Making Speech-Based Assistive Technology Work for a Real User

William Li[1], Don Fredette[2], Alexander Burnham[2], Bob Lamoureux[2], Marva Serotkin[2], Seth Teller[1]

[1] EECS/CSAIL, Massachusetts of Institute of Technology, Cambridge, USA
[2] The Boston Home, Boston, USA

wli@csail.mit.edu, dfredette@thebostonhome.org, aburnham@thebostonhome.org, bob@lamoureux.com, mserotkin@thebostonhome.org, teller@mit.edu

Abstract

We present a customized speech-activated email system that is the product of efforts focused on a single target user with high speech recognition error rates. The system, which includes off-the-shelf and custom hardware and software, allows the user to use speech to send emails with recorded audio attachments. Over the past 16 months, our target user has sent and received hundreds of emails and has integrated the system into his daily life. Key factors contributing to the long-term adoption of the device include our extended efforts to understand the target user over multiple years, iterative design, and the collaboration of our multidisciplinary team of assistive technology (AT) designers, clinicians, software developers, and researchers. Overall, we ask: if we set our sights on developing and supporting a technology that someone will actually use daily, what can we learn? We share our approach, system design, user observation and findings, with implications for speech-based AT research and development.

Index Terms: speech interfaces, usability, assistive technology

1. Introduction

Functional access to computers and other devices can help people with physical impairments stay connected with others, access information, or control the environment. For many individuals who cannot use touch-based interfaces like keyboards and mice, automatic speech recognition (ASR) could be a viable alternative access method. However, ASR systems can be challenging to use for individuals who have speech difficulties, since such systems are typically not trained on, or designed to be used by, people in these relatively small populations. These technical challenges mean that ASR-based assistive technology (AT) often falls short of its potential as an access equalizer for people with disabilities [Young2010].

The present paper describes a system that has enabled a single individual, an adult wheelchair user with advanced secondary progressive multiple sclerosis (SPMS), to send emails without assistance on a regular basis. We offer details on the multi-year process required to design and implement speech-based email system that has made a positive impact in his daily life. Where commercial off-the-shelf components existed and were appropriate, we tried and incorporated them. Our work has involved rehabilitation technology staff, clinicians, family members, and researchers who worked to understand his context, needs, and preferences in order to develop an appropriate, long-lasting AT intervention.

Our approach differs from most academic research on speech recognition for individuals with disabilities, which often prioritizes novel algorithms, new models, or superiority over baselines in short-term user studies. While we certainly do not dismiss these contributions – we follow these research paradigms most of the time ourselves – our deviation is deliberate. Specifically, in this work, we ask: What is required to *actually* deploy speech-based assistive technology and have tangible impact on a user's life? What can we learn from this implementation process?

This paper goes beyond describing an end product – we also discuss the target user's context and our design process. We introduce the target user (Section 2) and his past AT usage (Section 3), then describe the speech-based email system (Section 4). We provide details on how staff and clinicians, family and friends, students in a design-based assistive technology course, researchers, and, most importantly, our target user himself were involved in identifying the shortcomings and utility of various AT interventions. Section 5 discusses our findings: our target user's actual email usage over a 16-month period. We discuss our insights and their implications for researchers and practitioners in Section 6.

2. User and design constraints

Our work occurred at The Boston Home (TBH), a residence and center for care for adults with multiple sclerosis and other progressive neurological conditions. The 96 residents at TBH receive nursing, medical, physical therapy, speech-language pathology, and assistive technology services on site, in addition to an array of social, artistic, and residential activities.

2.1. Description of target user

Our target user is a middle-aged male living with advanced SPMS. He is a power wheelchair user, has minimal control of his arms and no active movement in his legs due to spastic quadriplegia, and vision challenges due to SPMS-associated optic neuritis. Meanwhile, he has high cognitive function, good working memory, and generally an eagerness to try new AT.

Given these limitations, ASR could be a promising access channel. However, our target user's speech is not recognized accurately by existing, large-vocabulary speech recognizers. Challenges include abnormally strained vocal quality, reduced respiratory support for duration and intensity of phonation, variable pitch control (vocal fry) over the course of a single utterance, and dialectical variation from standard American English, which he acquired as a second language in adulthood. Our target user's successes and difficulties of using ASR-based AT is discussed in Section 3.

2.2. Goal: Computer and email access

Our target user seeks greater independence. Any device that allows him to rely less on other individuals can have a positive impact. Our current goal is to enable independent (and thus private) computer access, particularly to email, which would help him better stay in touch with friends and family.

Our close interaction with our target user allowed us to define some key characteristics of our eventual system. The

need for system training by the user and adjustment by outside experts should be minimized, even though his abilities can fluctuate over time. Meanwhile, the appearance and user interface of any solution is very important, particularly those that require mounting hardware on the target user's wheelchair, body, or living space.

3. Other assistive technology usage

Our team is intimately familiar with our target user's past and current AT. This knowledge helped us understand what might work for email access. We describe both speech and non-speech devices to illustrate where ASR has been used and where other channels were more appropriate.

Wheelchair control: Our target user operates a power wheelchair using proximity switches embedded in his headrest. He has independent control of driving, adjusting speed, tilting the chair, and changing modes. The headrest proximity switches have proven to be a robust access pathway for the target user's wheelchair. By using switches to operate in different modes and by activating combinations of switches to perform different functions, he can control dozens of wheelchair functions independently.

Television control: Our target user has an InVoca 3.0 Voice Activated Remote Control for controlling his television. This commercially available device allows users to program custom keywords that are transmitted as infrared signals, similar to any conventional TV remote control. It rests in a custom-built wooden stand on our target user's wheelchair tray, and he can instruct a caregiver to place the remote control in its recharge cradle (which is not on the wheelchair) at night.

The InVoca has worked well, even in environments with television or other ambient noise. Its major limitation is that it can only handle approximately 20 words or phrases. In addition, fluctuations in our target user's voice (even the common cold) can present significant challenges.

Telephone control: The target user has a voice-activated telephone system. Typically, a caregiver helps him don a headset connected to his landline telephone. From that point onwards, he uses a breath-activated switch to cycle through a preset list of telephone numbers. One of these preset numbers is tied to a commercially available voice recognition virtual assistant service, which contains an extended address book. This setup allows him to dial more than 50 contacts.

Our target user has had considerable success with this system and continues to use it for telephone calls, but the need for outside assistance reduces its convenience and his privacy. Furthermore, since our target user likes to communicate with family and friends in different time zones, it is not always feasible to coordinate mutually agreeable phone scheduling. An asynchronous communication medium like email could be useful for staying in touch with these contacts.

Spoken dialogue system: Our target user participated in a study that evaluated an assistive probabilistic dialogue system. This work hypothesized that that using confirmation questions to clarify the user's intent would help improve dialogue success rates for high-error speakers (the concept error rate of our target user in this study was 56.7%). As described in [4], the system helped the user complete more dialogues successfully in a supervised experimental setting, compared to a simpler baseline. While promising, the dialogue system would need to be deployed in a longer study to determine whether it is sufficiently useful for our target user.

3.1. Computer access

Our target user has tried numerous devices for desktop computer access with mixed success. While each of these technologies had drawbacks, they contributed to our insight into the user's preferences and abilities.

First, despite training commercial speech recognition software (Nuance Dragon NaturallySpeaking 7.0, and later, 10.0) with our target user's speech and adjusting the settings to the best of our ability, such software packages were too unreliable to allow him to use a desktop computer effectively. Our target user would often have to resort to time-consuming, lower-level mouse-scrolling commands instead of faster shortcut commands. Moreover, some software programs, such as browser-based Google Gmail, were not optimized for speech-based access, thereby increasing the failure rate.

We also tested non-speech access channels. Our target user tried using a head mouse, in which an infrared camera follows an infrared-reflecting sticker controlled by head movements, combined with an onscreen keyboard like Dasher [2]. Despite his use of headrest proximity switches, this method proved challenging: he experienced rapid onset of fatigue, double vision, and exacerbation of facial pain from SPMS-associated trigeminal neuralgia from the head and neck movements required to operate the headmouse successfully.

To address these speech-recognition and user-interface challenges, a team of undergraduate students in a semester-long course called Principles and Practice of Assistive Technology (PPAT) focused on how to make a desktop-computer setup more usable for our target user [3]. They evaluated different microphone stands, computer setups, and speech recognition software in our target user's bedroom. By working closely with the target user, the team determined that a desktop computer with the target user's large television set as a display would be a workable solution. Their work contributed to the groundwork for our current solution, which we describe next.

4. Email system description

The current system is situated in our target user's bedroom and allows him to keep in touch with friends and family through emails. Our customized email client has two components that make it effective for the target user: first, the user interface is optimized for speech-based access, with the ability to skip down to the desired message, open messages, reply, and delete messages with single voice commands. Second, to overcome speech recognition limitations, the emails are in the form of 20, 30 or 45-second *audio messages,* not transcribed text, that are sent as an attachment. Figure 1 shows a schematic of the entire user, hardware, and software setup, while Figure 2 shows the actual setup in his bedroom at TBH.

4.1. Hardware: Computer, screen, and audio capture

A large, flat-screen television serves as the display for a Windows 7-based computer. The target user also watches television on this screen, so he is comfortable viewing it for extended time periods.

Voice input occurs through two audio capture devices: First, we use the aforementioned InVoca device to switch between the cable television services and computer display inputs. As before, this device sits on the target user's wheelchair tray. Second, to record audio email messages, we use a Microsoft Kinect device which includes an array

microphone. Although a close-talking microphone or headset could result in a clearer voice signal, these alternatives would require more precise positioning and outside assistance. We found that the Kinect's built-in mechanisms to improve speech capture (such as sound localization and beamforming) worked well for the target user's needs.

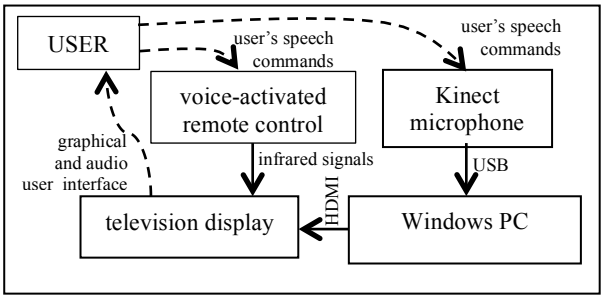

Figure 1: *Schematic of speech-based email client.*

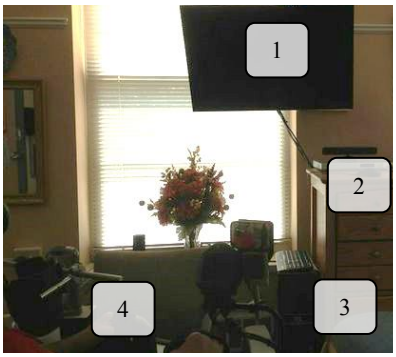

Figure 2: *Actual bedroom setup with 1) television, 2) Kinect, 3) computer, and 4) wheelchair with voice activated remote control*

4.2. Speech recognition and customized email client

Our system uses Windows Speech Recognition, which has well-supported Kinect application programming interfaces (APIs) to process the Kinect's audio stream. Based on the current mode of the software (browsing or composing messages), a small set of custom grammar files are dynamically loaded. Setting this constraint dramatically improves the recognition rate since the grammar is targeted to the task at hand. The grammar is set to recognize vocabulary for one of about 45 pre-determined phrases required for the custom email software to function.

We developed a customized email client for our target user. As shown in Figure 3, the user interface shows a green square to indicate that the speech recognizer is active; a text box displaying the currently recognized speech (which is "Go" in Figure 3), and the "From" and "Subject" headers for several email messages. At the end of an utterance, the email client parses the recognized speech and also shows a percentage confidence score for the utterance in large text.

The target user can move the active message (highlighted in light blue with a triangle on the left side) with commands such as "Move down #" (where "#" is between 1 and 10) to skip to the desired message. He can then say "Open message" to view the message body, and "What does it say" in order to activate the Windows 7 voice synthesizer, which reads the emails to him when he is too tired or his eyes are not focusing clearly. The system reads the subject, sender and body of the message and recognizes when to stop reading the message body when the signature or quoted text is reached. Finally, he can reply to messages or choose from a pre-determined address book of contacts, all with further voice commands.

The email client automatically scans all attachments and includes them directly inline when displaying the message. This makes it easier for the user to view picture attachments without having to click or double click as with traditional email readers. It also detects links to sites such as YouTube and places large icons on the toolbar, allowing the user to easily navigate off to these external sites from the email client. New contacts are automatically added to the contact list simply when emails are received from a new individual. The system also automatically archives all picture attachments into a folder hierarchy so that the target user can replay slideshows of all these photos whenever he wants.

5. Results: Current usage

The speech-based email system has been used continuously by our target user since February 2012. Between February 2012 and June 2013, the system has handled 460 received messages and 210 sent messages. In peak weeks, he has sent 10 to 20 emails to his contacts. These usage statistics are noteworthy because our target user had *never sent emails without assistance before the creation of this system.* While the system is not perfect and the speech recognition sometimes falters, the benefits of email communication have made this system acceptable for our target user.

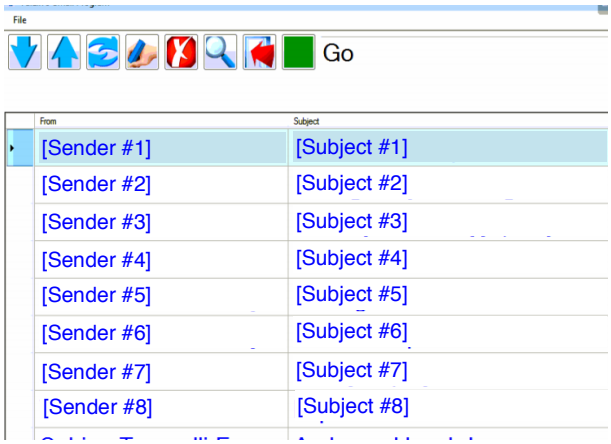

Figure 3: *User interface, with first message selected (senders and subjects anonymized). The upper right message box shows the top speech hypothesis.*

5.1. Observations on Usage

Through long-term user observations and unstructured interviews, we have learned about how our target user interacts with the system. Typically, he does not reply to every message, but rather replies once to every few messages from a given person so that the sender knows he has read the emails. This behavior is feasible because the user has a small group of contacts who appear to be sending him emails regularly.

It is worth noting that our target user still uses the telephone because it enables immediate, two-way communication. While we have not done formal monitoring, it appears that the email system has augmented, not replaced, his

telephone usage. He especially values messages with photo attachments of friends and family, which cannot be transmitted by telephone. He also receives many emails containing comic strips, jokes, and YouTube video links.

The target user has become adept at interpreting the user interface's visual cues and using these cues to adapt his behavior accordingly. For example, the hypothesized utterance and the large percentage confidence score are both displayed on the television screen. These cues allow him to see whether he needs to speak differently, adjust the microphone, reduce background noise, or report a bug.

6. Discussion

The process of developing the email system has yielded significant insights into developing customized speech-based assistive technology.

6.1. Factors for success

We believe that there were three main reasons why our target user has adopted the email system:

6.1.1. Design for a single user

Our approach focused intensely on our target user. Our success metric – and our singular goal while developing the system – was to enable him to communicate more frequently with friends and family. As a result, our work was tailored very specifically to the target user's abilities, preferences, environment, and feedback. Instead of focusing on an innovation that could potentially generalize across many users, our work deliberately was driven by our sole target user. Interestingly, it may be that some elements of system could be useful to other people, meaning that, in the process of seeking measurable impact on our target user, we have identified some generalizable components or ideas.

6.1.2. Multidisciplinary collaboration

Our team of authors has backgrounds in AT research, rehabilitation technology, speech-language pathology, speech recognition, and software development. In addition, some of our team members are staff or clinicians in the residential-care setting itself, which helped ensure that necessary issues or adjustments could be dealt with in a timely manner. The time and skills of each of these individuals were essential to the success of this project. The project would not have succeeded without any of the hardware and software components, readily available onsite support and physical care, and extensive speech therapy and training.

6.1.3. Frequent and long-term interaction with the user

The current system is the product of many years of interacting with our target user and learning from his AT usage patterns. For example, it is clear why the InVoca voice-activated remote control continues to be used: it is robust, requires little outside assistance or intervention to be operated, and enables him to watch television independently. In contrast, steep learning curves, reliability issues, and interface challenges made other speech technologies less appropriate. We considered these experiences as we developed the current system.

Perhaps more importantly, working with our target user over several years has allowed us to develop a working relationship that extends beyond simply being a research subject for new technologies. Whenever possible, we strived to incorporate his motivations, ideas, and direction, and we based our design decisions on in-home user observation. Such an approach may bear intrinsic value when working with people with disabilities, who often find mismatches between their abilities and existing technology. More directly, frequent communication and design iteration has helped us understand the subtleties that separate AT non-use from AT adoption.

6.2. Limitations

The purpose of this paper is to document the process leading to the development of a usable speech-based email client for our single target user. Our goal was not to develop a system that would necessarily work for other users. It may be the case that other users would find the limitations of our system unacceptable, or that their speech recognition error rates would be too high to use it successfully. Answering this question would only be possible with a study involving more users.

Clearly, the current system has limited functionality. The features that we did prioritize, though, made it possible for our target user to communicate with friends and family. Interestingly, through his extended usage, he has suggested feature ideas, including the ability to place pre-defined sets of sentences into emails for simple messages or pre-downloading attachments while he is sleeping so that emails load faster during the day. As a next step, our target user is interested in adding video calling capabilities. A separate grammar for Skype functions should make it possible to implement this feature without compromising speech recognition accuracy.

While actually deploying useful AT can be time-consuming and difficult, our efforts have helped us remain connected to the realities of users. Our work suggests relevant areas of inquiry for this user population, including the need to adapt acoustic models to speakers who may not be able to access a close-talking microphone, speech recognition that is robust to environmental noise in healthcare settings, and graphical user interfaces tailored to people who may have co-occurring vision or other impairments.

7. Conclusions

We described the implementation of a system that uses speech recognition to allow a single user to communicate via email with friends and family. The process of developing this assistive technology was made possible by embracing the target user's goals, focusing on a practical solution, learning from past devices and technologies, and drawing from our diverse professional backgrounds and skills. Building real-world, actual implementations of working assistive devices could help define worthwhile research efforts and illuminate the characteristics of successful assistive technology.

8. References

[1] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99-112, 2010.

[2] D.J. Ward et al, "Dasher – a data entry interface using continuous gestures and language models," in Proc. UIST, 2000, pp. 129-137.

[3] Principles and Practice of Assistive Technology (PPAT), Fall 2011. http://courses.csail.mit.edu/PPAT/fall2011.

[4] W. Li et al, "Probabilistic dialogue modeling for speech-enabled assistive technology," in Proc. SLPAT, 2013.