Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor Control

Susumu Harada[1], Jacob O. Wobbrock[2], Jonathan Malkin[3], Jeff A. Bilmes[3], James A. Landay[1]

[1]Computer Science and Engineering
DUB Group
University of Washington
Seattle, WA 98195 USA

[2]The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA

[3]Electrical Engineering
College of Engineering
University of Washington
Seattle, WA 98195 USA

{harada, landay}@cs.washington.edu, wobbrock@u.washington.edu, {jsm, bilmes}@ee.washington.edu

ABSTRACT
We conducted a 2.5 week longitudinal study with five motor impaired (MI) and four non-impaired (NMI) participants, in which they learned to use the *Vocal Joystick*, a voice-based user interface control system. We found that the participants were able to learn the mapping between the vowel sounds and directions used by the Vocal Joystick, and showed marked improvement in their target acquisition performance. At the end of the ten session period, the NMI group reached the same level of performance as the previously measured "expert" Vocal Joystick performance, and the MI group was able to reach 70% of that level. Two of the MI participants were also able to approach the performance of their preferred device, a touchpad. We report on a number of issues that can inform the development of further enhancements in the realm of voice-driven computer control.

Author Keywords: Longitudinal study, speech recognition, voice-based interface, motor impairment, pointer control.

ACM Classification Keywords: 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*.

INTRODUCTION
In the United States, there are over a quarter of a million people with spinal cord injuries, 47% of whom are quadriplegic[1] (i.e., with significantly restricted use of their upper limbs and hands). According to the United Spinal Association, about 70% of the people with spinal cord injuries are unemployed.[2] For these individuals with limited mobility and motor control, access to a computer may be one of the few options available to them for achieving greater independence, obtaining or retaining employment, staying connected with people and information around them, and expressing themselves creatively [17]. These issues extend to people with other motor impairments as well, including the 46 million adults in the United States diagnosed with arthritis, the 1 million with Parkinson's disease, and the 50,000 children and adults with muscular dystrophy.[3]

Speech as a Primary Input Modality
Various assistive technology solutions have been developed over the years to make computers more accessible to users with disabilities. Among them, speech recognition holds great potential for users with motor impairments due to the hands-free interaction it affords without significant investment in specialized hardware. Speech recognition technology has been steadily improving, leading to accurate commercial dictation engines such as Nuance's *Dragon Naturally Speaking* software.[4] However, speech-based control of computers has not yet reached a point where it can provide the same level of access to application functionality afforded by the keyboard and mouse.

A key component that is missing in today's speech-based input technology is the analogue to direct manipulation that has made the mouse such a successful input device. While speech recognition systems excel at enabling spoken text entry and command-and-control-style interaction, they lack the facility to perform continuous fluid control such as the kind of pointing afforded by the mouse.

This ability to emulate direct manipulation using the human voice is essential for users with limited hand control, especially for those who depend on such capability to successfully operate modern computer applications for their employment and daily well-being. Such tasks may arise when using diagramming or drawing tools, selecting unnamed items or regions in a user interface, performing continuous browsing tasks such as panning, scrolling, and zooming, or even controlling various games and social applications that require fluid input, such as *Second Life*.

[1] http://www.sci-info-pages.com/facts.html
[2] http://unitedspinal.org/pdf/scd%20fact%20sheet.pdf
[3] http://www.hmc.psu.edu/healthinfo/
[4] http://www.nuance.com/naturallyspeaking/

To address these limitations, several research systems have recently been developed that use the non-speech parameters of human vocalization such as loudness, pitch, and vowel sounds for continuous control of the mouse cursor [4,8,13,15,19].

While promising, it is not yet clear whether these systems are indeed practical or usable. A key piece of knowledge we lack is how the target population of people with motor impairments can use these systems, and in particular, what issues they encounter in the process of learning them.

In this paper, we focus our attention on a promising option among the voice-based mouse emulators, the *Vocal Joystick* [4] (Harada et al. [9] present a comparison of various voice-based mouse emulator systems). We investigate how people with motor impairments develop their skills to control the mouse pointer as they learn to use the Vocal Joystick over multiple sessions spanning 2.5 weeks (see Figure 1 for an example of one of our participants learning to draw using her voice).

We present our findings from a longitudinal study we conducted to reveal the learning curve of the Vocal Joystick. We also analyze the space of voice-driven UI interaction, and discuss ways in which this space may be enhanced through the expressivity offered by the Vocal Joystick. There are certain challenges associated with evaluating such a novel input system, especially with the target population of users with disabilities over an extended period of time [6]. We present the lessons we learned and how they inform the design and enhancement of voice-based direct manipulation systems.

In the following sections, we will first explore in more detail the limitations of current speech-based computer control and the functionality offered by the Vocal Joystick. Next, we describe the stages involved in learning to use the Vocal Joystick and the design of our longitudinal investigation. Finally, we present the results and observations from the study and discuss lessons learned and propose directions for future research in this area.

SPEECH-BASED COMPUTER CONTROL

For speech to become a fully functional modality for operating the typical personal computer today, several key functions need to be supported:

1. Text entry – Ability to input textual information quickly and accurately.
2. Commands – Ability to execute all commands available on a system.
3. Direct manipulation – Ability to manipulate objects and perform mouse-like operations fluidly.

Much research has been poured into the first criteria in the advancement of automated speech recognition (ASR). Commercial products such as Dragon Naturally Speaking and the Windows Vista speech recognizer include command and control capabilities to address the second criteria, although there is still a large number of functionality that remains inaccessible through speech commands. The third criterion is still virtually unaddressed.

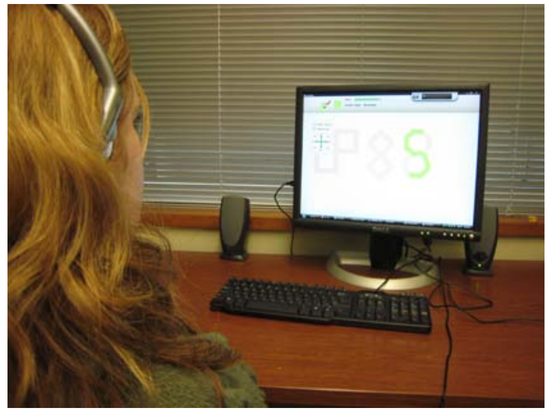


Figure 1: One of our participants with muscular dystrophy using her voice to draw with the *VoiceDraw* program [10].

This current state of speech-based input may be analogous to a user being given only a squishy keyboard and no mouse for providing input into a computer. The user may be able to enter text with relatively acceptable accuracy. If the user knows the various keyboard shortcuts, he may be able to access various menu items and issue commands, and to switch between applications. He may also be able to control the mouse pointer using the arrow keys to move the pointer around in the four cardinal directions, with constant or possibly incrementally variable speed. With such a setup, performing mainly text-entry oriented tasks such as composing email messages or editing documents may be feasible, but other common tasks that typically demand the use of a mouse may be extremely difficult or nearly impossible, such as drawing or creating diagrams, manipulating a scrollable or zoomable interface, and so on.

Even with such limitations, people such as Philip Chavez, a self-described "voice painter," have been using command-based pointer control to painstakingly create digital artwork using Microsoft Paint and commands such as "move mouse upper left… faster… stop" [10]. This illustrates how, for certain individuals, hands-free control of the computer through voice is one of the few options available to them, and that they are willing to expend a great amount of effort in trying to learn to use it and become proficient.

What we need for speech to be adopted as a first-class citizen of the input modalities is for it to afford the level of direct manipulation offered by the mouse. In the long term, it would be ideal if an entire user interface paradigm could be designed that is optimal for voice interaction. This should be pursued as a research area. However, the reality for thousands of people with motor impairments is that mouse and keyboard interfaces are pervasive and they need a solution that can give them access to such interfaces.

THE VOCAL JOYSTICK

There have been a number of research prototypes that have attempted to harness non-speech vocal parameters for input.

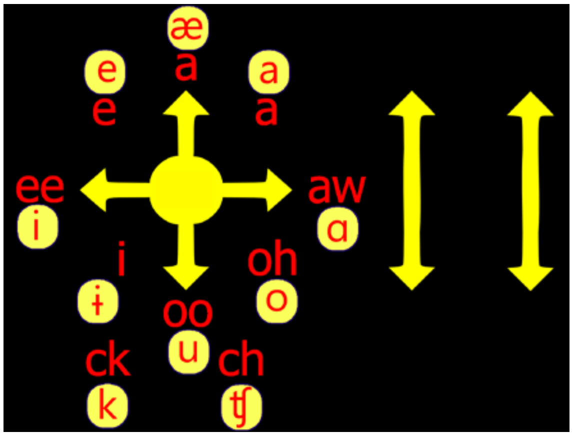


Figure 2: The "compass" shows the sounds mapped to each direction in the Vocal Joystick. The red vowels in each word approximate the sound corresponding to that direction (represented by the corresponding IPA alphabet in the adjacent circles). The Vocal Joystick also tracks loudness and pitch, as well as discrete non-vowel sounds such as "ck" and "ch".

However, as reviewed by Harada et al. [9], most of them are quite limited. The Vocal Joystick system [4] offers the greatest flexibility among these choices, by offering the closest emulation of mouse pointer control.

The key distinguishing feature of the Vocal Joystick engine is its exploitation of the continuous vowel space as the input domain. It originated from the observation that human beings can produce a fluid array of vowel sounds by smoothly varying the shape of their mouth and the position of the tongue. As shown in Figure 2, various vowel sounds are assigned to radial directions, and while the user vocalizes a sound, the mouse pointer continues to move in the corresponding direction, changing direction and speed as the user changes sound and loudness, respectively.

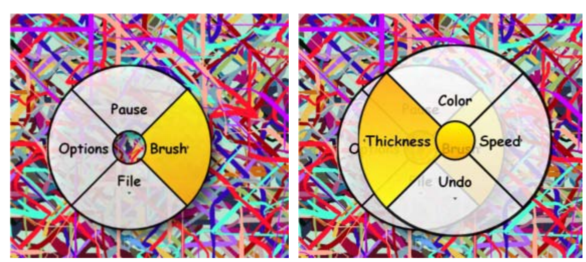
The assignment of the eight vowel sounds to each radial direction in the Vocal Joystick may seem arbitrary, but it is grounded in the relationship of the sounds created by the mouth to the position of the tongue in the sagittal plane (i.e., the up/down forward/backward position of the tongue). To map a distinct sound to each of the eight radial directions and to provide the ability to naturally transition from one sound to the next sound corresponding to the adjacent direction, the vowels along the periphery of the International Phonetic Alphabet (IPA) vowel map [5] were used. The eight particular vowels used in the Vocal Joystick were chosen because they represent the eight most distinct sounds in the vowel map periphery that are present in as many of the world's major languages as possible.

Figure 3: A screenshot of the VoiceDraw application [10] showing (a) the status bar, (b) help overlay, and (c) canvas area. The first author created this painting using only his voice in about 2.5 hours.

Applications of the Vocal Joystick

The Vocal Joystick engine has been successfully used in a number of applications beyond mouse pointer control. Although the Vocal Joystick application has been designed primarily to control the mouse pointer in a 2-D continuous space, the underlying engine's capability to classify vowel sounds and extract loudness and pitch information can be used for various other controls that may not have any relationship to 2-D space. For example, any subset of the eight vowels may be used to simulate distinct buttons for selecting among up to eight choices. One may also use only a pair of vowels to simulate a 1-D slider. Loudness and pitch can also be used to manipulate a continuous value. Due to this flexibility in the application of the Vocal Joystick signal, once the user masters the directional vowel sounds and loudness/pitch control, they will then be able to extend that skill beyond pointer control to a variety of interactions.

In *VoiceDraw* (Figure 3), the directional vowel mapping was used to control the paint brush, but it also took advantage of the other vocal parameters such as loudness to control brush thickness [10]. The program also extended the 2-D vowel mapping to control a custom widget called the vocal marking menu (Figure 4), in which the user can use a sequence of "voice gestures" to quickly select an item from a hierarchical menu. House et al. [12] applied the Vocal



(a) (b)

Figure 4: The vocal marking menu supports menu navigation using only non-speech vocalizations. The menu is invoked by issuing the discrete sound "ck". (a) The user finishes uttering "aww" (right), and is about to open the submenu by issuing the discrete sound "ch"; (b) the user finishes uttering "eee" (left) within the submenu and is about to execute the command by issuing the discrete sound "ch".

[5] http://en.wikipedia.org/wiki/Vowel

Joystick mapping to control of a robotic arm, where the various joint angles were controlled in either direction using a pair of vowel sounds from the vowel compass as well as pitch inflections. *VoicePen* [11] augmented a digital stylus by using a pair of vowel sounds to control a virtual slider to smoothly manipulate parameters such as brush stroke thickness and opacity as well as zoom level while the user controlled the stylus.

As these examples demonstrate, the degree of flexibility and expressivity of control attained once the user learns the fundamental controls of the Vocal Joystick can extend far beyond pointer control. We outline the key characteristics of the Vocal Joystick engine and its capabilities below:

- Quick response – Unlike word-based commands such as "move mouse left" in which the user has to complete the utterance before the system can act upon it, the Vocal Joystick can process and respond to vocalized sound every 10 milliseconds.
- Continuous variation of a parameter value – Due to the rapid sampling rate and ability to capture continuously varying features such as pitch and loudness, the Vocal Joystick can provide these as inputs into applications for manipulating continuous parameter values.
- Simultaneous multidimensional control – As human vocalization allows for modifying the vowel sound, loudness, and pitch at the same time, these parameters can be processed by Vocal Joystick to manipulate multiple parameters simultaneously.
- Transferable 2-D mapping – Once the 2-D mapping of the Vocal Joystick is learned, it can be transferred to other analogous mappings such as the vocal marking menu, or to more diverse applications such as controlling a robotic arm or using only one of the dimensions for linear slider control.

STAGES OF LEARNING THE VOCAL JOYSTICK

The flexibility afforded by the Vocal Joystick is also accompanied by a set of unfamiliar controls that the user must learn. Based on our past observations, we identified the following four stages that a user might go through to acquire the skills needed to use the Vocal Joystick.

Vowel Production

First, the user needs to be able to produce each of the vowels being used (typically either four- or eight-vowel mode) distinctly and consistently. Due to the way in which the vowels were chosen, some of the sounds may not exist in certain languages or dialects. The system is flexible to a degree in being able to accommodate variations in individual pronunciations of the vowel sounds through the use of the adaptation process, so the key factor is that the user is able to produce four (or eight) distinct sounds in the proximity of the original vowel sounds rather than having to be able to produce the original sounds exactly.

Vowel Direction Mapping

Once the user is able to produce the vowel sounds, they need to memorize which vowel sound corresponds to which direction. As mentioned before, there is a carefully considered rationale behind the choice of the vowel sounds and their positioning relative to each other, but the actual decision of how these vowels should be oriented in the space of radial directions was arbitrary. Therefore, the user will need to become accustomed to this mapping through repeated exposure and memorization.

Loudness Control

Currently in the Vocal Joystick, the loudness of the vocalization is mapped to the velocity of the corresponding pointer movement. At the present, we use an exponential mapping between the power of the audio signal registered through the microphone and the resulting pointer velocity [14]. Because most people have never experienced using the loudness of their voice to directly manipulate a user interface, they will need to learn this mapping.

Smooth Transitions

Being able to produce the right vowel sound and control its loudness should allow the user to effectively move the pointer to a desired target. However, if the task demands that the user follow a curvilinear trajectory, it is necessary that the user be able to smoothly transition from one vowel to another and control the rate of transition in conjunction with their loudness. Such a skill is essential in order to perform tasks such as creating drawings or playing games, which require continuously varying motion.

LONGITUDINAL STUDY OF THE VOCAL JOYSTICK

There has been prior evidence to suggest that the Vocal Joystick can be used by "expert" users to perform mouse-oriented tasks effectively, as demonstrated by our videos of web surfing, game playing, and even robotic arm control,[6] and by the artwork created using the VoiceDraw program (Figure 3) [10]. In all but one of these instances, the "expert" users were the creators of the Vocal Joystick (the exception being Philip Chavez).

It has also been shown that the Vocal Joystick can be used by novices with very little training. Seven users with no prior experience with Vocal Joystick were able to learn the vowel mapping and perform basic mouse tasks to browse through a web site and navigate an online map [4]. When children between the ages of 7 and 18 were given the opportunity to try out the VoiceDraw application during a public exhibit, with only a few minutes of training, they were able to create highly expressive drawings [10].

The unanswered questions that remain are:

- How long does it take people to reach the "expert" level of performance on Vocal Joystick?
- Can the Vocal Joystick be used effectively by our primary target group of people with motor impairments?

[6] http://www.vocaljoystick.org/video_demos.htm

MI group

ID	Gender	Age	Impairment	Time with impairment	Effect on mouse usage
P01	M	52	Multiple sclerosis	7 years	Fatigue and pain
P02	M	51	Idiopathic neuropathy	Since childhood	Fatigue
P03	F	20	Muscular dystrophy	Since birth	Fatigue, difficulty moving
P04	F	30	Cerebral palsy (CP), Fibromyalgia (FM), Dyslexia	Since birth (CP), 13 years (FM)	Fatigue and spasm, hard to move and slow
P05	F	57	Parkinson's disease	16 years	Erratic movements and lack of reflex

NMI group

ID	Gender	Age
P06	F	30
P07	M	23
P09	M	19
P10	F	20

Table 1: Basic demographic information about the participants in the longitudinal study. The table on the left lists the participants with motor impairments (MI group), and the table on the right lists the participants without motor impairments (NMI group).

- What issues related to the usability of Vocal Joystick have we not discovered yet due to the limited duration of usage by users up to this point?

To answer these questions, we conducted a longitudinal study involving both motor impaired and non-impaired participants spanning 10 sessions for each participant.

Foci of Our Study

Our primary focus in this investigation is not to evaluate the immediate usability of the Vocal Joystick system, but rather to assess the learning experience and benefit that it could yield to users. This is particularly driven by the fact that the system is primarily targeted for individuals with motor impairments, for whom there may not be many alternatives for efficient access to computers. Because of this, these users may be willing to tolerate a steep learning curve and longer time investment if the ultimate outcome is a significant increase in their ability to use computers.

Therefore, it will not be appropriate to evaluate the system using a single-session. It would also not suffice to evaluate the system only with people who do not have a targeted disability, which happens unfortunately often in assistive technology research [2,3,5,16]. Such choices could be likened to a hypothetical scenario in which the inventor of the violin decided to test his new instrument by having people with no musical training play with it for half an hour.

In the ideal case, a tool should be both immediately effective for novice users and yield high long-term gains as the user becomes more experienced. However, care needs to be taken to ensure that the desire to improve the immediate effectiveness of the tool does not lead to premature rejection of alternatives that hold longer-term potential.

Another focus of our study is on the learning process that each user goes through in acquiring the skills to use the tool. Although a highly controlled and structured study would be ideal from a comparative point of view, we are interested in uncovering the issues that each user encounters and what works best for them in facilitating the learning process. We do not wish to force the subjects through a rigid set of protocols and sacrifice the quality of individualized observations for the sake of obtaining statistically pure data. Therefore, during our study, when a participant had specific issues or difficulties, we worked with them to identify the source of the issue and to find the way to address it that was most suited for that participant.

Participants

We recruited ten participants for our longitudinal study (one had to drop out due to visa issues). Of the remaining nine, five had some form of motor impairment that affected the use of the hands in controlling a mouse (MI group), and the others had no motor impairments (NMI group). All were native English speakers. Table 1 shows basic demographic information about the participants. Participants P01 through P05 form the MI group, and P06 through P10 form the NMI group. Additional descriptions of the MI group participants are presented next. Sears and Young [18] provide a more in-depth coverage of some of the physical impairments in the context of computer technology.

P01 has had multiple sclerosis for 7 years, a progressive disease of the central nervous system that affects his ability to type or use the mouse for prolonged periods. He uses Dragon Naturally Speaking at work for dictating text, and mentioned that when he does try to type using the keyboard, his hands feel arthritic and he easily gets tired and sore. He stated that during exacerbations (i.e., sudden worsening of symptoms), his hands feel like they are on fire and he cannot grasp anything. He also noted that he frequently has to clear his throat, as multiple sclerosis can affect the larynx, but he was able to vocalize normally.

P02 has had idiopathic neuropathy since childhood, a disorder affecting the peripheral nerves leading him to experience pain and tingling in his hands when symptoms surface. He mentioned that his medication helps the pain from becoming too severe, and he is able to finger type on a keyboard and operate the mouse, although with increasing discomfort over time.

P03 has had muscular dystrophy since birth, a progressive muscle disorder that affects her range of mobility and manual control. She uses a powered wheelchair, and is able to move her arms but is unable to turn her palms face down or fully extend her elbows. She tried using speech recognition software a number of years ago, but stopped using it due to low accuracy. She can use the keyboard with both hands by using the backs of her fingers, but she mentions that she can only do so for several minutes before she gets very fatigued. She can also use the mouse by gripping it with the back of her two hands, but she finds it very hard to move and also tiring. She prefers to use a touchpad on her laptop computer, which she can operate by using the knuckle on her finger.

P04 has had multiple conditions that affect her motor abilities. She has had spastic cerebral palsy throughout her life, a non-progressive condition that affects her muscle strength and also causes her to spasm frequently when attempting to use her hands or arms. She has also had fibromyalgia for 13 years, a chronic condition characterized by widespread pain in the muscles. She also uses a powered wheelchair with an electronic ventilator that periodically pumps air out through a breathing tube near her headrest. In our sessions, she only needed the ventilator occasionally, and therefore was able to turn it off for most of the time, although we found that even when it was on, the Vocal Joystick was not affected. Generally, her speech ability is unimpaired, although she does occasionally experience shortness of breath. She also has dyslexia, which makes it hard for her to process written cues such as the words in the Vocal Joystick vowel compass. She is also able to use a keyboard and a mouse, although it is extremely tiring for her. She prefers instead to use a touchpad on her laptop computer.

P05 has had Parkinson's disease for 16 years, a degenerative nervous system disorder that has reduced her flexibility and reflexes, and now affects both sides of her body. She has mild resting tremors, and her voice tended to tremble and be monotonic, a common symptom of the condition. She can use the mouse, but complains that she often ends up clicking the buttons unintentionally, and has a hard time with continuous motion.

Study Setup

The next section outlines the general procedure for each session that was designed before the start of the study. However, due to the significant individual differences among participants, especially within the MI group, the actual procedure within each session for each participant was varied to accommodate the specific difficulties that the participant was facing at the time, such as by spending extra time on vowel coaching.

Procedure

The study took place in a lab over a period of 10 sessions for each participant. Each session was one hour long, except for the first and last sessions being 90 minutes long due to system introduction and the final comparative assessment using each participant's preferred pointing device. For each participant, the time between consecutive sessions was at least 3 hours and no more than 48 hours. The participants were compensated for their time by being paid \$25 for the first session, \$10 each for the subsequent eight sessions, and \$145 for the 10^{th} session.

During the first session, participants were introduced to the Vocal Joystick control method and to the vowel compass, and were shown how to use the vowel feedback tool (Figure 5a). In the vowel feedback tool, the user can click on the speaker icon below each word to see and hear a video of the corresponding sound being pronounced. The user can also vocalize and see the system's recognition result, indicated by the yellow arrow. For our study, after the review of vowel sounds, the participant's vowel utterances were collected using the Vocal Joystick application to build the initial user profile.

The general structure of the rest of the sessions was as follows. At the beginning of the session, the participant was tested on their vowel recall to see if they remembered the mappings of the sounds to directions. Next, they used the vowel feedback tool to review the vowel sounds and practice tuning their vocalization using the feedback. They were given the option of readapting the Vocal Joystick user profile if they felt that the system was not responding well.

(a) Vowel feedback tool **(b) Target acquisition task test screen** **(c) Steering task test screen**

Figure 5: The three stages that the participants went through during each study session. (a) The speaker icon below each word can be clicked to both hear and see the video of the corresponding sound. The participant can also vocalize themselves and see the system's recognition result, indicated by the yellow arrow. (b) The screen shows the horizontal targets in the Fitts' law reciprocal target acquisition task. (c) The screen shows the circular tunnel in the steering task condition with the entry and exit target shown as the vertical bar and the trail of the pointer movement shown in the tunnel.

Following the vowel testing, the participants engaged in two stages of tasks; target acquisition stage, and steering stage. A vowel compass printout was placed next to the monitor if they needed it for reference. During the target acquisition stage, the participants first practiced by navigating through a series of web sites using the Vocal Joystick. During this phase, they were free to adjust the sensitivity of the Vocal Joystick or readapt the user profile if desired. After the practice phase, they engaged in a Fitts' law reciprocal target acquisition task (Figure 5b, as utilized by Harada et al. [9]). In this task, the participant is presented with two bars of a certain width separated by a certain distance, and is asked to click on the bars alternating as quickly and accurately as possible. The factors and levels for the task were as follows:

- Distance (D): {150, 280 pixels}
- Width (W): {30, 80 pixels}
- Angles (□):□{0, 45, 90, 135 degrees}

The combination of the target amplitudes and widths yielded four indices of difficulty (ID): 1.52, 2.17, 2.59, and 3.37 bits. Because the task was reciprocal, the four angles covered all eight cardinal and ordinal directions. For each of the D×W×□ conditions, the participants were presented with 4 trials, where two trials compose one round trip set of clicks between the targets. Since we aggregated the angles in our analysis, this yielded 16 trials per ID per participant per session.

During the steering stage, the participants first practiced by playing a game called *FishTales*,[7] and they traced a set of figure eight paths in VoiceDraw (Figure 1). After the practice phase, they engaged in a steering task [1] (Figure 5c) in which the participants were asked to steer the pointer through circular tunnels of varying widths and radii. The factors and levels for the task were as follows:

- Tunnel radius (R): {100, 200 pixels}
- Tunnel width (W): {100, 140 pixels}

The combination of the tunnel radii and widths yielded four indices of difficulty (ID) of 4.49, 6.28, 8.98, and 12.57 bits. For each of the R×W conditions, the participants were presented with 8 trials, where half of those trials where clockwise and the other half were counter-clockwise. We aggregated the rotation direction in our analysis.

Although we used the test frameworks for Fitts' law and the steering law in our sessions, the main objective here was to collect comparable task completion times across sessions, and not specifically to extract a model fit for each session (except for the final session). We do have prior evidence to suggest, however, that Fitts' law is indeed a good predictor of speed-accuracy tradeoff for the Vocal Joystick [9]. Similar verification has yet to be made for the steering law.

[7] http://www.funny-games.biz/fishtales.html

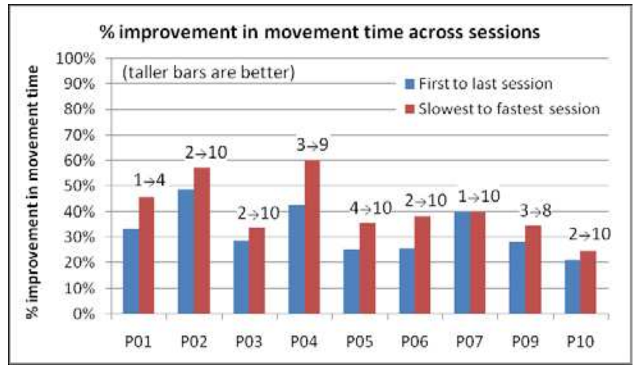


Figure 6: The percent improvement in average movement time between the first and the last session (left bar) and between the slowest and the fastest session (right bar) for each participant. Numbers above the right bars indicate the session numbers corresponding to the slowest and fastest sessions.

Equipment Setup

The study sessions were conducted using a Dell Optiplex GX280 desktop computer running Windows Vista Business with an Intel Pentium 4 processor clocked at 3.4GHz with 1.5GB of RAM. The computer was connected to a Dell 2001FP 20" monitor displaying 1280×960 pixels at a resolution of 96 dpi. A Plantronics DSP400 USB headset microphone was used for sound input.

Results

We were able to collect a significant amount of data from our 99 hours with nine participants. In the following sections, we highlight some of the key results from the longitudinal user study, associating each stage of the study session to the stages of learning presented earlier.

Vowel Recall (Vowel Production and Direction Mapping)

All participants were able to memorize the vowel-to-direction mapping during the 10-session period. On average, the participants were able to correctly recall all eight vowels and conduct the entire session without the aid of the vowel compass after the 5th session.

The accuracy of vowel *production*, on the other hand, varied widely among participants, such that although they could recall what the sound should be for a particular direction, they had difficulty vocalizing it in such a way that the Vocal Joystick consistently recognized it as the intended sound. The next subsection provides further detail.

Target Acquisition (Loudness Control)

Figure 6 shows the percent improvement in average movement time for the target acquisition task between the first and last sessions, as well as between the sessions with the slowest and fastest average movement times. Overall, participants demonstrated improvement in their performance over the 10-session period of at least 20%. Among the NMI group, average improvement in movement time ranged from 25% to 49% (34% to 60% if comparing slowest to fastest sessions), and 21% to 40% (24% to 40% for slowest to fastest) among the MI group.

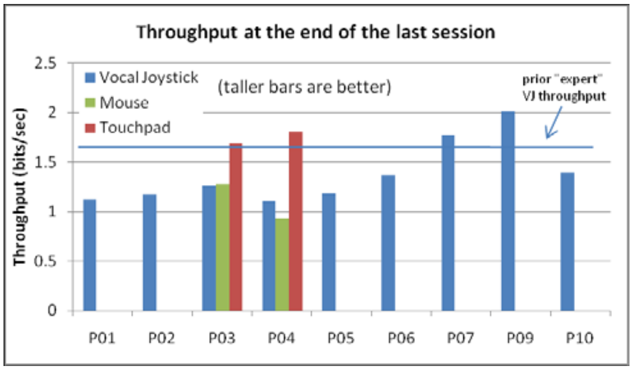


Figure 7: The Fitts' throughput measure for the Vocal Joystick at the end of the 10th session. For P03 and P04 whose preferred pointing device is the touchpad, their touchpad and mouse throughputs are also shown. All other participants' mouse throughput ranged from 3.9 to 6.2 bits/sec (mean of 5.1).

Figure 7 shows the final Fitts' throughput achieved by each participant after the last session using the Vocal Joystick. The section for P03 and P04 also contains the throughput that each of these participants achieved using a mouse and their preferred device, a touchpad. For the rest of the participants, their preferred device (mouse) throughput was measured but is omitted from the figure due to scale. The MI group's mouse throughput was in the range of 3.9 to 4.6 bits/sec. For the NMI group, that range was 4.9 to 6.2 bits/sec. Note that by the final session, P03 and P04's Vocal Joystick throughput had equaled or exceeded their mouse throughput, and reached 75% and 61% of their preferred touchpad throughput, respectively.

Two of our NMI participants have exceeded our previously observed expert Vocal Joystick throughput of 1.65 bits/sec [9] (Figure 7). The average throughput for the NMI group is 1.64 bits/sec, which is comparable to the prior expert VJ throughput. The average throughput for the MI group is slightly lower at 1.17 bits/sec, or 70% of the NMI group. There may be a number of reasons for this difference, which needs to be investigated in more detail, including the difference in age and amount of experience with computers.

If we were to project the same rate of learning as exhibited during our study into the future by fitting a power curve to the per-session data, then P03 and P04 are projected to attain the previously set "expert" Vocal Joystick throughput level after 36 and 17 more sessions, respectively. If they attempt to attain the same throughput as their touchpad, P03 will only need another 8 sessions and P04 another 11 sessions. This shows that the Vocal Joystick has the potential of offering comparable performance to current devices without the need for physical manipulation or an unreasonably long practice period. Figure 8 shows the trend of the average movement time for each group. It is not clear why there was a slight upward trend around the third quarter of the sessions.

By the end of the last session, participants were able to successfully use the Vocal Joystick to navigate a website. A sample clip showing a participant navigating through Yelp.com and interacting with their Google maps web control, as well as another participant successfully tracing a figure eight in VoiceDraw, can be seen on our web site.[8]

Steering (Smooth Transitions and Loudness Control)

The steering task proved to be challenging for the MI group, most of whom had difficulty getting consistent speed response as they changed their vowel sounds, resulting in high error rates. The NMI group fared better, exhibiting mean improvement in task completion time of 32% between the first and last sessions and 43% between the slowest and fastest sessions.

Difficulties Faced by the Participants

Although there were various issues specific to each individual, several prominent findings surfaced across a majority of the participants in the areas of vowel production and loudness control. We describe each in detail below.

Vowel Production

A great number of participants had difficulty consistently and distinctly vocalizing the sounds represented by "a", "ɑ" and "ɨ" in Figure 2. Depending on the region of origin, the native pronunciation of these words does not contain the intended sound, and the user is unable to distinguish it from the adjacent sound. This is one of the difficult tradeoffs that the Vocal Joystick system has to make. To provide the expressiveness and the ability to smoothly move at arbitrary angles, the system needs the user to be able to produce as many distinct vowel sounds as possible. However, not everyone can produce or even perceive as many distinct sounds, and there are significant differences among individuals depending on their origin and dialect. Although the system adapts to the user's sounds, it operates best when the sounds are close to the originally trained sounds, which were chosen to be maximally distinct in acoustic space. In order to attempt to deal with the sounds that were giving them trouble, participants came up with a variety of alternate representations to help them make the correct sound, such as "all" and "'ol" for "ɑ," and "good" and "err" for "ɨ".

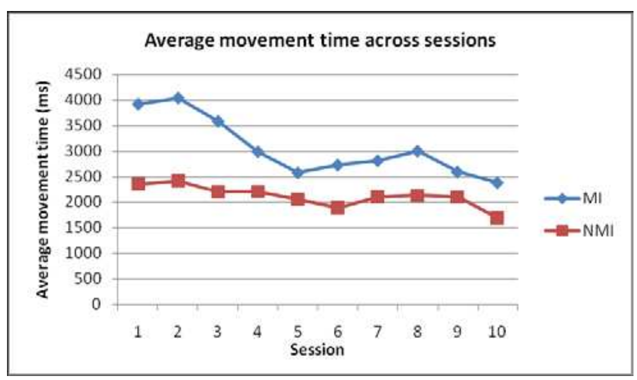


Figure 8: The average movement time over each session for the participants with (MI) and without (NMI) motor impairments.

[8] http://www.vocaljoystick.org/videos/chi2009/

Loudness Control

Some users did not have much control over the loudness of their voice, making it challenging to perform tasks that required movement at various levels of scale, such as moving over a long distance or clicking on a small target.

There was also an issue where the speed response of the pointer varied significantly under certain situations, and the movement along a particular direction was significantly faster or slower than the other directions, even though the user was vocalizing at a relatively fixed loudness. For this reason, most of the participants did well in the VoiceDraw tracing task where the brush was set to be at a fixed speed, but had difficulty with the steering task where the pointer speed was controlled by loudness.

These issues all point to one of the most important areas of improvement for the Vocal Joystick, which is the need for a better mapping between the loudness of the utterance and the resulting speed of the pointer.

Observing Each Motor Impaired Participant

To better understand specific issues that may affect our target user group, we highlight some key observations specific to each participant in the MI group below.

There were a number of times when P01 had to clear his throat. When he did, it was picked up by the microphone, but did not seem to affect the pointer movement. He also mentioned that it really helped him to be able to both hear the vowel sound and see the mouth shape being made when trying to learn the vowels, in reference to the vowel feedback tool (Figure 5a).

P02's loudness control was limited at first, but with practice he learned to modulate it to control the speed. He had quite a hoarse and loud voice, which appeared to throw off the Vocal Joystick engine quite a bit. As he progressed, his vowel quality improved as well.

P03 had quite a soft voice and shorter vocalizations, which seemed to be restricted by her reduced lung capacity. This led her to move the pointer in small segments, which did not leave much room for smooth vowel sweep transitions.

One of the concerns we had was regarding P04's ventilator, but the Vocal Joystick was relatively unaffected by the sound from the ventilator, which made a puffing sound every ten seconds. Another challenge for P04 was her dyslexia, which made it difficult for her to process the vowel compass and produce the sound given a direction stimulus. She was able to perform fine when she was given extra time to think of the desired sound.

She also commented that, "I'm trying to work with the system too, since most people don't have someone who would be able to make small adjustments each time." This reflects the attitude of a number of people with disabilities who are willing to invest more time in learning a system, especially if there are no other alternatives due to situational, monetary, or availability constraints.

During the first several sessions of the study, P05 had significant difficulty producing the sounds corresponding to *right*, *bottom-right*, and *bottom-left* on the vowel compass. After much vowel coaching over the first five sessions, she was able to produce distinct sounds for all directions and get the system to recognize them consistently except for *right*, with which she continued to have difficulty. Her voice had the tendency to tremble, especially when she tried to sustain a sound, causing a drop in recognition accuracy.

Other Observations

It appeared that those with prior music or voice lesson experience (P03: piano; P06: poetry reading/voice class; P09: violin; P10: choir) seemed to have less trouble making the vowel sounds. Those without such background (P01, P02, and P05) found the vowel training tool's video, sound, and arrow feedback to be especially helpful.

We also observed that a number of times, the system would temporarily stop responding, or the movement of the pointer became erratic. The experimenter could identify that for many of these instances the main cause was due to the participant vocalizing too loudly or with extraneous sounds at the beginning of the utterance.

FUTURE WORK

There needs to be explicit and constant feedback available to the user that reflects the loudness of the user's vocalization as detected by the system, as well as the system's confidence level regarding whether the utterance is one of the vowel sounds or a non-vowel sound. The issue of providing concise, meaningful feedback that the user can process to make appropriate adjustments is an important topic we will be pursuing in our future research.

One of the most salient issues with the system that surfaced as a result of our study was the need for a better mapping between the vocalization's loudness and the mouse pointer's speed. Although this issue has been investigated before [14], a more thorough study needs to be conducted to examine this relationship between loudness and speed with a greater number of users, drawing from the design of similar devices such as isometric joysticks. An intuitive interface also needs to be developed to allow the user to adjust the speed mapping to their preference.

CONCLUSION

Speech and voice-based user interface control holds great promise for enabling fluid hands-free computer interaction, especially for users who have limited motor abilities. However, current speech-based tools available today lack a key component for fully realizing this potential, namely the ability to replicate the expressivity and direct manipulation metaphor afforded by the mouse.

The Vocal Joystick engine may be the key technology that could bring the flexibility and expressivity of the mouse to voice-based interaction. Our goal in this study was to better understand the issues faced by users with motor impairments in learning a novel voice-based input modality. We also wanted to know how quickly various aspects of the skills are acquired and how proficient people can get in using non-speech vocalization to control the mouse pointer.

Over the 10 session period, the participants were able to learn the vowel mappings and showed marked improvement in their target acquisition performance. At the end of the ten session period, the NMI group reached the same level of performance as the previously measured “expert” Vocal Joystick performance, and the MI group was able to reach 70% of that. Two MI participants P03 and P04 approached the performance of their preferred device.

These findings demonstrate the value that the Vocal Joystick technology can provide in expanding the domain of voice-driven computer interaction. The Vocal Joystick application is available for public download[9] and the API will be available soon. This opens up the possibility of creating many novel applications that leverage the great capacity of human voice for both users with and without disabilities. We must strive to prevent the digital divide between these two groups from growing wider.

ACKNOWLEDGMENTS

We thank all of our study participants for their time and patience in working with our system. This work is supported in part by the National Science Foundation under grants IIS-0326382 and IIS-0811063. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Accot, J. and Zhai, S. Beyond Fitts' law: models for trajectory-based HCI tasks. In *Proc. CHI 1997*, ACM Press (1997), 295-302.
2. Bates, R. Enhancing the performance of eye and head mice: a validated assessment method and an investigation into the performance of eye and head based assistive technology pointing devices. *PhD Thesis*, De Montfort University, May 2006.
3. Bates, R. and Istance, H.O. Why are eye mice unpopular? A detailed comparison of head and eye controlled assistive technology pointing devices. In *Universal Access in the Information Society 2*, 3 (2003), 280-290.
4. Bilmes, J.A., Li, X., Malkin, J., Kilanski, K., Wright, R., Kirchhoff, K., Subramanya, A., Harada, S., Landay, J.A., Dowden, P., and Chizeck, H. The Vocal Joystick: A voice-based human-computer interface for individuals with motor impairments. In *Proc. HLT/EMNLP 2005*, ACL (2005), 995-1002.
5. Capilouto, G.J., Higginbotham, D.J., McClenaghan, B., Williams, H.G., and Dickerson, J. Performance investigation of a head-operated device and expanded membrane cursor keys in a target acquisition task. In *Technology and Disability 17*, 3 (2005), 173-183.

[9] Visit http://www.vocaljoystick.org for more information.

6. Coyne, K.P. (2005) Conducting simple usability studies with users with disabilities. In *Proc. HCI Int'l 2005*, Lawrence Erlbaum Assoc. (2005). On proceedings CD.
7. Dai, L., Goldman, R., Sears, A., and Lozier, J. Speech-based cursor control: a study of grid-based solutions. In *SIGACCESS Accessibility and Computing*, 77-78 (2003), 94-101.
8. de Mauro, C., Gori, M., Maggini, M., and Martinelli, E. Easy access to graphical interfaces by voice mouse. Technical report, Università di Siena (2001). Available from the author at: maggini@dii.unisi.it.
9. Harada, S., Landay, J.A., Malkin, J., Li, X., and Bilmes, J.A. The Vocal Joystick: evaluation of voice-based cursor control techniques. In *Proc. ASSETS 2006*, ACM Press (2006), 197-204.
10. Harada, S., Wobbrock, J.O., and Landay, J.A. VoiceDraw: a hands-free voice-driven drawing application for people with motor impairments. In *Proc. ASSETS 2007*, ACM Press (2007), 27-34.
11. Harada, S., Saponas, T.S., and Landay, J.A. VoicePen: augmenting pen input with simultaneous non-linguistic vocalization. In *Proc. ICMI 2007*, ACM Press (2007), 178-185.
12. House, B., Malkin, J., and Bilmes, J.A. The VoiceBot: a voice controlled robot arm. In *Proc. CHI 2009*, ACM Press (2009), to appear.
13. Igarashi, T. and Hughes, J.F. Voice as sound: using non-verbal voice input for interactive control. In *Proc. UIST 2001*, ACM Press (2001), 155-156.
14. Malkin, J., Li, X., and Bilmes, J.A. Energy and loudness for speed control in the Vocal Joystick. In *IEEE Automatic Speech Recognition and Understanding Workshop*, (2005), 409-414.
15. Mihara, Y., Shibayama, E. and Takahashi, S. The migratory cursor: Accurate speech-based cursor movement by moving multiple ghost cursors using non-verbal vocalizations. In *Proc. ASSETS 2005*, ACM Press (2005), 76-83.
16. Mahmud, M., Sporka, A.J., Kurniawan, S.H., and Slavík, P. A comparative longitudinal study of non-verbal mouse pointer. In *Proc. INTERACT 2007*, Springer Berlin (2007), 489-502.
17. Riemer-Reiss, M.L. and Wacker, R.R. Factors associated with assistive technology discontinuance among individuals with disabilities. In *Journal of Rehabilitation 66*, 3 (2000), 44-50.
18. Sears, A. and Young, M. Physical disabilities and computing technologies: an analysis of impairments. In *the Human-Computer interaction Handbook*, Lawrence Erlbaum Associates (2003), 482-503.
19. Sporka, A.J., Kurniawan, S.H., and Slavík, P. Acoustic control of mouse pointer. In *Universal Access in the Information Society 4*, 3 (2006), 237-245.