Evaluating Existing Audio CAPTCHAs and an Interface Optimized for Non-Visual Use

Jeffrey P. Bigham and Anna C. Cavender
Department of Computer Science and Engineering
DUB Group
University of Washington
Seattle, WA 98195 USA
{jbigham, cavender}@cs.washington.edu

ABSTRACT
Audio CAPTCHAs were introduced as an accessible alternative for those unable to use the more common visual CAPTCHAs, but anecdotal accounts have suggested that they may be more difficult to solve. This paper demonstrates in a large study of more than 150 participants that existing audio CAPTCHAs are clearly more difficult and time-consuming to complete as compared to visual CAPTCHAs for both blind and sighted users. In order to address this concern, we developed and evaluated a new interface for solving CAPTCHAs optimized for non-visual use that can be added in-place to existing audio CAPTCHAs. In a subsequent study, the optimized interface increased the success rate of blind participants by 59% on audio CAPTCHAs, illustrating a broadly applicable principle of accessible design: the most usable audio interfaces are often not direct translations of existing visual interfaces.

ACM Classification Keywords
K.4.2 Social Issues: Assistive technologies for persons with disabilities; H.5.2 Information Interfaces and Presentation: User Interfaces

General Terms
Human Factors, Design, Experimentation

Author Keywords
Audio CAPTCHA, Non-Visual Interfaces, Blind Users

INTRODUCTION AND MOTIVATION
The goal of a CAPTCHA[1] is to differentiate humans from automated agents by requesting the solution to a problem that is easy for humans but difficult for computers. CAPTCHAs are used to guard access to web resources and, therefore, prevent automated agents from abusing them. Current CAPTCHAs rely on superior human perception, leading to CAPTCHAs that are predominately visual and, therefore, unsolvable by people with vision impairments. Audio CAPTCHAs that rely instead on human audio perception were introduced as a non-visual alternative but are much more difficult for web users to solve. Part of the problem is that the interface has not been designed for non-visual use.

Most CAPTCHAs on the web today exhibit the following pattern: the *solver* is presented text that has been obfuscated in some way and is asked to type the original text into an answer box. The technique for obfuscation

[1]Completely Automated Public Turing test to tell Computers and Humans Apart

CHI 2009, April 49, 2009, Boston, Massachusetts, USA.

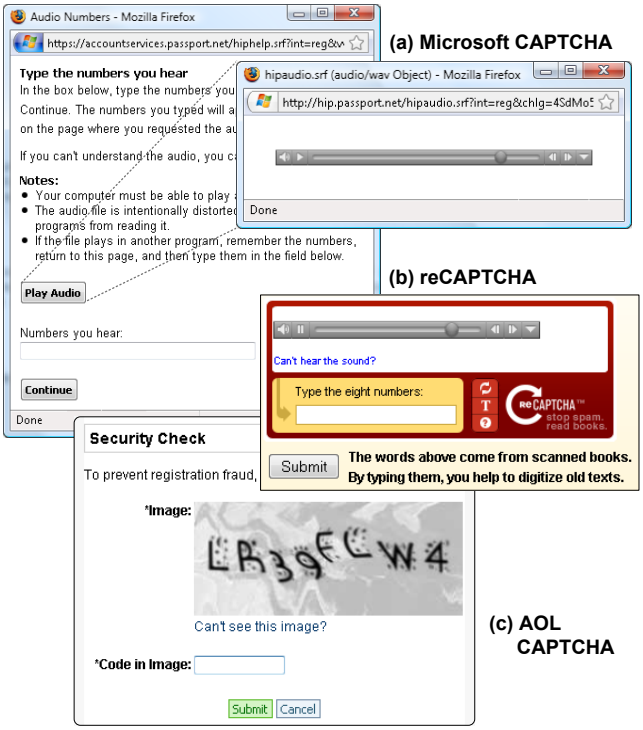

Figure 1. Examples of existing interfaces for solving audio CAPTCHAs. (a) A separate window containing the sound player opens to play the CAPTCHA, (b) the sound player is in the same window as the answer box but separate from the answer box, and (c) clicking a link plays the CAPTCHA. In all three interfaces, a button or link is pressed to play the audio CAPTCHA, and the answer is typed in a separate answer box.

is chosen such that it is difficult for automated agents to recover the original text but humans should be able to do so easily. Visually this most often means that graphic text is displayed with distorted characters (Figure 1). In audio CAPTCHAs, this often means text is synthesized and mixed in with background noise, such as music or unidentifiable chatter. Although the two types of CAPTCHAs seem roughly analogous, the usability of the two types of CAPTCHAs is quite different because of inherent differences in the interfaces used to perceive and answer them.

Visual CAPTCHAs are perceived as a whole and can be viewed even when focus is on the answer box. Once focusing the answer box, solvers can continue to look at visual CAPTCHAs, edit the answer that they provided, and verify their answer. They can repeat this process until satisfied without pressing any keys other than those that form their answer. Errors primarily arise from CAPTCHAs that are obfuscated too much or from careless solvers.

Audio playback is linear. A solver of an audio CAPTCHA first plays the CAPTCHA and then quickly focuses the answer box to provide their answer. For sighted solvers, focusing the answer box involves a single click of the mouse, but for blind solvers, focusing the answer box requires navigating with the keyboard using audio output from a screen reader. Solving audio CAPTCHAs is difficult, especially when using a screen reader.

Screen readers voice user interfaces that have been designed for visual display, enabling blind people to access and use standard computers. Screen readers often speak over playing CAPTCHAs as solvers navigate to the answer box, speaking the interface but also talking over the CAPTCHA. A playing CAPTCHA will not pause for solvers as they type their answer or deliberate about what they heard. Reviewing an audio CAPTCHA is cumbersome, often requiring the user to start again from the beginning, and replaying an audio CAPTCHA requires solvers to navigate away from the answer box in order to access the controls of the audio player. The interface to audio CAPTCHAs was not designed for helping blind users solve them non-visually.

Audio CAPTCHAs have been shown previously to be difficult for blind web users. Sauer *et al.* found that six blind participants had a success rate of only 46% in solving the audio version of the popular reCAPTCHA [18], and Bigham *et al.* observed that none of the fifteen blind high school students in an introductory programming class were able to solve the audio CAPTCHA guarding a web service required for the course [3]. In this paper, we present a study with 89 blind web users who achieved only a 43% success rate in solving 10 popular audio CAPTCHAs. On many websites, unsuccessful solvers must try again on a new CAPTCHA with no guarantee of success on subsequent attempts, a frustrating and often time-consuming experience.

Given its limitations, audio may be an inappropriate modality for CAPTCHAs. Developing CAPTCHAs that require human intelligence that computers do not yet have seems an ideal alternative, but the development of such CAPTCHAs has proven elusive [7]. CAPTCHAs cannot be drawn from a fixed set of questions and answers because doing so would make them easily solvable by computers. Computers are quite good at the math and logic questions that can be generated automatically. Audio CAPTCHAs could also be made more understandable, but that could also make them easier for computers to solve automatically.

The new interface that we developed improves usability without changing the underlying audio CAPTCHAs. By moving the interface for controlling playback directly into the answer box, a change in focus (and thus a change in context) is not required. Using the new interface, solvers have localized access to playback controls without the need to navigate from the answer box to the playback controls. Solvers also do not need to memorize the CAPTCHA, hurry to navigate to the answer box after starting playback of the CAPTCHA, or solve the CAPTCHA while their screen readers are talking over it. Solvers can play the CAPTCHA without triggering their screen readers to speak, type their answer as they go, pause to think or correct what they have typed, and rewind to review - all from within the answer box.

Because popular audio CAPTCHAs have similarities in their interfaces, our optimized interface can easily be used in place of these existing interfaces. Both the ideas and interface itself are likely to be applicable to CAPTCHAs yet to be developed. Finally, the design considerations explored here have application to improving a wide range of interfaces for non-visual access.

This paper offers the following four contributions:

- A study of 162 blind and sighted web users showing that popular audio CAPTCHAs are much more difficult than their visual counterparts.
- An improved interface for solving audio CAPTCHAs optimized for non-visual use that moves the controls for playback into the answer box.
- A study of the optimized interface indicating that it increases the success rate of blind web users on popular CAPTCHAs by 59% without altering the underlying CAPTCHAs.
- An illustration via the optimized interface that usable interfaces for non-visual access should not be directly adapted from their visual alternatives without considering differences in non-visual access.

RELATED WORK

CAPTCHAs were developed in order to control access to online resources and prevent access by automated agents that may seek to abuse these resources [22]. As

their popularity increased, so did the concern that the CAPTCHAs used were primarily based on the superiority of human visual perception, and therefore excluded blind web users. Although audio CAPTCHAs were introduced as an accessible alternative, the interface used to solve them did not consider the lessons of prior work on optimizing interfaces for non-visual use.

Making CAPTCHAs Accessible

Audio CAPTCHAs were introduced soon after their visual alternatives [22, 9], and have been slowly adopted by web sites using visual CAPTCHAs since that time. Although the adoption of audio CAPTCHAs has been slower than that of visual CAPTCHAs, many popular sites now include audio alternatives, including services offered by Google and Microsoft. Over 2600 web users have signed a petition asking for Yahoo to provide an accessible alternative [25]. The reCAPTCHA project, a popular, centralized CAPTCHA service with the goal of improving the automated OCR (Optical Character Recognition) processing of books also provides an audio alternative. Although audio CAPTCHAs exist, their usability has not been adequately examined.

Researchers have quantified the difficulty that users have solving both audio and visual CAPTCHAs. For instance, Kumar *et al.* explored the solvability of visual CAPTCHAs while varying their difficulty on several dimensions [6]. Studies on audio CAPTCHAs have been smaller but informative. For instance, Sauer *et al.* conducted a small usability study (N=6) in order to evaluate the effectiveness of the reCAPTCHA audio CAPTCHA [18]. They noted that participants in the study employed a variety of strategies for solving audio CAPTCHAs. Four participants memorized the characters as they were being read and then entered them into the answer box after the CAPTCHA had finished playing and one participant used a separate note taking device to record the CAPTCHA characters as they were read. They noted that the process of solving this audio CAPTCHA was highly error-prone, resulting in only a 46% success rate. The study presented in the next section expands these results to a diverse selection of popular CAPTCHAs in use today and further illustrates the frustration and strategies that blind web users employ to solve audio CAPTCHAs.

The usability of CAPTCHAs for human users must be achieved while maintaining the inability of automated agents to solve them. Although visual CAPTCHAs have had the highest profile in attempts to break them, audio CAPTCHAs have recently faced similar attempts [20]. As audio CAPTCHAs are increasingly made the target of automated attacks, changes that make them easier to understand will be less likely to be adopted out of concern that they will make automated attacks easier as well. Changing the interface used to solve a CAPTCHA, however, only impacts the usability for human solvers.

The audio CAPTCHAs described earlier are currently the most popular type of accessible CAPTCHA, but they are not the only approach pursued. Holman *et al.* developed a CAPTCHA that pairs pictures with the sounds that they make (for instance, a dog is paired with a barking sound) so that either the visual or audio representation can be used to identify the subject of the CAPTCHA [8]. Tam *et al.* proposed phrased-based CAPTCHAs that could be more obfuscated than current audio CAPTCHAs but remain easy for humans to solve because human solvers will be able to rely on context [20]. The improvements provided by our optimized interface to audio CAPTCHAs could be adapted to both of these new approaches should they be shown to be better alternatives.

Other Alternatives

Because audio CAPTCHAs remain difficult to use and are not offered on many web sites, several alternatives have been developed supporting access for blind web users. Many sites require blind web users to call or email someone to gain access. This can be slow and detracts from the instant gratification afforded to sighted users. The WebVisum Firefox extension enables web users to submit requests for CAPTCHAs to be solved, which are then forwarded to their system to be solved by a combination of automated and manual techniques [24]. Because of the potential for abuse, the system is currently offered by invitation only and questions remain about its long-term effectiveness. For many blind web users the best solution continues to be asking a sighted person for assistance when required to solve a visual CAPTCHA.

Combinations of (i) new approaches to creating audio CAPTCHA problems and (ii) interfaces targeting non-visual use promise to enable blind web users to independently solve CAPTCHAs in the future. This paper demonstrates the importance of the interface.

Targeting Non-Visual Access

The interface that we developed for solving audio CAPTCHAs builds on work considering the development of non-visual interfaces. Such interfaces are often very different than the interfaces developed for visual use even though they enable equivalent interaction. For instance, in the math domain, specialized interfaces have been developed to make navigation of complex mathematics feasible in the linear space exposed by non-visual interfaces [16]. Emacspeak explores the usability improvement resulting from applications designed for non-visual access instead of being adapted from visual interfaces [17].

With the increasing importance of web content, much work has targeted better non-visual web access. For instance, the HearSay browser converts web pages into semantically-meaningful trees [15] and, in some circumstances, automatically directs users to content in a web page that is likely to be interesting to them [11]. Trail-

Blazer suggests paths through web content for users to follow, helping them avoid slow linear searches through content [5]. A common theme in work targeting web accessibility is that content should be accessed in a semantically meaningful way and functionality should be easily available from the context in which it most makes sense.

The aiBrowser for multimedia web content enables users to independently control the volume of their screen reader and multimedia content on the web pages they view [12]. Without the interface provided by aiBrowser, content on a web page can begin making noise (for instance, playing a song in an embedded sound player or Flash movie) making screen readers difficult to hear. This audio clutter can make navigating to the controls of the multimedia content using a screen reader difficult, if controls are provided for the multimedia content at all. One of the goals of our optimized interface to audio CAPTCHAs is to prevent CAPTCHAs from starting to play before the user is in the answer field where they will type their answers - a major complaint of our study participants concerning how audio CAPTCHAs work currently. Just as with the aiBrowser, the goal is, in part, to give users finer control over the audio channel used by both their screen readers and other applications.

Work in accessibility has also explored the difference between accessibility and usability. Many web sites are technically accessible to screen reader users, but they are inefficient and time-consuming access. Prior work has shown that the addition of heading elements to semantically break up a web page or the use of skip links to enable users to quickly skip to the main content of a page can increase its usability [21, 23]. Audio CAPTCHAs are accessible non-visually, but their usability is quite poor for most blind web users. Our new interface helps to improve usability.

EVALUATION OF EXISTING CAPTCHAS

Many web services now offer audio CAPTCHAs because they believe them to be an accessible alternative to visual CAPTCHAs. However, the accessibility and usability of these audio CAPTCHAs has not been extensively evaluated. Our initial study aims to evaluate the accessibility of existing audio CAPTCHAs and search for implications we could use to improve them. We did this by gathering currently used CAPTCHAs from the most popular web services and presented them to study participants to solve. During the study, we collected tracking data to investigate the means by which both sighted and blind users solve CAPTCHAs. The tracking data we collected allowed us to analyze the timing (from page load to submit) of every key pressed and button clicked, and search for problem areas and possible improvements to existing CAPTCHAs.

Existing Audio CAPTCHAs

To gather existing audio CAPTCHAs for our study, we used Alexa [1], a web tracking and statistic gathering service, to determine the most popular web sites visited from the United States as of July 2008. Of the top 100, 38 used some form of CAPTCHA, and of those less than half (47%) had an audio CAPTCHA alternative. For our study, we chose to only include sites offering both visual and audio CAPTCHAs and avoided sites using the same third party CAPTCHA services.

Features of Audio CAPTCHAs

	AOL	Auth-orize	Craigslist	DIGG	Face-book	Google	MS-Live	PayPal	Slash-dot	Veoh
Assistance Offered	no	no	yes	no	no	no	no	no	yes	no
Beeps Before	3	0	0	0	1	3	0	0	0	1
Background Noise	voice	none	music	static	voice	voice	voice	static	none	voice
Challenge Alphabet	A-Z 0-9	A-Z 0-9	A-Z	A-Z	0-9	0-9	0-9	A-Z	Word	0-9
Duration (sec)	10.2	5.1	9.3	6.9	24.7	40.9	7.1	4.3	3.0	25.1
Repeat	no	no	no	no	no	no	no	no	yes	no

Figure 2. A summary of the features of the CAPTCHAs that we gathered. Audio CAPTCHAs varied primarily along the several common dimensions shown here.

Using this method we chose 10 unique types of CAPTCHAs that represent those used by today's most popular websites: AOL (aol), Authorize.net payment gateway service provider (authorize), craigslist.org online classifieds (craigslist), Digg content sharing forum (digg), Facebook social utility (facebook), Google (google), Microsoft Windows Live individual web services and software products (mslive), PayPal e-commerce site (paypal), Slashdot technology-related news website (slashdot), and Veoh Internet television service (veoh). For each of the 10 CAPTCHA types we downloaded 10 examples, resulting in a total of 100 audio CAPTCHAs used for the study (Figure 2).

Several of these sites attempted to block the download of the audio files representing each CAPTCHA although all of them were in either the MP3 or WAV format. Many sites added the audio files to web pages using obfuscated Javascript and would allow each to be downloaded only once. These techniques at best marginally improve security, but can often hinder access to users who may want to play the audio CAPTCHA with a separate interface that is easier for them to use.

Study Description

To conduct our study, we created interfaces for solving visual and audio CAPTCHAs mimicking those we observed on existing web pages (Figure 3). The interface for visual CAPTCHAs consisted of the CAPTCHA image, an answer field, and a submit button. The interface for solving audio CAPTCHAs replaced the image with a play button that when pressed caused the audio CAPTCHA to play. These simplified interfaces preserve the necessary components of the CAPTCHA interface, enabling interface components to be isolated from the surrounding content. Solving CAPTCHAs in real web

Figure 3. An interface to solving audio CAPTCHAs modeled after those currently provided to users to solve audio CAPTCHAs (Figure 1).

pages may be more difficult as there are additional distractions, such as other content, and the CAPTCHA may need to be solved with a less ideal interface, for instance using a pop-up window.

Our study was conducted remotely. As Petrie *et al.* observed, conducting studies with disabled people in a lab setting can be difficult, but remote studies can produce similar results [13]. Blind users in particular use many different screen readers and settings that would be difficult to replicate fully in a lab setting, meaning the remote studies can better approximate the true performance of participants.

Participants were first presented with a questionnaire asking about their experience with web browsing, experience with CAPTCHAs and the level of difficulty or frustration they present, as well as demographic information. They were then asked to solve 10 visual CAPTCHAs and 10 audio CAPTCHAs (for sighted participants) or 10 audio CAPTCHAs (for blind participants). Each participant was asked to solve one problem randomly drawn from each CAPTCHA type, and the CAPTCHA types were presented in random order to help avoid ordering effects.

For this study, participants were designated as belonging to the blind or sighted condition based on their response to the question: "How do you access the web?" The following answers were provided as options: "I am blind and use a screen reader," "I am sighted and use a visual browser," and "Other." In this paper, blind participates will refer to those who answered with the first option and sighted participants to those who answered with the second option.

Participants were given up to 3 chances to correctly solve each CAPTCHA, but of primary concern was their ability to correctly solve each CAPTCHA on the first try because this is what is required by most existing CAPTCHAs.

To instrument our study, we included Javascript tracking code on each page of the study that allowed us to keep track of the keys users typed and other interaction with page elements. This approach is similar to that provided by the more general UsaProxy [2] system which records all user actions in the browser when users connect through its proxy. This approach has also been used before in studies with screen reader users [4].

The data recorded enabled us to make observations, including the time required to answer the CAPTCHA, how many times the CAPTCHA was played, how many mistakes were made in the process of answering a CAPTCHA, and the number of attempts required. The full list of the events gathered and the information recorded for each is shown below:

- **Page Loaded** - the web page has loaded.
- **Focused Play** - participant selected the play button.
- **Pressed Play** - participant pressed the play button.
- **Blurred Play** - participant moved away from the play button.
- **Answer Box Focused** - participant entered the answer box either by clicking on it or tabing to it.
- **Answer Box Blurred** - participant exited the answer box either by clicking out or moving away.
- **Key Pressed** - participant pressed a keyboard key.
- **Focused Submit** - submit button was selected.
- **Pressed Submit** - submit button was pressed.
- **Blurred Submit** - participant moved away from the submit button without pressing it.
- **Incorrect Answer** - the answer provided by the participant is incorrect, leading the participant to be presented with a 2nd or 3rd try.

Personally identifying information was not recorded.

Results

Of our 162 participants, 89 were blind and 73 were sighted; 56 were female, 99 were male, and 7 chose not to answer that question; and their ages ranged from 18 to 69 with an average age of 38.0 ($SD = 13.2$).

Before participating in our study, blind and sighted participants showed differing levels of frustration toward the audio and visual CAPTCHAs they had already come across. Participants were asked to rate the following questions on a scale from *Strongly Agree* (1) to *Strongly Disagree* (5) or opt out by answering "I have never independently solved a visual[audio] CAPTCHA" for the following questions: "Audio CAPTCHAs are frustrating to solve." and "Visual CAPTCHAs are frustrating to solve."

For the question about audio CAPTCHAs, averages from the two groups were similar, 2.73 ($SD = 1.3$) for blind participants and 2.82 ($SD = 1.4$) for sighted participants. Far more sighted participants opted out; however, as only 7.87% of blind participants opted out compared to 44.44% of sighted participants who opted out ($\chi^2 = 69.13$, $N = 161$, $df = 1$, $p < .0001$). This shows that nearly half of our sighted participants had never solved an audio CAPTCHA before, but those who had

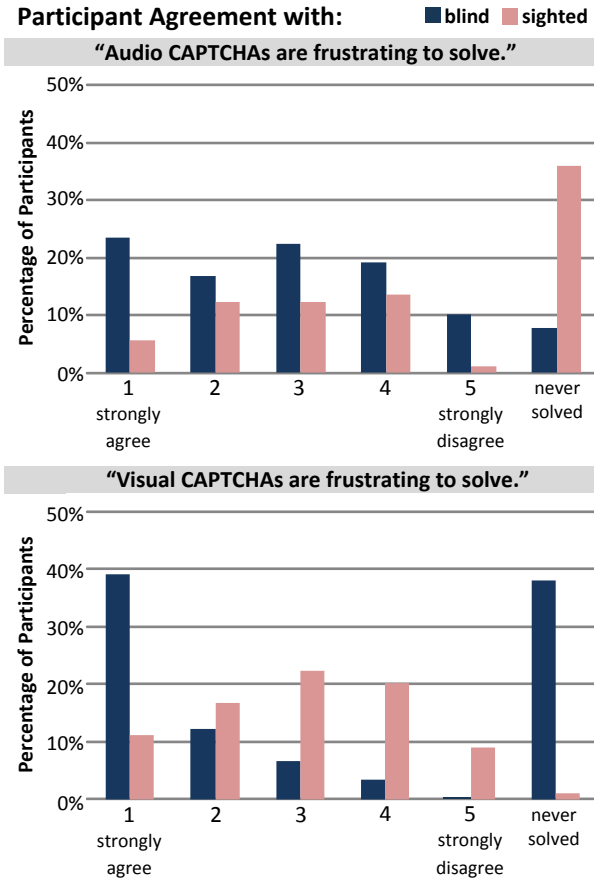

Figure 4. Percentage of participants answering each value on a Likert scale from 1 *Strongly Agree* to 5 *Strongly Disagree* reflecting perceived frustration of blind and sighted participants in solving audio and visual CAPTCHAs. Participants could also respond "I have never independently solved a visual[audio] CAPTCHA." Results illustrate that (i) nearly half of sighted and blind participants had not solved an audio or visual CAPTCHA, respectively, (ii) visual CAPTCHAs are a great source of frustration for blind participants, and (iii) audio CAPTCHAs are also somewhat frustrating to solve.

were nearly as frustrated by them as blind participants. For the question about visual CAPTCHAs, blind participants averaged 1.58 ($SD = 0.9$) with 38.2% opting out and sighted participants averaged 2.98 ($SD = 1.2$) with only 1.4% opting out ($\chi^2 = 14.21$, $N = 161$, $df = 1$, $p = .0002$). This shows that more than a third of blind participants said they had never solved a visual CAPTCHA and the others found them very frustrating with a rating very close to (1) *Strongly Agree*. This rating may mean that some of our participants who checked the "I am blind and use a screen reader" box did have some vision and had tried to solve visual CAPTCHAs before or perhaps some participants found the required phone call to technical support, the added step of waiting for an email, or the task of finding a sighted person for help to be extremely frustrating. These results are summarized in Figure 4.

The data gathered from the Javascript tracking code were analyzed using a mixed-effects model analysis of

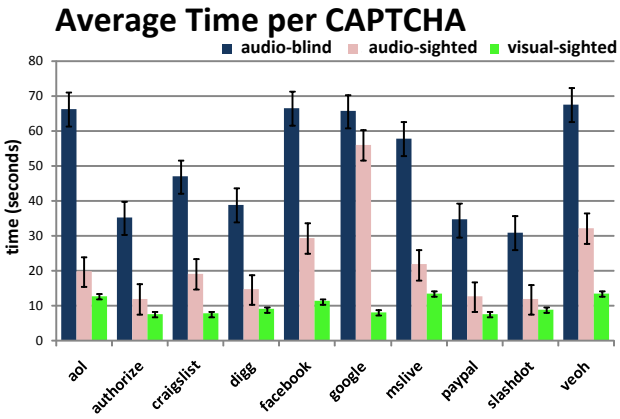

Figure 5. The average time spent by blind and sighted users to submit their first solution to the ten audio CAPTCHAs presented to them. Error bars represent ± 1 standard error (SE).

variance with repeated measures [10, 19]. Condition (blind or sighted), CAPTCHA type (audio or visual), and CAPTCHA source, were modeled as fixed effects, with Condition and CAPTCHA type combined as a fixed effect group with three possible values (blind-audio, sighted-audio, and sighted-visual). Participant was modeled correctly as a random effect. Mixed-effects models properly handle the imbalance in our data due to not all participants solving both audio and visual CAPTCHAs. Mixed-effects models also account for correlated measurements within participants. However, they retain large denominator degrees of freedom, which can be fractional for unbalanced data.

Sighted participants solving visual CAPTCHAs were much faster than blind participants solving audio CAPTCHAs. On average, their respective completion times were more than 5 times faster. Sighted participants averaged 9.9 seconds ($SD = 1.9$) and blind participants averaged 50.9 seconds ($SD = 1.8$), ($F_{1,232.1} = 243.9$, $p < .0001$). This may have been expected, but sighted participants also outperformed blind participants on audio CAPTCHAs with average completion times of 22.8 ($SD = 1.9$), or about twice as fast as our blind participants ($F_{1,232.4} = 113.9.0$, $p < .0001$). The timing data alone show the drastic inequalities in current CAPTCHAs for blind web users (Figure 5).

The largest differences were observed in success rates. The sighted participants in this study successfully solved nearly 80% of the visual CAPTCHAs presented to them (on the first try). This resembles the 90% previously reported [6][2]. These same participants, however, were only able to solve 39% of audio CAPTCHAs on the first try, demonstrating again the higher difficulty of solving audio CAPTCHAs. And while it did take blind participants longer (see above), blind and sighted participants

[2]The lower observed success rate may reflect the trend of CAPTCHAs having become more difficult in order to thwart increasingly-sophisticated automated attacks.

were on par when it came to solving the audio CAPTCHAs *correctly*. Blind participants solved 43% of audio CAPTCHAs presented to them successfully on the first try, although the difference between blind and sighted was not significant ($\chi^2 = 3.46$, $N = 161$, $df = 1$, $p = .06$). Second and third tries rarely helped in finding a correct answer (Figure 6).

Even though blind participants were on par (slightly better, but not significantly so) at solving audio CAPTCHAs correctly, they took twice as long to do so. So, what occupied the remaining time? This extra time may have been spent listening to the CAPTCHA (on average, blind participants clicked played 3.6 ($SD = 0.1$) times whereas sighted participants clicked play 2.5 ($SD = 0.1$) times ($F_{1,232.1} = 52.2$, $p < .0001$)) or they may have spent more time navigating to and from the text box. Blind participants entered the text box on average 2.9 ($SD = .1$) times whereas sighted participants entered the text box on average 2.4 ($SD = 0.1$) times ($F_{1,232.2} = 10.2$, $p < .001$).

Discussion

Recruiting participants to take part in studies can be especially difficult when looking for participants with specific characteristics, such as participants who use a screen reader. Despite this, we had very little trouble recruiting participants for this study (as reflected by the large number of responses). Our post on an online mailing list for blind web users was greeted with a flurry of responses - both positive and negative. Many seemed pleased to find that this problem was being worked on and many doubted that audio CAPTCHAs could ever be improved. Our first anecdotal evidence that CAPTCHAs were a widely-acknowledged problem were the number of responses, many of which were written with what appeared to be significant emotion.

Audio CAPTCHAs were anecdotally a great source of frustration to both blind and sighted participants in our study. Many sighted participants had no prior experience with audio CAPTCHAs and told us that they were much more difficult than they expected. In fact, one participants said, "After going through this exercise, I've changed my opinion that audio CAPTCHA is a good alternative solution for people who are blind." Many participants, but perhaps especially the blind participants, expressed exacerbation toward CAPTCHAs: "I understand the necessity for CAPTCHAs, but they are the only obstacle on the Internet I have been unable to overcome independently."

Clearly, some types of audio CAPTCHAs are much more difficult to solve than others and some features were better received than others. For example, "The random-letters, random-numbers ones were completely impossible for me to solve. I couldn't tell the difference between c/t/v/b, for example. Those with human-intelligible context (e.g. 'c as in cucumber') were far easier and less stressful."

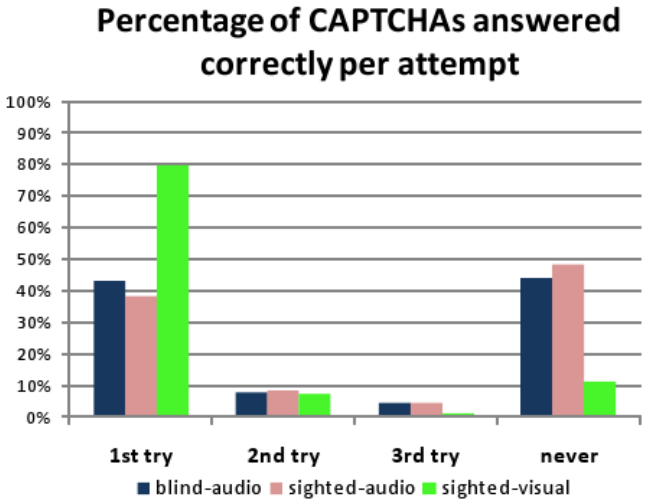

Figure 6. The number of tries required to correctly answer each CAPTCHA problem illustrating that (i) multiple tries resulted in relatively few corrections, (ii) the success rates of blind and sighted solvers were on par, and (iii) many audio CAPTCHAs remained unsolved after three tries.

While some of the frustration from solving CAPTCHAs seemed to stem from the difficulty of deciphering distorted audio, for blind people, much of the frustration comes from interacting with the CAPTCHA with their screen reader. For example, "It will always be hard to activate the play button, jump to the answer edit box, silence a screen reader and get focused to listen and enter data accurately." This process takes time and often content in the beginning of the CAPTCHA is missed: "At the beginning of the captcha, give me time to get down to the edit box and enter it. My screen reader is chattering while I'm getting to the edit box and the captcha is playing."

Instead of trying to navigate while the CAPTCHA plays, some people try to memorize the answer, wait for the play to finish, and then move to the text box and start typing. But, this presents an entirely new challenge: "I heard them, but could not remember them. And if I tried to type them out [while] listening, my screen reader interfered with my listening." This process resembles what one might expect sighted users to do if the visual CAPTCHA and the answer box were located on different pages and only one could be viewed at a time.

The types of interaction problems discovered as part of this study motivate a new interface design with simple improvements that could greatly increase the usability of audio CAPTCHAs

IMPROVED INTERFACE FOR NON-VISUAL USE

The comments of participants identified two main areas in which audio CAPTCHAs could be improved. As expected, one area was the audio itself – the speech representation should be made clearer, background noise

Figure 7. The new interface developed to better support solving audio CAPTCHAs. The interface is combined within the answer textbox to give users control of CAPTCHA playback from within the element in which they will type the answer.

reduced, and additional contextual hints provided in order to make audio CAPTCHAs easier to solve. The audio characteristics of a CAPTCHA were important in determining its difficulty but are difficult to change because they directly determine how resistant the CAPTCHA will be to automated attacks. Audio CAPTCHAs have recently become a more popular target for automated attacks, for example reCAPTCHA was shown likely to be vulnerable to automated attack [20].

The second area of difficulty mentioned by participants was the interface provided for solving audio CAPTCHAs. Users found the current interfaces cumbersome and sometimes confusing to use. Unlike many improvements to the CAPTCHA itself, improvements to the interface do not affect the resistance to automated attacks. As long as the interface does not embed clues to the answer of the CAPTCHA, then it can be modified in whatever way is best for users.

The navigation elements used to listen to and answer an audio CAPTCHA can be distracting, forcing users to either miss the beginning of the CAPTCHA or memorize the entire CAPTCHA before typing the answer. Participants reported that they appreciated CAPTCHAs that began with a few beeps (as 4 of the 10 CAPTCHAs did) because this allowed them time to move from the "Play" button to the answer box. This suggested that a more usable interface would not require users to navigate back and forth. Our interface optimized for non-visual use addresses this navigation problem by moving the controls for playback into the answer box, obviating the need to navigate from playback controls to the answer box because they are now one in the same.

By combining the playback controls and the answer box into a single control, the interface is designed to present less of a hurdle for users to overcome, enabling them to focus on answering the CAPTCHA. Many participants mentioned that using the current interface required them to play through an entire audio CAPTCHA to review a specific portion. Even when controls other than "play" are available, users do not use them because they require them to navigate to the appropriate control and then back again to the answer box. Based on this feedback, we added simple controls into the answer box that enabled users to both play/pause the CAPTCHA and to rewind or fast-forward by one second without additional navigation (Figure 7).

Through several rounds of integrative design with several blind participants, we refined this new interface. For example, we initially used various control key combinations to control playback of the CAPTCHA (such as CTRL+P for play), we found that the shortcuts that we chose often overlapped with shortcuts available in screen readers. We briefly considered using the single key "p" for play, but this overlaps with the alphabet used in many popular CAPTCHAs meaning our interface could not be used with them.

On the suggestion of a blind participant, we chose to use the following individual keys for the playback controls: comma(,) for rewind, period(.) for play/pause, and forward slash(/) for fast-forward. These were not included in the alphabets of any of the CAPTCHAs that we considered (Figure 2) and are located in that order in standard American keyboards. For users of keyboards with different layouts, the keys could be similarly chosen to avoid collision with screen reader shortcuts and characters used in language-specific CAPTCHAs, and such that they are conveniently located on popular local keyboards.

Integration Into Existing Websites

An advantage of altering the interface used to solve CAPTCHAs instead of attempting to make CAPTCHA problems themselves more usable is that a new interface can be independently added to existing web sites. We have written a Greasemonkey script [14] that detects the reCAPTCHA interface and replaces the interface used to solve its audio CAPTCHA with our optimized interface.

For web sites in which this is not currently possible, web developers could add this interface into their sites without concern that the new interface will expose them to additional risk of automated attack. All of the currently-used CAPTCHAs considered in the study in the previous section can be used directly with our optimized interface.

EVALUATION OF NEW CAPTCHA INTERFACE

We evaluated our new interface for solving audio CAPTCHAs with the optimizations for screen reader users based on the insights from our initial study.

Study Design

To evaluate the new interface for audio CAPTCHAs we repeated the study described earlier but with the new interface. Below is a snippet from the instructions that were given to participants before the study began:

We are testing a different interface for solving CAPTCHAs. Please take some time to familiarize yourself with the new interface. Keys for controlling playback are as follows:

- *Typing a period in this box will cause the CAPTCHA to play, and pressing it again will pause playback*

- *Typing a comma will rewind the CAPTCHA by 1 second and then continue playing.*
- *Typing a forward slash will fast forward the CAPTCHA by 1 second and then continue playing.*

These keys work only when the textbox used to answer the CAPTCHA problem has focus. This allows you to control the CAPTCHA directly from the box into which you will enter your answer. The control key characters will not be entered into the box and are only used to control playback of the CAPTCHA.

No participants reported difficulty learning the new interface.

Results

This study included 14 blind participants: 2 were female, 10 were male, and 2 chose not to answer that question; their ages ranged from 22 to 59 with an average age of 36.1 ($SD = 10.2$).

We again used a mixed-effects model analysis of variance with repeated measures to analyze our data. While our optimized interface did not have a significant effect on the time required to solve CAPTCHAs (participants averaged 50.9 seconds ($SD = 2.4$) with the original interface and 47.3 seconds ($SD = 5.9$) with the optimized interface ($F_{1,101.3} = 0.31$, $p = n.s.$), it did have significant and positive effects on the number of tries required to solve CAPTCHAs and the observed success rate of participants.

With our optimized interface, participants were able to reduce the average number of attempts required to solve the CAPTCHAs from 2.21 ($SD = 0.4$) with the original interface to 1.56 ($SD = 1.2$) ($F_{1,100} = 20.3$, $p < .0001$). Perhaps more importantly, participants solved over 50% more CAPTCHAs correctly on the first try with the optimized interface than they did with the original interface: 42.9% ($SD = 0.2$) were correctly solved on the first try using the original interface and 68.5% ($SD = 0.5$) were correctly solved on the first try using the optimized interface ($F_{1,100} = 22.3$, $p < .0001$). These improvements can be seen in Figure 8.

Discussion

Participants in this study were generally enthusiastic about the new interface to audio CAPTCHAs that we created, leading one participant to say, "I really liked the interface provided here for answering the captchas. I think it could really be benefitial [sic] if widely used." Some participants felt that while the new interface offered an improvement, audio CAPTCHAs were still frustrating. For example, one participant said, "... sometimes the audio captchas are still so distorted that it's hard to solve them even then."

In general, while audio CAPTCHAs remained challenging for users, they were both more accurate with the new interface (answered incorrectly less often) and required fewer attempts to find the right answer. Because the new interface does not affect the security of the underlying CAPTCHA and can be easily adapted to new CAPTCHAs we hope this interface will become the default in the near future.

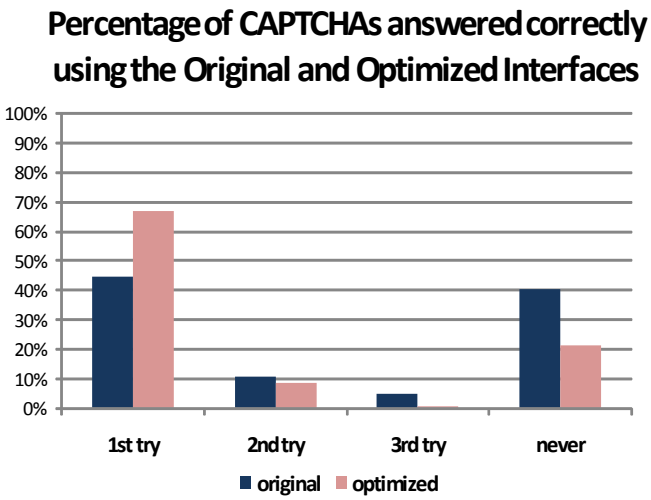

Figure 8. The percentage of CAPTCHAs answered correctly by blind participants using the original and optimized interfaces. The optimized interface enabled participants to answer 59% more CAPTCHAs correctly on their first try as compared to the original interface.

FUTURE WORK

In this work, we have demonstrated the difficulty of audio CAPTCHAs and offered improvements to the interface used to answer them that can help make them more usable. We plan to explore other areas in which interface changes may improve non-visual access, and consider how the lessons we learned in this work may generalize beyond the interfaces to audio CAPTCHAs.

Future work may explore how audio CAPTCHAs could be created that are easier for humans to solve while still addressing the improved automatic techniques for defeating them. The ten audio CAPTCHAs explored in our study exhibited a wide variety of dimensions on which they varied, but yet remained quite similar in design. Perceptual CAPTCHAs face many problems, including that (i) none are currently accessible to individuals who are both blind and deaf and (ii) automated techniques are becoming increasingly effective in defeating them. An important direction for future work is addressing these problems.

CONCLUSION

Creating an interface optimized for non-visual access presents challenges that are very different than those targeting visual access. Our study with blind participants demonstrated that existing audio CAPTCHAs are inadequate alternatives and that their frustration is due in part to the interface provided for solving them.

Based on this feedback, we optimized the interface to solving audio CAPTCHAs for non-visual use by localizing the playback interface to the answer box.

Although we did not change the audio CAPTCHAs themselves, users in our subsequent study were able to successfully solve CAPTCHAs on their first try 59% more of the time. This dramatic improvement can be directly used in existing interfaces to CAPTCHAs without impacting the ability of the CAPTCHA to protect access from automatic agents. Because of the incredible differences in non-visual access, the interface can make all the difference when developing applications designed to be accessed non-visually.

ACKNOWLEDGEMENTS

This work has been supported by National Science Foundation Grant IIS-0415273. We thank Jennison M. Asuncion, Lindsay Yazzolino, Sambhavi Chandrashekar, and Annuska Perkins for their comments and assistance in recruiting participants. We especially thank our anonymous participants for their insightful feedback.

REFERENCES

1. Alexa web search – data services, 2008. http://www.alexa.com.
2. R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move - user activity tracking for website usability evaluation and implicit interaction. In *Proc. of the 15th Intl. Conf. on World Wide Web (WWW '06)*, pages 203–212, New York, NY, 2006.
3. J. P. Bigham, M. B. Aller, J. T. Brudvik, J. O. Leung, L. A. Yazzolino, and R. Ladner. Inspiring blind high school students to pursue computer science with instant messenging chatbots. In *Proc. of the 39th SIGCSE Tech. Symp. on Computer Science Education (SIGCSE '08)*, pages 449–453, Portland, OR, USA, 2008.
4. J. P. Bigham, A. C. Cavender, J. T. Brudvik, J. O. Wobbrock, and R. E. Ladner. Webinsitu: A comparative analysis of blind and sighted browsing behavior. In *Proc. of the 9th Intl. ACM SIGACCESS Conf. on Computers and Accessibility (ASSETS '07)*, pages 51–58, Tempe, AZ, USA, 2008.
5. J. P. Bigham, T. Lau, and J. Nichols. TrailBlazer: Enabling blind web users to blaze trails through the web. In *Proc. of the 12th Intl. Conf. on Intelligent User Interfaces (IUI '09)*, Sanibel Island, FL, USA, 2009.
6. K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski. Designing human friendly human interaction proofs (HIPS). In *Proc. of Computer Human Interaction (CHI '05)*, pages 711–720, 2005.
7. P. B. Godfrey. Text-based captcha algorithms. In *First Workshop on Human Interactive Proofs*. Unpublished Manuscript, 2002. http://www.adaddin.cs.cmu.edu/hips/events/abs/godfreyb_abstract.pdf.
8. J. Holman, J. Lazar, J. H. Feng, and J. D'Arcy. Developing usable captchas for blind users. In *Proc. of the 9th Intl. ACM SIGACCESS Conf. on Computers and accessibility (ASSETS '07)*, pages 245–246, New York, NY, USA, 2007.
9. G. Kochanski, D. Lopresti, and C. Shih. A reverse turing test using speech. In *Proc. of the Intl. Conf. on Spoken Language Processing (ICSLP '02)*, pages 1357–1360, 2002.
10. R. C. Littell, G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. *SAS System for Mixed Models*. SAS Institute, Inc., Cary, North Carolina, USA, 1996.
11. J. Mahmud, Y. Borodin, and I. V. Ramakrishnan. CSurf: A context-driven non-visual web-browser. In *Proc. of the Intl. Conf. on the World Wide Web (WWW '07)*, pages 31–40.
12. H. Miyashita, D. Sato, H. Takagi, and C. Asakawa. AiBrowser for multimedia: introducing multimedia content accessibility for visually impaired users. In *Proc. of the 9th Intl. ACM SIGACCESS Conf. on Computers and accessibility (ASSETS '07)*, pages 91–98, New York, NY, USA, 2007.
13. H. Petrie, F. Hamilton, N. King, and P. Pavan. Remote usability evaluations with disabled people. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI '06)*, pages 1133–1141, New York, NY, USA, 2006.
14. M. Pilgrim, editor. *Greasemonkey Hacks: Tips & Tools for Remixing the Web with Firefox*. O'Reilly Media, 2005.
15. I. V. Ramakrishnan, A. Stent, and G. Yang. Hearsay: Enabling audio browsing on hypertext content. In *Proc. of the 13th Intl. Conf. on the World Wide Web (WWW '04)*, pages 80–89, New York, NY, USA 2004.
16. T. V. Raman. *Auditory User Interfaces: Toward the Speaking Computer*. Kluwer Academic Publishers, Boston, MA, 1997.
17. T. V. Raman. Emacspeak - A Speech Interface. In *Proc. of the SIGCCHI Conf. on Human Factors in Computer Systems (CHI '96)*, pages 66–71, Vancouver, BC, Canada, 1996.
18. G. Sauer, H. Hochheiser, J. Feng, and J. Lazar. Towards a universally usable captcha. In *Proc. of the 4th Symp. On Usable Privacy and Security (SOUPS '08)*, Pittsburgh, PA, USA, 2008.
19. C. Schuster and A. Von Eye. The relationship of ANOVA models with random effects and repeated measurement designs. *Journal of Adolescent Research*, 16(2):205–220, 2001.
20. J. Tam, J. Simsa, D. Huggins-Daines, L. von Ahn, and M. Blum. Improving audio captchas. In *Proc. of the 4th Symp. on Usability, Privacy and Security (SOUPS '08)*, Pittsburgh, PA, USA, July 2008.
21. J. Thatcher, P. Bohman, M. Burks, S. L. Henry, B. Regan, S. Swierenga, M. D. Urban, and C. D. Waddell. *Constructing Accessible Web Sites*. glasshaus Ltd., Birmingham, UK, 2002.
22. L. von Ahn, M. Blum, and J. Langford. Telling humans and computer apart automatically: How lazy cryptographers do AI. *Communications of the ACM*, 47(2):57–60, February 2004.
23. T. Watanabe. Experimental evaluation of usability and accessibility of heading elements. In *Proc. of the Intl. Cross-Disciplinary Conf. on Web Accessibility (W4A '07)*, pages 157 – 164, 2007.
24. WebVisum Firefox Extension, 2008. http://www.webvisum.com/.
25. Yahoo Accessibility Improvement Petition, 2008. http://www.petitiononline.com/yabvipma/.