POSTER: Assessing the Usability of the new Radio Clip-Based Human Interaction Proofs

Jonathan Lazar, Jinjuan Feng, Olusegun Adelegan, Anna Giller, Andrew Hardsock, Ron Horney, Ryan Jacob, Edward Kosiba, Gregory Martin, Monica Misterka, Ashley O'Connor, Andrew Präck, Roland Roberts, Gabe Piunti, Robert Schober, Matt Weatherholtz, Eric Weaver

Department of Computer and Information Sciences, Universal Usability Laboratory
Towson University
Towson, MD, 21252 USA
Contact: jlazar@towson.edu

Abstract

Human Interaction Proofs (HIPs) are widely adopted to protect websites and online accounts from the voracious attacks of automated programs. Unfortunately, HIPs have become one of the major accessibility obstacles for individuals with visual disabilities. Audio-based HIPs were developed to address the challenge. Previous studies suggest that those audio-based HIPs, although theoretically accessible for blind users, are very hard to use and have low task success rates. We conducted an empirical study to evaluate the efficiency and accuracy of the new radio-clip based CAPTCHA. Based on the data collected, blind users still find the audio CAPTCHAs hard to use.

1. INTRODUCTION

Human interaction proofs (HIP) have become a necessary part of everyday interaction on the web. A HIP is a security technique used to limit the amount of spam and bots. They are a form of "challenge", such as identifying distorted text or sound clips, which humans can typically understand, but computerized speech and image recognition cannot understand. While these various HIPs serve an important need, it is well documented that they are very hard to use. Typically, HIPs pose a great problem for blind users, and they have been cited as the greatest security-related problem for blind users [4]. One well-known example is the CAPTCHA (Completely Automated Public Turing tests to tell Computers and Humans Apart). The original CAPTCHAs were completely visual. A later version of the CAPTCHA added an audio version. Because a CAPTCHA typically is a gatekeeper, the usability of CAPTCHAs is of paramount importance. The goal of this paper is to evaluate the usability of the new audio reCAPTCHA.

A CAPTCHA consists of a series of letters, numbers, or a combination of both, that has been distorted. The earlier CAPTCHAs were only visual, but an audio version was later added for the sake of accessibility. The audio version used 8 numbers, spoken by different voices, and with distortion. While the distortion was necessary to avoid being easily recognized by speech recognition, the distortion certainly makes it harder for users to successfully complete the CAPTCHA. Other approaches to HIPs exist, such as HIPUU, which require users to recognize either non-textual sound clips or images, such as birds, rain, and pianos [6]. Previous usability evaluations of the CAPTCHA [1, 6] found that blind users had trouble using the audio CAPTCHAs, and generally had a task success rate below 50%. The goal for a usable HIP should be that a human should be successful at least 90% of the time, but a bot should only be successful .01% of the time [2]. Given that, the CAPTCHA product/approach is still the best known and most-used HIP in the world. However, there have been recent changes to the CAPTCHA approach used by the reCAPTCHA project. Instead of random numbers and letters, the reCAPTCHA project is now using old printed material and radio clips as source material. The newer visual approach to CAPTCHAs presents two words--one word, a control word, which is known by the reCAPTCHA engine, and another word which is not known by the reCAPTCHA engine and was not successfully interpreted by OCR [7]. The user must solve half of the visual CAPTCHA successfully--the half already known by the reCAPTCHA engine. The other half of the CAPTCHA is the user digitizing an unknown word from old books or other printed material. While there is no documentation from the reCAPTCHA project for the exact mechanism of the audio CAPTCHA, it is assumed that the approach is similar: a portion of the radio clip is understood by the reCAPTCHA engine (the "control text"), where the user is actually being tested, and the other portion of the radio clip was not previously understood. So the human is helping to digitize that portion of the radio clip. In general, with these newer CAPTCHAs, the user is only being tested on half of the visual text or audio clip.

2. RESEARCH METHODOLOGY

There is a well-established research methodology for evaluating the usability of CAPTCHAs. The user listens to the audio clip, attempts to type in the text that they heard, and then the CAPTCHA engine states whether the response was correct or not. In general, the previous CAPTCHAs, as well as the HIPUU, used an Levenshtein edit distance of two [3, 6]. An edit distance is the number of changes that would need to be made to the text for it to be correct. For instance, if the edit distance on the 8-number audio CAPTCHA was 2, then only 6 out of 8 characters would need to be correct for the response to actually be considered correct. However, on the new reCAPTCHA, edit distance is greater than two. In the visual reCAPTCHA, there are two words, one of which is the actual test for the human, and the other word is not the test. In the new audio reCAPTCHA, there are generally more words, possibly 6-10 words in a phrase. Informal evaluation has shown that only approximately half of the words are being tested. It is safe to say that the edit distance is more than two letters or two words.

Because the edit distance in the newer CAPTCHAs is much larger, there is a new challenge in doing usability evaluations: only approximately half of what the user types actually needs to be correct for the CAPTCHA to note the response as being correct. The CAPTCHA engine could say that a response is correct, even when it was nowhere close to being correct, since only approximately half of the words are being checked. Therefore, the classification of the user response by the CAPTCHA engine as being "correct" or "incorrect" is no longer accurate. Therefore, a new approach needs to be taken in performing usability evaluations of the audio CAPTCHAs.

In previous usability evaluations of the audio CAPTCHA [6], time performance was noted for each reCAPTCHA clip, and whether the participant response was classified by the system as "correct" was also noted. The methodology used in this study was identical to the methodology used in [6]. However, since the classification of "correct" or "incorrect" in the new audio reCAPTCHA is questionable, we modified the evaluation method, so that for each audio clip, it was noted in the data collection both the researcher's perception of correctness, and reCAPTCHA's perception of correctness. That is, the researchers listened to the audio clip, and noted whether what they heard matched up with what the user typed into the reCAPTCHA edit box. The researcher also recorded the reCAPTCHA engine response of "correct" or "incorrect."

3. RESULTS

40 individuals took part in the study. 10 of these individuals were blind, and 30 of these individuals had no documented disabilities. When testing accessibility features, it is often necessary to have both users with disabilities, as well as users without any disabilities, taking part [6]. If an accessibility feature degrades the experience for users without disabilities, it will not be adopted by developers. Demographics of the participants are listed in table 1. Table 2 displays the results of the data collection, alongside the data collected from two previously published studies. Due to the small sample size and the different approaches to define correctness rate, no statistical comparison was conducted.

	Age	Gender
Blind users	Average: 35.5, Stdev 9.55,	6 males, 4 females
Visual users	Average: 38.9, Stdev 15.9,	15 males, 15 females

Table 1. Demographic information of participants

Usability tests	Success rate	Time perf. Correct (seconds)	Time perf. incorrect (seconds)
Old version [1] *Blind users (n = 89)*	*43%*	*50.9*	*N/A*
Old version [1] *Sighted users (n = 89)*	*39%*	*22.8*	*N/A*
Old version [6] *Blind users (n = 6)*	*46.6%*	*65.64*	*59.56*
New version Blind users (n =10)	46%/ 60%	35.75	39.1
New version Visual users (n =30)	60.6%/ 70%	34.8	34.9

Table 2. Correctness rate and task completion time

4. DISCUSSION

Compared to previous studies, the task success rate for the audio CAPTCHA has not changed for blind users. In the two previous studies, it was 43% and 46%. With the new radio clip CAPTCHA, the task success rate was also 46%. However, the time that it took blind users to successfully complete an audio CAPTCHA task dropped from 50.9/65.64 seconds to only 35.75 seconds.

The comparison between the visual users and blind users is also interesting. With the previous audio CAPTCHA, blind users successfully completed the task in an average of 50-65 seconds, and visual users successfully completed the task in 22.8 seconds. With the new radio-clip CAPTCHA, the blind users and the visual users took approximately the same time to successfully complete the test: 35.75 for the blind users and 34.8 seconds for the visual users. This result suggests that blind users were able to successfully complete the task faster than before, but it took visual users more time than before. Of course, the visual users can choose to use the visual CAPTCHA instead, which blind users cannot do.

To summarize, compared to previous audio CAPTCHAs, the radio clip CAPTCHA improved the task success rate for visual users while not improving the task success rate for blind users. While the efficiency (time performance) for blind users improved, the efficiency for visual users deteriorated. This study provides valuable insights to the usability and accessibility of the radio-clip CAPTCHA. However, the comparison results should be interpreted with caution due to the different approaches used to measure correctness rate.

5. REFERENCES

[1] Bigham, J. P. and A. C. Cavender (2009). Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use. Proceedings of the 27th international conference on Human factors in computing systems. Boston, MA, USA, ACM, 1829-1838.

[2] Chellapilla, K., K. Larson, et al. (2005). Designing human friendly human interaction proofs (HIPs). Proceedings of the SIGCHI conference on Human factors in computing systems. Portland, Oregon, USA, ACM, 711-720.

[3] Gilleland, M. (2009). "Levenshtein Distance, in Three Flavors." Retrieved 19, March 2009, 2009, from http://www.merriampark.com/ld.htm.

[4] Holman, J., J. Lazar, et al. (2008). Investigating the Security-Related Challenges of Blind Users on the Web. Designing Inclusive Futures. P. Langdon, J. Clarkson and P. Robinson. London, Spring-Verlag**:** 129-138.

[5] Lazar, J. (Ed.) (2007). Universal Usability. Chichester, UK: John Wiley and Sons.

[6] Sauer, G., J. Holman, et al. (2010, in press). "Accessible Privacy and Security: A Universally Usable Human-Interaction Proof." Universal Access in the Information Society.

[7] Von Ahn,L., Maurer, B., McMillen, C., Abraham, D., Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321, Sept. 12, 2008, 1465-1468.