick engine is designed as a linked library that outputs generic vocal pattern features which can then be used by another application to control any arbitrary parameter, such as operating a robotic arm or a powered wheelchair.

2.2 Usage model

Before the user first interacts with the Vocal Joystick system, they have the option of adapting the system to

their vocal characteristics. Although the default speaker-independent model works fine for some people, adaptation can greatly enhance the accuracy of the system. During adaptation, the user simply holds each of the vowel sounds for two seconds at their normal loudness.

Once adaptation is complete, the user can start controlling the cursor by making vowel sounds that correspond to the desired direction (Figure 1(b)). The speed of the cursor can be controlled by changing the loudness of the vowel; the softer the sound, the slower the cursor movement and vice versa.[1]

2.3 Key features

The Vocal Joystick system has several key distinguishing features that provide benefit to its users. First, recognition of the eight vowels is very robust and accurate compared to recognizing words under conventional speech recognition systems. Also, the instantaneous processing of every audio frame leads to much more immediate system response compared to systems that require a whole word or sequence of words to be recognized before an action is taken. In addition, because vocal characteristics such as vowel quality, volume and pitch can be changed by the user continuously, the system allows such continuous changes to be transfered directly onto the control parameters, resulting in smooth and responsive interaction. Finally, the only physical ability required of the user by the system is the ability to vocalize, and it requires minimal equipment at very low cost.

3. RELATED WORK

In the following two subsections we present an overview of the existing mouse alternative devices, and more specifically voice-based cursor control methods, and highlight the key characteristics of the Vocal Joystick that make it an attractive solution.

3.1 Mouse alternative devices

There is a variety of solutions available today for substituting conventional mouse input with another modality for individuals with limited motor abilities.

Mouth operated joysticks, such as the QuadJoy from SEMCO[2] and Integra Mouse from Tash Inc.[3], enable joystick functionality to be operated using the mouth. Clicking is typically performed through a sip and puff switch integrated into the joystick. The need to keep the joystick in the mouth as well as fatigue are cited as common issues with such devices. The Vocal Joystick on the other hand can be operated without any physical contact with the device.

Eye trackers provide an alternative that requires only the ability to move the eyes. Although these devices provide an attractive alternative for individuals with very limited mobility, it requires specialized hardware to be attached to the computer, and often involve extra steps in the interaction process to work around the "Midas touch" problem [10]. The Vocal Joystick provides are more natural interaction mode, as the cursor moves only while the user is vocalizing and stops as soon as the vocalization is stopped. The user can also shift their visual attention to any part of the interface while operating the Vocal Joystick.

Head operated devices, such as the HeadMaster Plus from Prentke Romich Company[4], require the user to wear a tracking object (usually an ultrasonic sensor or an infrared reflective sticker) and for there to be a sensing/transmitting device located near the screen. Clicking can be performed either through dwelling (fixating on a point for longer than a fixed duration) or through the use of an additional device such as the sip and puff switch. However, such devices (as well as the eye trackers and mouth operated joysticks) can cost orders of magnitudes more than the Vocal Joystick, which only requires a relatively inexpensive microphone.

Some software based techniques, such as the CrossScanner from RJ Cooper & Associates[5], can be operated with much simpler input devices (typically a single switch) and cost significantly less than the solutions listed above. Unfortunately, these techniques can be limited in the control they offer. For example, the CrossScanner requires the user to wait for a scanning line to pass over the desired target before issuing a click and does not support continuously changing the cursor position. The Vocal Joystick provides immediate control over the speed and direction of the cursor through continuous variation of volume and vowel quality.

3.2 Voice-based cursor control

A number of systems have been proposed, both in academic research as well as in commercial products, to enable control of the mouse cursor using speech input.

Igarashi et. al. proposed several techniques for using non-verbal voice features such as utterance duration, pitch, and discrete sound frequency to control the rate of change of some interface elements such as when scrolling a map [9]. For example, to scroll the map down, the user would say "move down" followed by a continuous production of any sound, during which the map would scroll as long as the sound was present. Scrolling speed changed as the user varied the pitch. This method is similar to the Vocal Joystick in that it utilizes the non-verbal aspect of the voice input (e.g., pitch and utterance duration), but it still requires the user to issue discrete verbal commands to initiate the cursor movement in a specific direction.

The Migratory Cursor technique [15] also combines discrete verbal command and non-verbal utterance for positioning the cursor. This technique augments the standard cursor with a row or column (depending on the movement direction) of ghost cursors that are numerically labeled. The user hones in on the desired target by using the numeric labels on the line of ghost cursors to specify the coarse-level coordinate (e.g., by uttering "move left, eight"), and uses non-verbal vocalization to precisely adjust the position (e.g., saying "ahhhh" during which the cursor continues to move left at a slow pace). This technique also lacks the ability to fluidly and continuously change the movement direction without having to issue discrete commands.

The *SUITEKeys* [14] interface provides cursor control based on a constant velocity cursor that is initiated by voice commands such as "move mouse down" and "move mouse two o'clock". The cursor continues to move (although it is not specified at what speed and whether it is constant) until the user says "stop" or issues a button press command such as "click left button". The system also provides a way to move

[1]Video demonstrations of the Vocal Joystick system are available at http://ssli.ee.washington.edu/vj/

[2]http://www.quadjoy.com/

[3]http://www.tashinc.com/

[4]http://www.prentrom.com/

[5]http://www.rjcooper.com/cross-scanner/

the cursor in some direction by a specified amount, or to set its position to a specified coordinate. As with the Migratory Cursor technique, this interface also suffers from the discontinuous nature of discrete command recognition and the inability to control the movement speed efficiently.

Building on the work of Kamel and Landay [11], Dai et al. presented and compared two different versions of the grid-based cursor control method [4]. In a grid-based system, the screen is overlaid with a 3×3 grid that is numbered "1" through "9". The user recursively drills down into each grid by saying the corresponding numbers until the desired target is below the center of the grid. The two systems compared differed in whether there was only one active cursor at the center of the middle grid, or nine active cursors at the center of each of the grids. Their results showed that the nine cursor version resulted in significantly faster task times for acquiring targets of various sizes and distances. Although the grid-based approach can be quite efficient in moving the cursor to a particular point on the screen, it does not allow the user to move the cursor continuously across the screen, as is necessary when performing tasks such as drawing.

One of the widely used commercial speech recognition packages, *Dragon NaturallySpeaking*®[6], offers several methods for speech-based cursor control. One of them is the MouseGrid™, which is essentially the one cursor version of the grid-based method presented in [4]. The other is based on a constant velocity cursor similar to *SUITEKeys*. For example, the user would say "move mouse up" and the cursor would start moving up at a fixed velocity (the default is roughly 4 pixels per second). The user can then issue commands to change the speed (e.g., "much faster"), direction (e.g., "left"), to stop the motion ("stop"), or to click the mouse button ("click"). There are three levels of commands for changing the cursor speed ("faster", "very fast", and "much faster", and corresponding ones for decelerating). The cursor movement is also jerky, updating its position roughly four times a second, thereby skipping over a number of pixels when the velocity is greater than the default.

The system that is most similar to the Vocal Joystick is Voice Mouse [5], which uses different vowel sounds associated with each cursor movement direction. In this system, the user utters a vowel sound corresponding to the desired one of the four directions (/a/ for up, /e/ for right, /i/ for down, and /o/ for left), and the cursor starts moving in that direction. Once the cursor starts moving, it is governed by "inertial motion" and the user does not have to continue vocalizing. The cursor speed initially starts out slow and gradually accelerates with time. The cursor is stopped by uttering the same vowel sound again, and click is performed by uttering a two-vowel command (/a-e/). One major difference between the Voice Mouse and the Vocal Joystick is that with Voice Mouse, the cursor does not start moving until a vowel sound has been produced for a minimum duration and recognized (which could take around four seconds from the start of the vocalization). The Vocal Joystick on the other hand starts the cursor movement as soon as the user starts producing the vowel sound. The inability to control the cursor speed and to change its direction without stopping it are also major limitations of the Voice Mouse system.

[6]http://www.nuance.com/naturallyspeaking/

4. FITTS' LAW STUDY

To better understand the properties of the Vocal Joystick as a device for target acquisition, we first present a study with expert users to determine whether the Vocal Joystick can be modeled by Fitts' law, a well adopted model of human motor performance for movement tasks.

Before we describe the motivation behind and the setup of the study, we will briefly overview the key underlying concepts of the Fitts' law model.

4.1 Fitts' law

Fitts' law [7] provides a mathematical formulation, based on Shannon's fundamental theorem of communication systems [17]. It models the human motor system as a channel with a certain bandwidth (measured in bits per second) that is used to transmit information in performing a movement task of a certain index of difficulty (measured in bits). The model assumes a fairly simple 1-D target acquisition task where the goal is to move from a starting position to a target at a distance A (referred to as the amplitude) whose size is W (referred to as the width) along the movement direction. The model relates the amplitude and width to the movement time (MT) in the following form:

$$MT = a + bID \tag{1}$$

where $ID = log_2(2A/W)$ is referred to as the index of difficulty of the task (measured in bits), and a and b are empirically determined regression coefficients that characterize the particular modality used. The inverse of the slope coefficient, $1/b$, is referred to as the index of performance (IP), the information capacity of the motor system involved (measured in bits per second). Higher IP (or lower slope coefficient b) indicates that the particular motor system is more efficient at the target acquisition task. We note that the model predicts that the movement time should not be affected if the target amplitude and width are scaled by an equal factor.

For our study, we chose to use a minor variation of the formulation as proposed by MacKenzie [12] that has been shown to be slightly more accurate and theoretically sound [3], where ID is defined as:

$$ID = \log_2(A/W + 1) \tag{2}$$

4.2 Motivation for the study

In our previous preliminary user study [1], we were able to get a general idea of the order of magnitude of the difference in performance between the mouse and the Vocal Joystick for web page and map navigation tasks with novice users. In this study, we set out to conduct a more thorough analysis of the performance characteristics of the Vocal Joystick in the context of target acquisition tasks. In particular, we were interested first in determining whether the Vocal Joystick used as a pointing device can be modeled by a Fitts' law performance model, and second in calculating the Index of Performance (IP) of the Vocal Joystick with respect to the mouse.

There were several reasons why we tested the Vocal Joystick against the Fitts' law model. First, if we determine that the Vocal Joystick can indeed be modeled by Fitts' law, it will allow us to predict performance times for various target acquisition related tasks. This is useful in determining whether the Vocal Joystick is usable in interacting with

a particular interface that may have specific minimum performance requirements. Second, the model will allow us to generalize results from future experiments with fewer conditions (given the same general population). Finally, and perhaps most interestingly, determining the relative ratio of the Vocal Joystick's index of performance to that of the mouse will allow the Vocal Joystick to be compared to various other input devices that have been analyzed with respect to the mouse (or any other well studied device) using Fitts' law experiments. According to MacKenzie [12], variations in the exact setup and procedure of each Fitts' law experiment makes cross-device comparison of the absolute values of the index of performance difficult. However, the relative ranking and the ratio of the performance indices within each study tend to be quite consistent across studies, which suggests that analyzing the performance index of a new input device with respect to the well studied mouse can provide a way to compare it relative to other devices that already have known relationships to the mouse.

4.3 Experiment

4.3.1 Participants

Since the goal of this study was to characterize optimal Vocal Joystick performance, we had four "expert" Vocal Joystick users participate. As the Vocal Joystick system has not yet been widely distributed to the public, all experts were part of the research team. None of the participants had any motor impairments. We believe that this population still provides us with a good estimate for the optimal performance. In the future, we plan to conduct a more thorough study where we train and study a larger number of users from a variety of population groups including those with various motor impairments.

Our definition of an "expert" user was based on the fact that all four participants were well accustomed to all of the vowel sounds used in the Vocal Joystick, were very familiar with the speed control response behavior of the system, and have used the system over an extended period of time (over a month).

4.3.2 Apparatus

The experiment was conducted on a Dell Latitude D600 laptop with 1.6 GHz Intel Pentium M processor and 512 MB of RAM running Windows XP. The LCD screen had a diagonal size of 14.1 inches and the display resolution was set to 1400×1050 pixels. The experiment was conducted in full screen mode, with the user's head situated about two feet from the screen. Input into the system was either an external mouse or a microphone. An external optical USB mouse was used for the mouse condition, and software acceleration for the mouse was turned off in the operating system. A Sennheiser Headset Microphone connected to an Andrea USB sound pod was used for the Vocal Joystick condition.

4.3.3 Design

The study followed a fully-crossed, 2×4×3×8 within-subjects factorial design with repeated measures. The factors and levels were as follows:

- Modality (M) {Vocal Joystick, mouse}
- Index of difficulty (ID in bits) {2, 3, 4, 5}

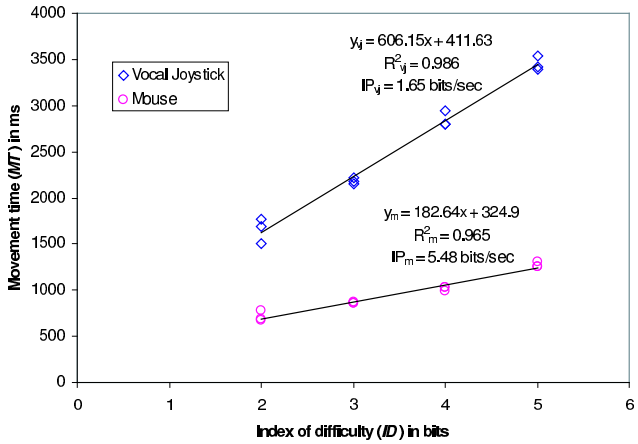

Figure 2: Linear regression of movement time (MT) versus index of difficulty (ID) for the Vocal Joystick and the mouse. Index of performance (IP)

- Target width (W in pixels) {12, 24, 32}
- Approach angle (θ in degrees) {0, 45, 90, 135, 180, 225, 270, 315}

As the target distance (A) is directly dependent on the index of difficulty and target width, it was not varied independently, and the following values were derived (A=36, 72, 84, 96, 168, 180, 224, 360, 372, 480, 744 and 992 pixels). Angle of 0 degrees corresponded to the direction pointing to the right of the screen, with the angle increasing in counter-clockwise order. For each of the 192 conditions, the participants performed 3 trials. The trials were grouped by modality, and the order of the modality was counterbalanced across participants. Within each modality, the order of the conditions were randomized.

4.3.4 Procedure

At the beginning of each trial, a target bar of width W (and infinite "height", as the bar continued beyond the edge of the screen) was presented on the screen at a distance of $A/2$ from the center of the screen. The cursor was positioned directly opposite of the target across the center of the screen, also at a distance of $A/2$ from the center of the screen. The trial was initiated and the timer started as soon as the cursor moved away from its original position. The participants were instructed to attempt to acquire the target as quickly and accurately as possible. When the cursor was moved above the target, the target bar changed color to indicate that the cursor was above the target. The participant acquired the target by pressing the space bar under both modalities (in the case of mouse, they were told to use the hand unoccupied by the mouse). This was done to normalize the difference in the clicking modality between the two devices, since we were primarily interested in the movement time to the target.

4.4 Results

In the rest of the paper, data we report as significant are at $p < .05$ level, unless otherwise specified.

To test for model fit against Fitts' law, linear regression analysis was performed on movement time against ID (Fig-

Table 1: Relative *IP*s across devices. Higher number represents more "efficient" device.

Device	Relative *IP*
Trackball [6]	1.11
Mouse	1.00
Touchpad [6]	0.88
Joystick [2] (isometric; velocity control)	0.43
Joystick [6] (displacement; velocity control)	0.42
Vocal Joystick	0.30

ure 2). For both modalities, *ID* significantly predicted movement times ($p < .001$) and also explained a significant portion of variance in movement times ($p < .001$). In the figure, within a particular modality and an index of difficulty, aggregate movement time for each of the three widths were plotted as separate points. The fact that the three points within each *ID* are close to each other reflects the fact that the data follows Fitts' law, as change in width given a fixed *ID* should not affect the movement time (see Equation 2).

The normalized *IP*s of the Vocal Joystick as well as several other devices that have been studied with respect to the mouse are shown in Table 1. Here, numbers less than 1 represent devices that are less efficient than the mouse, and those greater than 1 represent devices that are more efficient than the mouse. As can be seen, the Vocal Joystick is fairly close in terms of relative *IP* to the conventional velocity control joysticks.

Discussion of the results for the Fitts law study is presented in section 6 along with the discussion of the comparative evaluation presented next.

5. COMPARATIVE EVALUATION

The Fitts' law study provided us with a model of the Vocal Joystick's optimal performance characteristics as it pertains to target acquisition tasks. We next sought to investigate how novice users would perceive the Vocal Joystick as compared to existing speech-based cursor control methods, namely the two methods found in *Dragon NaturallySpeaking*®. One of them was the Mouse Grid (*MG*), where the screen was recursively subdivided into nine grids, and the other we refer to as the Speech Cursor (*SC*), the technique in which the cursor was controlled by saying "mouse move (direction)" to start the movement.

5.1 Experiment

5.1.1 Participants

We recruited nine participants (five males and four females) to participate in the informal comparative evaluation, ranging in age from 18 to 25. None of the participants had any motor or speech impairments, and all but two were native English speakers.

5.1.2 Aparatus

The equipment setup used was the same as that in the expert study. The version of *Dragon NaturallySpeaking*® used was version 8 Professional.

5.1.3 Design

For each of the three cursor control methods (SC, VJ, MG), the participant was exposed to a random sequence

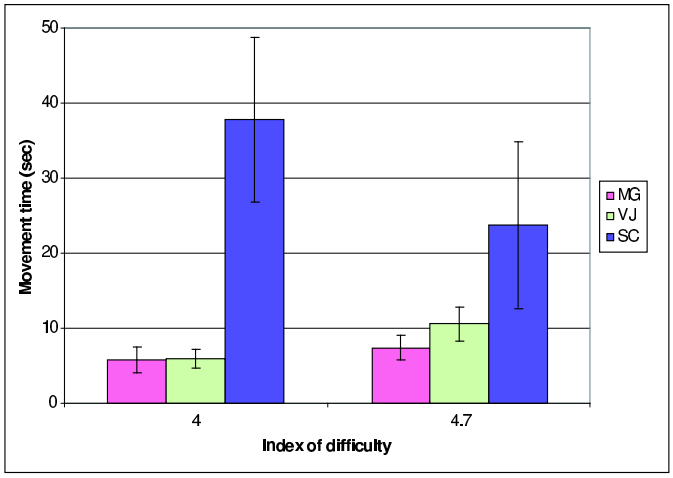

Figure 3: Movement times for Mouse Grid (MG), Vocal Joystick (VJ) and Speech Cursor (SC) for novice users.

of 16 trials, comprised of 2 target sizes (26 pixels and 52 pixels) × 8 directions, and was asked to acquire the target as quickly and accurately as possible. The participants used the four-way version of the Vocal Joystick to accommodate for the limited amount of time available to train the required vowels.

Due to the fact that the Vocal Joystick and Speech Cursor used for these conditions only allow movement along the four cardinal directions, distance to diagonal targets were measured using Manhattan distance and thus resulted in index of difficulty of 4 for non-diagonal targets and 4.7 for diagonal targets. As the emphasis was less on exhaustively covering all possible target conditions but rather on evaluating performance and preference under scenarios more closely resembling actual usage, we chose the target sizes that roughly corresponds to the size of common interface elements such as buttons and links.

5.1.4 Procedure

Each participant was introduced to the three systems in counterbalanced order, and for each system, they first went through a training session to adapt the system to their voice and then were given five minutes to practice the particular cursor control method. After the training session, they were asked to perform a series of target acquisition tasks similar to those used in the expert user study but with much fewer conditions and circular targets.

At the end of each set of trials for a particular control method, the participants were asked to fill out a questionnaire and rate the method on a 7 point Likert scale based on the 10 categories shown in Figure 4.

After all three methods were completed, we had the participants perform a simple path following task using Speech Cursor and the Vocal Joystick in counterbalanced order, where they were asked to trace a circle with a diameter of 600 pixels as quickly and as accurately as possible.

5.2 Results

Figure 3 shows the comparison of mean target acquisition time for each of the three modalities. There was a signif-

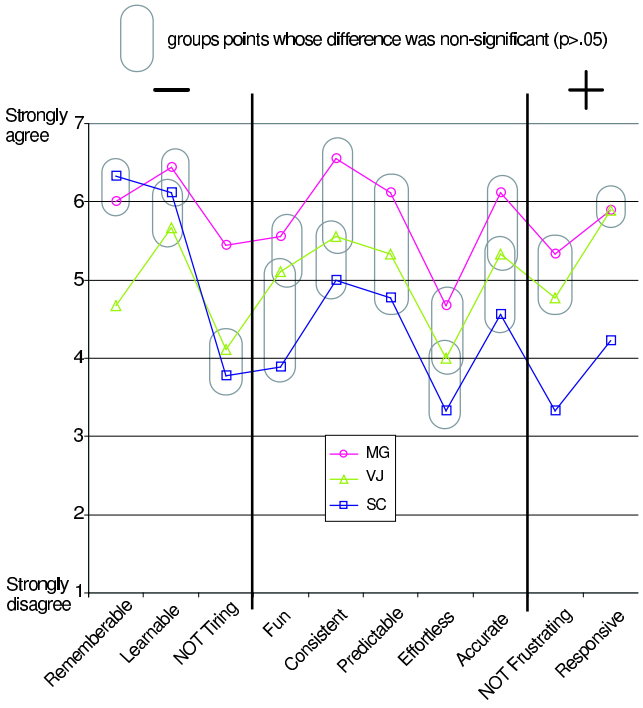

Figure 4: Novice users' subjective rating of the three cursor control methods: Mouse Grid (MG), Vocal Joystick (VJ) and Speech Cursor (SC). Items in the region labeled "+" are those in which VJ had a significantly better rating than another technique, and vice versa for the "-" region. VJ did not have any significantly different rating for the items in the unlabeled region.

icant difference between the Speech Cursor and the other two modalities for both indices of difficulties. There was no significant difference between the Vocal Joystick and the Mouse Grid for either of the *ID*s.

Figure 4 shows the aggregated user rating of the three control methods for each of the 10 dimensions. The ovals group points within each dimension whose difference was not significant. Therefore, any two points within a category that do not lie within the same oval are significantly different. The figure shows that the Mouse Grid was rated most favorably on most categories, but also that the Vocal Joystick ratings were not significantly different from it.

It is interesting to note that despite the Speech Cursor being rated higher than the Vocal Joystick in terms of how rememberable the control method was, it was rated to be significantly more frustrating to use than the Vocal Joystick. Based on the observation of the participants and the qualitative feedback we gathered during the post-session informal interview, participants indicated that the default Speech Cursor speed setting of 2 was too slow and desired a greater dynamic range on the speed change commands. Also, five of the nine users had substantial difficulty getting the system to recognize some of the direction words (about five unrecognized commands per trial). This may be due to the extremely short acoustic training period provided (three minutes) for the Mouse Grid and Speech Cursor. This points to a strength of the Vocal Joystick, which was able to perform

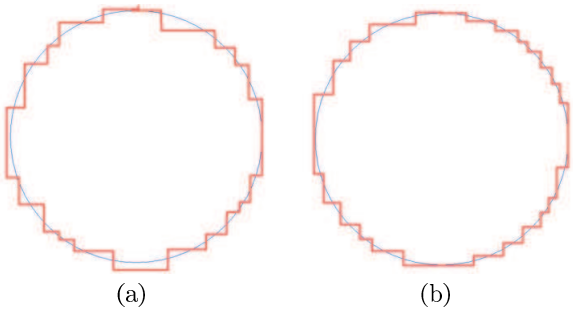

Figure 5: One of the participants' result from the tracing task using (a) the Vocal Joystick and (b) the Speech Cursor. It took the participant 45 seconds to complete (a) and 150 seconds to complete (b).

significantly better (Figure 3) given even shorter acoustic training time (eight seconds) total.

Figure 5 shows a result of the path following task from one of the participants. Figure 5(a) was traced using the Vocal Joystick, and Figure 5(b) was done using the Speech Cursor. Although the accuracy of the tracing looks quite comparable, the trace using the Vocal Joystick took only 45 seconds, whereas the Speech Cursor trace took 150 seconds. The average time across those who completed the path following task was 49 seconds for the Vocal Joystick and 155 seconds for Speech Cursor.

6. DISCUSSION

The Fitts' law study showed that the expert index of performance of the Vocal Joystick is roughly a third of that of a conventional mouse, and that it is comparable to a standard velocity control joystick. This is promising, as it indicates that as the user gains proficiency, they can potentially approach the performance of a hand-operated joystick. It remains to be seen what the learning curve looks like for the Vocal Joystick as a new user starts learning how to use it. Further investigation also needs to be conducted to see how much difference there will be with the population of individuals with various motor impairments.

The comparative evaluation revealed that novice users were indeed able to learn to use the Vocal Joystick given a short training period, and that they were able to achieve performance levels comparable to that of Mouse Grid. Mouse Grid is an attractive option from the point of view of its simplicity and reliability, however it cannot be used for non-discrete selection tasks such as path following. For such tasks, functionality offered by Speech Cursor is necessary, and given the significant difference in both target acquisition times and tracing time, the Vocal Joystick seems to be a viable alternative to these existing speech based cursor control methods.

7. CONCLUSION

We presented the Vocal Joystick system, a system that enables a user to continuously control the mouse cursor using their voice. We presented an overview of related systems aimed at providing mouse cursor control without the use of the hand, and discussed the benefits that the Vocal Joystick provides over these systems. Two user evaluations were pre-

sented to investigate expert performance of the system and to understand novice user performance and preference with respect to existing speech-based solutions. We were able to validate and determine the performance parameters that characterize the Vocal Joystick under the Fitts' model of human motor performance. Evaluations with novice users revealed that the Vocal Joystick can be effectively operated even with limited training, and that its performance beats or is comparable to existing speech-based cursor control methods.

We plan to continue our evaluation of the the Vocal Joystick system by recruiting more participants, in particular those with various motor impairments, to try out our system. We are also looking into studying the learning curve of the Vocal Joystick through a longitudinal study to determine the level of training necessary for people to achieve sufficient proficiency. Such a longitudinal study will also help reveal any potential issues with fatigue of the vocal cord after prolonged use. Through our evaluations so far we have not encountered any major complaints of vocal fatigue from our participants.

Also, as mentioned in [8], applications such as drawing and games play an important role in enriching the lives of people especially those whose range of activities may be limited due to some disability. We believe the Vocal Joystick is well suited for such applications, and will be exploring ways in which the system can be best used to support them.

8. ACKNOWLEDGEMENTS

We would like to thank all the participants who volunteered to take part in the user evaluation.

This material is based upon work supported by the National Science Foundation under grant IIS-0326382. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

9. REFERENCES

[1] J. A. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchhoff, A. Subramanya, S. Harada, J. A. Landay, P. Dowden, and H. Chizeck. The Vocal Joystick: A voice-based human-computer interface for individuals with motor impairments. In *Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing*, October 2005.

[2] S. K. Card, W. K. English, and B. J. Burr. Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys, for text selection on a crt. pages 386–392, 1987.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] L. Dai, R. Goldman, A. Sears, and J. Lozier. Speech-based cursor control: a study of grid-based solutions. *SIGACCESS Access. Comput.*, (77-78):94–101, 2004.

[5] C. de Mauro, M. Gori, M. Maggini, and E. Martinelli. Easy access to graphical interfaces by voice mouse. Technical report, Università di Siena, 2001. Available from the author at: maggini@dii.unisi.it.

[6] B. Epps. Comparison of six cursor control devices based on Fitts' law models. In *Proceedings of the 30th Annual Meeting of the Human Factors Society*, pages 327–331. Human Factors Society, 1986.

[7] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47:381–391, 1954.

[8] A. J. Hornof and A. Cavender. EyeDraw: enabling children with severe motor impairments to draw with their eyes. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 161–170, New York, NY, USA, 2005. ACM Press.

[9] T. Igarashi and J. F. Hughes. Voice as sound: using non-verbal voice input for interactive control. In *UIST '01: Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 155–156, New York, NY, USA, 2001. ACM Press.

[10] R. J. K. Jacob. What you look at is what you get: eye movement-based interaction techniques. In *CHI '90: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 11–18, New York, NY, USA, 1990. ACM Press.

[11] H. M. Kamel and J. A. Landay. Sketching images eyes-free: a grid-based dynamic drawing tool for the blind. In *Assets '02: Proceedings of the fifth international ACM conference on Assistive technologies*, pages 33–40, New York, NY, USA, 2002. ACM Press.

[12] I. S. Mackenzie. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction*, 7:91–139, 1992.

[13] J. Malkin, X. Li, and J. Bilmes. Energy and loudness for speed control in the Vocal Joystick. In *IEEE Automatic Speech Recognition and Understanding Workshop*, November 2005.

[14] B. Manaris, R. McCauley, and V. MacGyvers. An intelligent interface for keyboard and mouse control – providing full access to PC functionality via speech. In *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference*, pages 182–188. AAAI Press, 2001.

[15] Y. Mihara, E. Shibayama, and S. Takahashi. The migratory cursor: accurate speech-based cursor movement by moving multiple ghost cursors using non-verbal vocalizations. In *Assets '05: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pages 76–83, New York, NY, USA, 2005. ACM Press.

[16] A. Sears and M. Young. Physical disabilities and computing technologies: an analysis of impairments. pages 482–503, 2003.

[17] C. E. Shannon and W. Weaver. *The mathematical theory of communicaton*. Urbana University of Illinois Press, Urbana, Illinois, 1949.