Behaviour & Information Technology, 2014
Vol. 33, No. 9, 919–928, http://dx.doi.org/10.1080/0144929X.2014.898793

Taylor & Francis
Taylor & Francis Group

Investigating the effects of sound masking on the use of audio CAPTCHAs

Abiodun Olalere[a*], Jinjuan Heidi Feng[a], Jonathan Lazar[a,b] and Tim Brooks[a]

[a]*Department of Computer & Information Sciences, Towson University, 8000 York Road, Towson, MD 21252, USA;* [b]*Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA 02138, USA*

(*Received 3 October 2013; accepted 19 February 2014*)

The SoundsRight Audio Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHA) was developed with the goal of providing a usable and secure audio CAPTCHA for people with visual impairments. Its design requires users to repeatedly identify a specific sound from a group of different sounds (e.g. baby crying and bird chirping) in real time. Adding background noise (sound masks) to the sounds may make it more difficult for automated software to recognise the sounds and therefore, improve security. However, the sound masks may also make it more challenging for human users to recognise the sound. We conducted a user study involving 20 blind participants and 20 sighted participants to investigate the effect of sound masks on the usability of the SoundsRight CAPTCHA. The results suggest that sound masks do have a significant impact on the failure rate and response time. Sighted participants had significantly a higher failure rate than blind participants and were more vulnerable to the negative effect of sound masks.

Keywords: usability; accessibility; security; CAPTCHA; audio; masking

1. Introduction

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) are an authentication mechanism, described as any test that can be automatically generated, which most humans can pass, but that current computer programmes cannot pass (von Ahn *et al.* 2004). CAPTCHAs presented in either visual or audio format have been widely used in websites to prevent Bots (Tam *et al.* 2009). People with visual impairments, however, are often confronted with the problem of solving visual CAPTCHAs as they interact with the Web. Studies have reported that CAPTCHAs pose a major problem to users who are blind or have low vision (May 2005, Edwards 2008, Holman *et al.* 2008, Schwartz 2011), because most websites only provide visual CAPTCHAs which are completely inaccessible to blind people. Audio CAPTCHAs, although accessible, are usually time consuming and hard to use (Chellapilla *et al.* 2005a, Bigham and Cavender 2009, Lazar 2009, Sauer *et al.* 2010b, Shirali-Shahreza and Shirali-Shahreza 2011).

Existing audio CAPTCHAs generally ask users to transcribe some spoken text, using a variety of techniques including distortion and multiple voices to defeat attacks via automated speech recognition systems. Previous research suggests that these audio CAPTCHAs are not easy to use, with successful task completion rates of less than 50% (Bigham and Cavender 2009, Sauer *et al.* 2010b). To address those problems, the SoundsRight Audio CAPTCHA was developed, and requires users to recognise a specific sound from a group of different sounds such as birds chirping or a piano playing, in real time (Lazar *et al.* 2012). While the SoundsRight Audio CAPTCHA proved to be easier to use than existing audio CAPTCHAs, with a task completion success rate above 90%, it has limitations regarding security. For example, it is potentially possible for sound recognition software to recognise the sound, especially when a plain sound is used without any distortion. Hence, a technique for sound distortion may be necessary to make the application more robust. However, when sound distortion is introduced, it may also make it harder for humans to recognise the sound. A good design has to consider the trade-off between security and usability and find a balance between the two.

This study examined the impact of sound masking on the use of the SoundsRight Audio CAPTCHA by both blind users and sighted users. Three types of sound masks (conversation, laughter, and orchestra) and three masking levels were investigated. The results help us understand the effect of sound masking on user performance, and can inform the design of usable and secure audio CAPTCHAs.

2. Literature review

The threats of website abuse from Bots, automated programmes, and scripts, demand effective security mechanisms with the ability to distinguish between computers and humans. Human interaction proofs (HIPs), also known

*Corresponding author. Email: aolale1@students.towson.edu

as CAPTCHA (von Ahn *et al.* 2004) are a software attempt at solving this problem. CAPTCHAs are any tests that can be automatically generated, which most humans can pass, but is difficult for computers to pass (von Ahn *et al.* 2004, Chellapilla *et al.* 2005b) even if the algorithms for such CAPTCHAs are made public (von Ahn *et al.* 2004).

Most visual CAPTCHAs are presented in the form of images with twisted or distorted texts (Javed *et al.* 2012) that users must correctly identify. According to Yan and El Ahmad (2008), all CAPTCHAs can be classified into three schemes: the text-based scheme, the sound (audio)-based scheme, and the image-based scheme. The text-based (visual) scheme makes use of sophisticated distortion of text images such that rendering becomes difficult for pattern recognition software to decode. The sound-based (audio) scheme requires users to solve specific speech/sound recognition tasks, while the image-based (visual) scheme requires users to perform image recognition tasks.

2.1. *Usability evaluation of CAPTCHAs*

Multiple studies (Mori and Malik 2003, Chellapilla and Simard 2004, Chellapilla *et al.* 2005b, Tam *et al.* 2009, Li *et al.* 2010, Goodin 2012) have been reported to prove the vulnerability of several Audio and Visual CAPTCHA mechanisms with success rates ranging between 47% and 99%. Those studies also made discoveries that help better understand human nature versus computers' with respect to the ease of solving a CAPTCHA as well as factors impacting the security of CAPTCHAs. For instance, Chellapilla and Simard (2004) found that humans are better at solving segmentation problems (e.g. locating or finding the position of a character in a CAPTCHA) than computers, but computers are better at recognition than human beings; and, using common distortion and clutter scenarios, computers are as good or better than humans at single character recognition.

Holman *et al.* (2008) and Sauer *et al.* (2010b) proposed an audio-visual CAPTCHA solution called Human Interaction Proof Universally Usable (HIPUU). The HIPUU 1.0 prototype presents a non-textual image and a corresponding (directly related) non-textual sound and users can utilise either the sound or the image. Usability evaluation showed that both blind and sighted participants achieved success rates of over 90%, and later versions of HIPUU included multiple challenges. Compared to existing studies on the task success rates of CAPTCHAs for blind users, HIPUU presented impressive results in the area of usability. However, the fact that HIPUU has no distortion in both the images and sound clips makes it vulnerable to automated attacks, and there is a limited solution set of corresponding images and sounds in pairs.

A usability evaluation of 10 different audio-visual CAPTCHAs was conducted by Bigham and Cavender (2009) with 89 blind users and 73 sighted users. Eighty per cent of the sighted individuals solved the visual CAPTCHAs on their first try, and the same sighted participants were only able to solve 39% of the audio CAPTCHAs, while 42.9% of the blind participants were able to solve the audio CAPTCHAs. The test was particularly difficult for blind participants because they are required to focus the cursor inside the textbox before typing the answers. Some participants tried to memorise all the answers as the audio CAPTCHA challenge played. While sighted participants were able to click inside the text box directly as the CAPTCHA was playing, blind participants had to wait till the end and then tabbed through the interface to locate the text box where they typed their answers. With a new interface designed to solve the typing problem, success rates improved from 42.9% to 68.5%. Essentially, the study focused on how the user interface impacts the success rates of users at solving CAPTCHAs, but does not address the problems caused by the distortion techniques that make both audio and visual CAPTCHAs difficult for many users to solve.

As pointed out by Bigham and Cavender (2009), interface accessibility is among the factors that affect users' success rate at solving CAPTCHAs. The SoundsRight Audio CAPTCHA solves the interface problems by presenting a real-time challenge that requires users to only respond to a specific sound by pressing the space bar each time that sound is played. This solution presents significant improvement in accessibility compared to other audio CAPTCHAs that require users to type spoken words or texts in order to solve the challenge.

2.2. *Security evaluation of CAPTCHAs*

Multiple research studies have documented the security weaknesses of audio CAPTCHAs (e.g. (Chellapilla *et al.* 2005a, Goodin 2012). Tam *et al.* (2009) found that humans are better at deciphering distorted utterances than computers, and suggested the use of phrases instead of random isolated words, a large vocabulary of audio challenges, and more sound distortions. Chandavale and Sapkal (2011) found that CAPTCHAs that contain the same speaker's voice, the same type of noise – especially noise distinct from human voice (e.g. running water) which would produce a different energy spike, and a finite vocabulary, would be easier for voice recognition software to detect. It is recommended that audio CAPTCHAs that contain a finite vocabulary, and background noise should have multiple speakers and noise similar to the speaker's voice, to make it harder for automated software to break them. In another study, Bursztein *et al.* (2011) deduced that a low (below 5) signal-to-noise ratio (SNR) would make the constant noise mask any spoken digits, and this could make the CAPTCHA unintelligible, but this would be easier to solve by humans, while computers would find it difficult to solve. He suggested that constant noise should only be used as background noise with a low SNR.

The successful cracking of Google's reCAPTCHA was said to be due to the fact that the background noise was

in sharp contrast to the six words in each challenge, and did not include sounds that registered at higher frequencies, making it easier to isolate each word by locating the regions where high pitches were mapped, when plotted on a spectrogram. The limited vocabulary of unique words used in reCAPTCHA was said to have a negative impact on security. Also having a negative impact was having fewer words (6 words per challenge) and shorter length (8 seconds challenge length) of a challenge. Therefore, Google has strengthened the security of reCAPTCHA by using a human voice uttering unintelligible sounds as background noise, making it difficult to isolate distinct words in each challenge. The puzzle has also been expanded from 6 words to 10 words in each challenge, and the challenge length increased from 8 to 30 seconds (Goodin 2012). It is however doubtful if the security improvement of reCAPTCHA translates into better user experience because earlier usability studies conducted on the audio reCAPTCHA before this latest security improvement showed that reCAPTCHA is not easy to use, as participants were only able to complete the test at a success rate of 46% (Sauer *et al.* 2010a), which falls far less than the 90% success rate suggested as appropriate for HIPs (Chellapilla *et al.* 2005a).

Distortion (in both visual and audio formats) is used to make it difficult for Bots to decipher CAPTCHA content, but human users would find it difficult to decipher over-distorted content. Even though users may have the opportunity for multiple attempts to be able to solve the CAPTCHA, it is often frustrating to users. While some efforts have been made at determining the appropriate distortion levels for visual CAPTCHAs, audio CAPTCHAs have not received the same attention (Yan and El Ahmad 2008).

2.3. Balancing security and usability

Since the goal of a CAPTCHA is to protect online resources from abuse by Bots, it is necessary for CAPTCHAs to be secure in order to achieve their goal. Security features, by their nature, are designed to disrupt the cognitive flow and impede the user from immediately completing their tasks. By the very nature of CAPTCHAs, the more secure they are, the less usable they become, and vice versa (Holman *et al.* 2008, Javed *et al.* 2012). Fidas *et al.* (2011) noted that the main purpose of a CAPTCHA is to protect the system and provider's resources, rather than to contribute to a positive user experience on interacting with the system. However, if CAPTCHAs are not usable by humans, then their essence, in fact, is defeated (Yan and El Ahmad 2008), and they are no longer a security feature, but simply a roadblock. As an example, there was a big news story when the CAPTCHA used on the US White House website was found to be impossible to use (not just hard) for blind people, and therefore, blind people were unable to sign a petition related to human rights for blind people (BBC 2013).

The trade-off between usability and security is inherent, and achieving a 'sweet spot', the spot in which the HIP is easily solvable by humans, but hard for computers to crack (Chellapilla *et al.* 2005a), is an on-going challenge. For example, one of the usability strengths of the HIPUU (Holman *et al.* 2008, Sauer *et al.* 2010b) is that there is no distortion, and therefore, the usability level is very high. This situation may have been unrealistic, from a security point of view. To address that challenge, Lazar *et al.* (2012) developed SoundsRight Audio CAPTCHA, which is usable with enhanced security due to the real-time nature of the CAPTCHA. In the Bigham and Cavender study (2009), the sound clips could be downloaded and replayed at the listener's preferred speed, which makes it possible for Bots to study and solve it. SoundsRight was designed so that it can only be listened to and solved in real time. Part of the security aspect of the SoundsRight CAPTCHA is that is takes every person (and any Bots that attempt it) approximately 45 seconds. You cannot complete it any faster or slower, therefore, Bots have no advantage, time-wise, over humans. Unlike the CAPTCHAs where there is an edit distance of two (Sauer *et al.* 2010b) meaning that errors are allowed, no errors are allowed with the SoundsRight CAPTCHA. If the user misses hitting the space bar on even one sound, they must start again with a different and new challenge set of 10 sound clips (Lazar *et al.* 2012).

Even though the SoundsRight Audio CAPTCHA achieved over 90% success rate for blind users, the use of plain sounds (without any form of distortion) may potentially hinder the security. Therefore, sound masking was added to the SoundsRight CAPTCHA and a user study was conducted to examine its impact on performance. The next section describes the design details of the application evaluated in the user study.

3. Application design

In this study, a challenge set used in a single test is made up of 10 sound clips (3 targets and 7 decoys), with target positions randomly selected. Users were not told about the number of targets or the position of target in each challenge set. A spoken delimiter ('. . . next sound.') is placed between two consecutive sound clips in the challenge set. The initial timestamp is taken each time a challenge starts and subsequent timestamps are recorded each time a user presses on the spacebar. When the sound clip is finished playing, the system returns the collected timestamps which are mathematically evaluated to determine the result. Each result (success/failure) is displayed to the user as shown in Figure 1.

After each challenge test is completed, the system displays a prompt asking the user to click the OK button to continue to the next test and the user is able to monitor the progress of the test from the test number updated and displayed as can be seen from Figure 1. Users are totally in control of the pace of transitions from one test to another

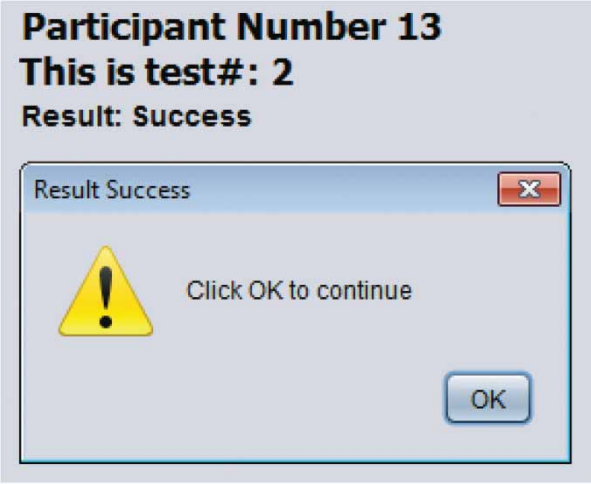

Figure 1. Test result and progress monitor.

(but not in control of the challenges within each specific test). This gives users the flexibility of pacing themselves, rather than being auto piloted. However, once a challenge test starts, users cannot stop until the challenge test is completed. At the end of every 10 challenge tests, the system specifically asks the user to take a break. The application was accessible for blind users to operate independently, utilising the JAWS (job access with speech) screen reader.

3.1. *Types of sound masking*

In order to evaluate the effect of sound masking, three types of sound masks were implemented in the application, specifically: conversation (conversations in a crowd), laughter (many people laughing), and orchestra (a crowd at an orchestra while the musicians are warming up – playing different audio instruments on stage) sounds. Each challenge set contains the same type of masking and different sounds (both target and decoys). During the test, the users need to differentiate three types of sounds – the target sound (which we ask the user to identify by pressing the space bar), the decoy sound (where the user should not press the space bar), and the masking sound (often known as 'distortion' or 'noise.').

3.2. *Depth of sound masking*

For this study, we define the depth of masking as the ratio between the volume of the target/decoy sound and the masking sound. From the security perspective, low masking depth (i.e. the target/decoy is the same volume or close in volume to the mask) is harder for Bots to solve due to the difficulty of isolating the target sounds, and therefore, is more secure (Bursztein *et al.* 2011). Three conditions were introduced to investigate the effect of the masking depth, namely 0, 2, and 4 dB. The volume of the mask remains constant at a normalised level of 0 dB for each challenge set. However, the target and decoy volume level change in relationship to the mask. For instance, if the challenge settings are such that the mask volume is at 10 dB, then the target and decoy sound volume will be 10 dB (i.e. add +0 dB), 12 dB (i.e. add +2 dB), and 14 dB (i.e. add +4 dB) under each condition. Note that 0 dB does not mean silence or the absence of sound, rather it is a normal threshold of hearing (Gelfand 2009). For our study, 0 dB is a comparative volume level. Audio normalisation was used to adjust the audio signal to a standard level (Beggs and Thede 2001).

4. Methods

The goal of our experiment is not to evaluate the security differences of the different masking levels, but rather to evaluate which masking depth levels and masking types have the least negative impact on usability.

4.1. *Participants*

A total of 40 participants, 20 blind and 20 sighted, took part in the study. All blind participants were recruited through the National Federation of the Blind, and for participation in the study, were required to be legally blind without any residual vision. All participants (blind and sighted) are required to be without any documented hearing loss or impairment. They must have lived in the USA for more than 10 years; this requirement was added because we found during the pilot study that people who have not lived long enough in the USA were not familiar with certain sounds in our sound clip set (e.g. the sound of a bagpipe). The educational levels of participants range from high school graduate to masters' degree level. For all the participants, we ensured that the test environments were quiet and without any form of noise or distraction and that participants were comfortably seated. The tests were taken by each participant independently on a laptop computer.

Of the blind participants, four were male and 16 were female (M = 10, SD = 8.49). On average, their age was 31.3 (SD = 13.11). They had used computers for an average of 16 years (SD = 7.91) and screen reader software for average of 12.65 years (SD = 7.24).

Of the sighted participants, 13 were male and seven were female (M = 10, SD = 4.24). The average age was 31.35 (SD = 13.92). They had used computers for 15.9 years on average (SD = 6.94).

4.2. *Procedure*

A within-group design was adopted and all participants completed tasks under four conditions: no noise masking, 0, 2, and 4 dB (0, 2, and 4 dB challenges were with audio mask). At the beginning of each test, the researchers explained the process, and answered questions from participants, if any. Next, the participant was asked to complete a demographic survey and perform two to three training tasks to get familiar with the application. After this, the main testing began. Each participant completed 10 challenge tests

independently under each of the four conditions, making a total of 40 challenge tests. Each test was randomly selected such that at the completion of 40 challenge tests, the user had completed 10 tests under each condition. After each 10 challenge tests, the system prompted the participant to take a break. We allowed the participants to take a break of up to 5 minutes, but some participants indicated that they were ready and wanted to start again sooner than 5 minutes. The order of the four conditions was balanced among the participants. The participants completed a satisfaction questionnaire at the end of the study.

4.3. Measurements

Each challenge had a fixed time (36 seconds long) and users must listen to the entire sound in each challenge set. Therefore, the time spent to complete each challenge was the same for all participants. The major measurement for performance was the failure rate and participants' subjective satisfaction ratings. We also collected data regarding the time it took to recognise the target sound (the time between the start of the target sound and when the participant clicked the space bar). However, the response time data are only available for sighted participants due to technical problems we experienced during the data collection.

4.4. Equipment

The test was conducted on two laptops. One was a Sony Vaio Laptop (Intel (R) Core(TM)2 Duo CPU 2 GHz and 4 GB RAM) running 64-bit Windows 7 Operating System and the other an IBM ThinkPad (1.8 GHz Core Duo) running 64-bit Windows 7 Operating System with JAWS version 11 (screen reader software) installed. While JAWS was installed, once the application began, the instructions were read using pre-recorded audio clips from the SoundsRight application. Only the progress markers (being tested now) and test results (passed or failed) were read by JAWS. The rest of the application was based on pre-recorded audio clips which were concatenated on the fly. The settings of JAWS were not modified by participants, with the exception of a few participants who modified the speed; however, since the sound clips were NOT spoken by JAWS, the speed of JAWS was irrelevant to the speed of the SoundsRight CAPTCHA challenges. Both laptops had built-in speakers which were used for the test.

4.5. Hypothesis

The goal of the study was to investigate the impact of masking on the use of SoundsRight CAPTCHA. The variables that we examined included the presence of visual impairment, the depth of masking, and the type of masking sound. We propose the following hypotheses for the study:

H1: Being blind or not does not have a significant impact on failure rates in solving Audio CAPTCHA.

H2: The presence of masking has no significant impact on failure rates in solving Audio CAPTCHA.

H3: The depth of masking has no significant impact on failure rates in solving Audio CAPTCHA.

H4: Mask types have no significant impact on failure rates in solving Audio CAPTCHA.

H5: The presence of masking does not have a significant impact on response time.

H6: The depth of masking has no significant impact on response time.

H7: Mask type does not have a significant impact on response time.

5. Results

5.1. Failure rate

5.1.1. Effect of disability

A factorial ANOVA test was conducted using the failure rate as the dependent variable. The two independent variables were the conditions of disability and the masking depth. The result suggests that both the conditions of disability and the masking depth had a significant impact on the failure rate ($F(1, 38) = 15.45$, $p < 0.001$; $F(3, 114) = 9.15$, $p < 0.001$).

Figure 2 demonstrates the average failure rates of both blind participants and sighted participants. The average failure rate for blind participants is 16.25%, with a minimum of 2.5% and maximum of 30%. The average failure rate for sighted participants is 30.75%, ranging from 12.5% to 60%. The failure rates of sighted participants are significantly higher than those of the blind participants. Therefore, hypothesis 1 is rejected.

5.1.2. Effect of masking depth

Figure 3 illustrates the failure rates of both blind participants and sighted participants under different conditions of masking depth.

Post-hoc Tukey's HSD test was used to determine whether the difference between the conditions is significant. The result shows that, for blind users, a significant difference exists between the following conditions:

- no masking condition and the 0 dB condition ($p < 0.05$),

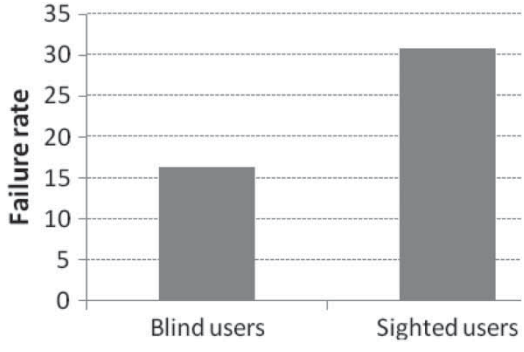

Figure 2. Average failure rates of blind participants and sighted participants.

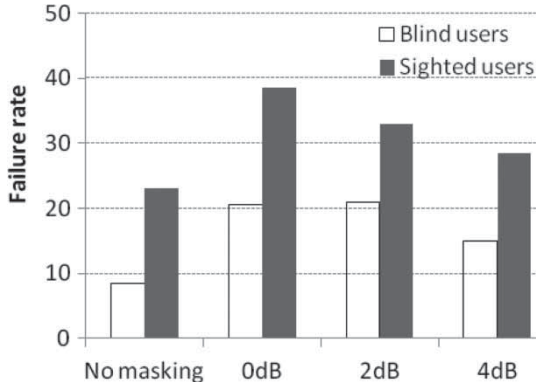

Figure 3. Average failure rates of blind participants and sighted participants under different conditions of masking depth.

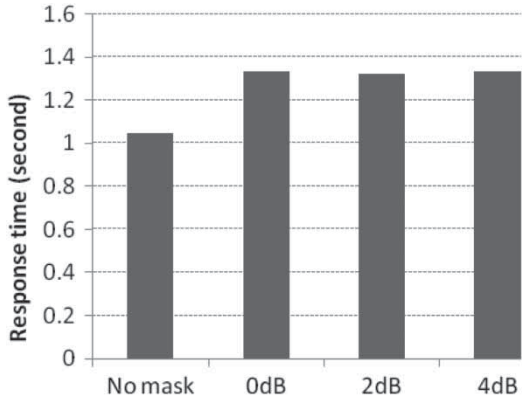

Figure 4. Average response time for sighted participants under the four conditions.

- no masking condition and the 2 dB condition ($p < 0.05$)

The failure rates under the 0 dB condition and the 2 dB condition are significantly higher than those of the no masking condition. There was no significant difference between any other conditions.

For sighted participants, one-tail Tukey's HSD test suggests a significant difference between no masking and the 0 dB ($p < 0.05$). The failure rate under the 0 dB condition is significantly higher than that of the no masking condition. There was no significant difference between any other conditions.

These results suggest that presenting noise masking does increase failure rates. In addition, the depth of the masking does affect the failure rate as well. When the depth of masking is larger (the 4 dB condition for blind users and the 2 and 4 dB conditions for sighted users), the masking did not have a significant impact on the failure rate. However, when the depth of masking is smaller (the 0 and 2 dB conditions for blind users and the 0 dB condition for sighted users), participants had higher failure rates. Therefore, both H2 and H3 were rejected.

5.1.3. Effect of masking type

A Repeated Measures test of ANOVA was conducted using the failure rate as the dependent variable and masking depth and masking type as the independent variables. No significant difference was found between the three types of masking for either the blind users ($F(2, 36) = 2.44$, n.s.) or the sighted users ($F(2, 34) = 2.33$, n.s.). Therefore, H4 is supported.

5.2. Response time

Response time is the time it took for the user to identify a sound. This measure provides insights on how easy it is for the user to recognise the sound. We only captured the response time data for sighted users due to technical problems. A Repeated Measures test of ANOVA was conducted using response time as the dependent variable and masking depth as the independent variable. The result suggests that masking depth has a significant impact on response time ($F(3, 57) = 27.24$, $p < 0.05$).

Figure 4 illustrates the average response time under the four conditions. A post-hoc Tukey's HSD test suggests that there is a significant difference in response time between the following conditions:

- no masking condition and the 0 dB condition ($p < 0.05$)
- no masking condition and the 2 dB condition ($p < 0.05$)
- no masking condition and the 4 dB condition ($p < 0.05$)

The response time under the 0 dB condition, the 2 dB condition, and the 4 dB condition is significantly longer than that of the no masking condition. There was no significant difference between the three masking depths. Therefore, H5 is rejected and H6 is supported.

A Repeated Measures test of ANOVA was conducted using response time as the dependent variable and masking depth and masking type as the independent variables. No significant difference was found between the three types of masking ($F(2, 38) = 2.14$, n.s.). Therefore, H7 is supported.

5.3. Subjective user satisfaction rating

Tables 1–3 summarise the subjective rating of both blind and sighted participants regarding the sound recognition, the general ease of use, and the willingness to adopt sound masking.

We asked both blind and sighted participants how easy it was to recognise the target sound in challenge sets with and without masking. Out of 20 blind participants, 18 participants strongly agreed, and 2 agreed that the target sounds in the challenge tests without masking were easy to recognise. Also, out of 20 blind participants, 5 participants strongly

Table 1. Subjective user experience on ease of sound recognition with andwithout masks.

		Participants rating				
		1	2	3	4	5
Blind participants	No mask	18	2	0	0	0
	With mask	5	7	7	1	0
Sighted participants	No mask	15	4	1	0	0
	With mask	0	11	7	1	1

Note: 1 represents strongly agree, and 5 represents strongly disagree.

Table 2. Subjective user experience on ease of use of SoundsRight CAPTCHA with and without masks.

		Participants rating				
		1	2	3	4	5
Blind participants	No mask	19	1	0	0	0
	With mask	9	4	3	2	2
Sighted participants	No mask	8	9	3	0	0
	With mask	0	9	7	4	0

Note: 1 represents strongly agree, and 5 represents strongly disagree.

Table 3. Subjective user experience on the question – 'I don't mind sound mask in audio CAPTCHA if it is for security'.

	Participants rating				
	1	2	3	4	5
Blind participants	11	4	4	1	0
Sighted participants	14	4	2	0	0

Note: 1 represents strongly agree, and 5 represents strongly disagree.

agreed, 7 agreed, 7 were neutral, and 1 participant disagreed that the target sounds with masking were easy to recognise. Out of 20 sighted participants, 15 participants strongly agreed, 4 participants agreed, and 1 participant was neutral that the challenge tests without masking were easier to recognise. On ease of target sound recognition in challenge tests with masking, out of 20 sighted participants, none of the participants strongly agreed, 11 participants agreed, 7 participants were neutral, 1 participant disagreed and 1 participant strongly disagreed. This result shows that participants (blind and sighted) obviously prefer solving the CAPTCHA without mask. The data pattern also shows that blind participants may tolerate solving CAPTCHAs with mask compared to sighted participants.

Table 2 shows the subjective user (blind and sighted) experience on ease of use of SoundsRight CAPTCHA with and without masks compared to other CAPTCHAs (audio or visual CAPTCHAs) they have previously used. Out of 20 blind participants, 19 participants strongly agreed that the challenge tests without masking are easier to use while 1 participant disagreed. Also, out of 20 blind participants, 9 participants strongly agreed, 4 agreed, 3 were neutral, 2 disagreed, and 2 strongly disagreed that the challenge tests with masking were easier to use.

For sighted participants, 8 participants strongly agreed that the challenge tests without masking were easier to use, 9 participants agreed, 3 participants were neutral, while none of the participants disagreed nor strongly disagreed. On challenge tests with masking, out of 20 sighted participants, none of the participants strongly agreed that the challenge tests with masking were easier to use, 9 participants agreed, 7 participants were neutral, and 4 participants disagreed. The results for each group are indicative of their experiences with CAPTCHAs. While most (18 out of 20) blind participants strongly agree that SoundsRight CAPTCHA (without mask) is easier to use compared to previous audio CAPTCHAs they have used, less than half (8 out of 20) of sighted participants strongly agree that the CAPTCHA is easier to use compared to previous visual CAPTCHAs they have used. This may be due to the fact that there are more visual CAPTCHAs than audio CAPTCHAs that sighted participants are exposed to. As a matter of fact, research in audio CAPTCHA is limited compared to visual CAPTCHA (Lazar *et al.* 2012) and most audio CAPTCHAs available are not very usable (Bigham and Cavender 2009). The pattern of responses observed in both groups (blind and sighted) under no masking condition is similar to that obtained under no masking condition in Table 1, and this clearly suggests that both groups prefer the CAPTCHAs without mask.

We also asked if participants would not mind solving CAPTCHA with sound mask if the sound mask is for security purpose. Table 3 shows the summary of users' responses. Out of 20 blind participants, 11 participants strongly agreed, 4 participants agreed, 4 were neutral while 1 participant disagreed. For sighted participants, out of 20 participants, 14 strongly agreed, 4 agreed, and 2 were neutral. This suggests that the majority of the participants are willing to pay the price of inconvenience cost by the masking in exchange for higher security. It also suggests that users should be informed of the necessity of adding sound masks.

5.4. Participants' comments

We asked participants if there are features they dislike or design changes they would suggest. Most blind participants would prefer not to use a sound mask, especially on low pitch sounds, and in situations where sound mask is used, they would prefer the target sound volume (target and decoy) to be much higher than the mask volume so that they can hear the sounds clearly. Sighted participants also did not like sounds with mask, and many complained that

they could not hear the target sound they expected to hear for certain target sounds. For instance, if the challenge test asks them to identify the sound of a pig, they expected to hear 'oink oink', but the pig sound used in the challenge sounded differently, with some base effect. Some sighted participants complained about being easily distracted or their minds wandering away during the test. It was also reported by sighted participants that sometimes they got 'trigger happy' or wrongly pressed the spacebar out of reflex during the test, which resulted in failures.

Other suggestions from participants for design changes include:

- play the target sounds before each challenge starts so that users can know what to listen for,
- allow more time to press spacebar in answer to a challenge test,
- provide the opportunity to pause and replay a sound during a test.

While some of these suggestions are reasonable, many cannot be implemented for security reasons. For instance, if a target sound is played before the start of each challenge, it may be easier for Bots to isolate the target sound from the mask and solve the challenge. Also, allowing too much time for users to identify the sound weakens the real-time argument of the application and creates opportunity for Bots in its solution attempt. Such practices may defeat the goal of CAPTCHA.

6. Discussion

6.1. *Impact of masking on usability*

The first stage of the SoundsRight Audio CAPTCHA study (Lazar *et al.* 2012) was conducted with only plain sounds (no masking), and the task failure rate was less than 10%. This study included plain sound (no masking) along with three mask types (conversation, laughter, and orchestra). The failure rate achieved by blind participants for the challenge tests with no masking is similar (8.5%) to the result previously obtained in Lazar *et al.* (2012). This is also the first study to evaluate the effectiveness of the SoundsRight CAPTCHA with sighted participants, and while the failure rate was higher for sighted participants, for challenge tests without masking, the success rate was still above 75%.

Our data show a significant difference in the failure rate between the challenge sets with masking and those with no masking in both groups (blind and sighted). The post-hoc Tukey's HSD test result between no masking and masking conditions for blind participants shows a significant difference in failure rates between no masking and 0 dB ($p < 0.05$), and no masking and 2 dB conditions ($p < 0.05$). No significant difference was found between no masking and 4 dB. A similar result is obtained in sighted participants where a one-tail Tukey's HSD test suggests a significant difference between no masking and the 0 dB ($p = 0.05$) condition and no difference between other conditions.

These findings suggest that the presence of masking does negatively impact the success rate, and the larger the distance between the mask volume and the challenge sound (decoys and target) volume, the higher the success rate, and vice versa. As the volume of the challenge sounds significantly increases over the sound mask volume, the effect of the sound mask fades out, making user sound recognition easier. When large depth of masking is provided, adding masking may not have any negative impact on the success rate.

For sighted users, the presence of mask also negatively affected the time it took to recognise the target sound. Even though the total time spent to solve the challenge remains the same, longer time spent to recognise the target sound suggests that the presence of mask may possibly cause higher cognitive load from the user.

Our results also show that mask types (conversation, laughter, and orchestra) have no significant impact on the failure rate in solving audio CAPTCHA for both the blind and sighted participants. This shows that it is the presence and comparative volume of sound mask, irrespective of the mask type, that significantly impacts the participants' failure rate in solving the audio CAPTCHA.

6.2. *Impact of masking on security*

One of the goals of adding masking was to improve the security of the SoundsRight Audio CAPTCHA. The previous stage of this study (Lazar *et al.* 2012) uses plain sounds (without masking). The addition of masking to the audio challenge in this study is considered a potential security improvement, without which sound recognition software may easily solve the audio challenge. The background (sound masks) used in this study fill a wide range of audio spectrum and are expected to be difficult to filter out without impacting the target sounds as well. The use of masking for each challenge set has increased the complexity of solving the puzzle compared to the first stage of this study that uses only 20 different plain sounds (without masking).

In an event where a Bot attempts to solve the CAPTCHA, the first task is to determine the target sound, and as the audio set is being analysed, the Bot will need to determine what background sound (mask) was used. This will be followed by an attempt to apply an algorithm to filter out the target sound, using some pattern matching to determine the target sound in the challenge set. Since the entire challenge set is masked, determining what target sound to look for, as well as matching the correct target, is expected to be difficult for Bots, compared to if plain sounds (without) masking are used. Simply put, the masking decreases the likelihood that a Bot will be able to solve it.

In production, the complexity can be further increased by increasing the number of sounds for each sound category and mask library. For instance, having about 20 different

sounds of birds and doing the same for other sound categories, and also increasing the library of background sounds (masks).

6.3. Difference between blind users and sighted users

Performance results between the blind and sighted participants are quite different. The difference provides interesting insights into sound recognition and the nature of user interaction in the two groups of users. Sighted participants in general had substantially a higher overall failure rate than the blind users (30% vs. 16%). Sighted participants were also more vulnerable to the negative impact of sound masking. While the 0 dB depth of masking (same volume level for both mask and target/decoy) does not affect the failure rate for blind users compared to 2 dB depth, it significantly increased the failure rate for sighted users compared to 2 dB depth.

The higher failure rates recorded with sighted participants might mean that blind users have more experience with audio output from computers. The reason for the higher failure rate may also be attributed to the comments captured in the subjective user experience survey. Sighted participants stated that they were often distracted through mind wonderings and they sometimes got 'trigger happy,' pressing the space bar even when they did not intend to do so. Whereas, blind participants did not seem to have problems with concentration, as their comments were more focused on sound attributes. This could be seen as similar to the findings of Lazar *et al.* (2007), which was that blind people were more efficient in responding to errors and frustrations on the web, as compared to visual people, because their coping strategies (e.g. looking for workarounds instead of rebooting like visual people) were more effective. Blind people may simply have more experience in listening carefully for auditory cues.

6.4. Future research

In this study, we kept the mask volume constant at a normalised level while adjusting the challenge set (decoys and target) volume. It would be interesting to examine the effect of masking the other way round; that is, keeping the challenge set volume at a normalised level while adjusting the mask volume. Also, instead of using only one mask type for an entire challenge set, it would be useful to examine the impact of masking when different sound masks are used for each sound clip in a challenge set. This is expected to increase the complexity of solving the challenge by Bots, but humans are expected to be able to solve the puzzle despite the complexity introduced by different masks selected at random.

7. Conclusion

This study explored the impact of masking on solving audio CAPTCHAs. The study was conducted with three types of masking (laughter, conversation, and orchestra sounds), and three depths of masking (0, 2, and 4 dB). A total of 40 participants (20 sighted and 20 blind users) participated in the usability testing. The evaluation results showed that low mask depths (when challenge volume is close to mask volume) decrease usability. The presence of mask in a challenge test can have a significant impact on usability, but the type of mask used does not seem to have a significant impact on usability. Furthermore, blind participants have an overall higher success rate in solving audio CAPTCHAs as compared to sighted participants. The results are useful in determining the appropriate mask and volume mix for the next stage of this research on improving the security of the SoundsRight Audio CAPTCHA. Even though audio CAPTCHAs are primarily used by people with visual impairments, it is useful to collect benchmark data on how the failure rates of blind participants compare with sighted participants, which can provide insights into the design of a universally accessible audio CAPTCHA.

Acknowledgements

We would like to thank the National Federation of the Blind for their assistance in recruiting participants for the study. We also want to thank all the participants.

References

von Ahn, L., Blum, M., and Langford, J., 2004. Telling humans and computers apart automatically. *Communications of the ACM*, 47 (2), 56–60.

BBC, 2013. *Blind Federation criticises Captcha security test* [Online]. Available from: http://www.bbc.co.uk/news/technology-22754006 [Accessed 31 July 2013].

Beggs, J. and Thede, D., 2001. *Designing web audio*. Sebastopol, CA: O'Reilly & Associates.

Bigham, J.P. and Cavender, A.C., 2009. Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use. *27th international conference on human factors in computing systems (CHI 2009)*, Boston, MA, 1829–1838.

Bursztein, E., *et al*., 2011. The failure of noise-based noncontinuous audio Captchas. *2011 IEEE symposium on security and privacy*, 22–25 May. New York: ACM, 19–31.

Chandavale, A. and Sapkal, A., 2011. An improved adaptive noise reduction for secured Captcha. *2011 fourth international conference of emerging trends in engineering & technology*, 18–20 November. Washington, DC: IEEE Computer Society, 12–17.

Chellapilla, K. and Simard, P.Y., 2004. Using machine learning to break visual human interaction proofs. *In*: L.K. Saul, Y. Weiss, and L. Bottou, eds. *Neural information processing systems (NIPS)*, vol. 17. Cambridge, MA: MIT Press, 265–272.

Chellapilla, K., *et al.*, 2005a. Designing human friendly human interaction proofs (HIPs). *ACM SIGGHI conference on human factors in computing systems*, vol. 1. New York: ACM, 711–720.

Chellapilla, K., *et al.*, 2005b. Computers beat humans at single character recognition in reading-based human interaction proofs (HIPs). *Second conference on email and anti-spam (CEAS)*, 2–2 July. Palo Alto, CA: Stanford University.

Edwards, J., 2008. *Beyond CAPTCHA: no bots allowed!* [Online]. Available from: http://www.sitepoint.com/captcha-problems-alternatives/ [Accessed 20 May 2012].

Fidas, C.A., Voyiatzis, A.G., and Avouris, N.M., 2011. On the necessity of user-friendly CAPTCHA. *2011 annual conference on human factors in computing systems (CHI '11)*. New York: ACM, 2623–2626.

Gelfand, S.A., 2009. *Essentials of audiology*. New York: Thieme Medical.

Goodin, D., 2012. *How a trio of hackers brought Google's reCAPTCHA to its knees* [Online]. Available from: http://arstechnica.com/security/2012/05/google-recaptcha-brought-to-its-knees/ [Accessed 11 June 2012].

Holman, J., Lazar, J., and Feng, J., 2008. Investigating the security-related challenges of blind users on the web. *In*: P. Langdon, J. Clarkson, and P. Robinson, eds., *Designing inclusive futures*. London: Springer, 129–138.

Javed, Y., Nazir, M., and Li, S., 2012. Captchæcker: reconfigurable CAPTCHAs based on automated security and usability analysis. *4th symposium on configuration analytics and automation (SafeConfig 2011)*, 31 October–1 November. Arlington, VA: IEEE.

Lazar, J., 2009. *Making CAPTCHA more accessible for the blind* [Online]. Available from: http://nfb.org/images/nfb/publications/bm/bm09/bm0901/bm090108.htm [Accessed 20 May 2012].

Lazar, J., *et al*., 2007. What frustrates screen reader users on the web: a study of 100 blind users. *International Journal of Human Computer Studies*, 22 (3), 247–269.

Lazar, J., *et al.*, 2012. The SoundsRight CAPTCHA: an improved approach to audio human interaction proofs for blind users. *SIGCHI conference on human factors in computing systems*. New York: CHI, 2267–2276.

Li, S., *et al.*, 2010. Breaking e-banking CAPTCHAs. *26th annual computer security applications conference (ACSAC '10)*. New York: ACM, 171–180.

May, M., 2005. *Inaccessibility of CAPTCHA* [Online]. Available from: http://www.w3.org/TR/turingtest/ [Accessed 23 February 2012].

Mori, G. and Malik, J., 2003. Recognizing objects in adversarial clutter: breaking a visual CAPTCHA. *Computer vision and pattern recognition (CVPR) conference*, 18–20 June. Washington, DC: IEEE Computer Society, I-134–I-141.

Sauer, G., *et al.*, 2010a. Accessible privacy and security: a universally usable human-interaction proof. *Universal Access in the Information Society*, 9 (3), 239–248.

Sauer, G., *et al.*, 2010b. Towards a universally usable human interaction proof: evaluation of task completion strategies. *ACM Transactions on Accessible Computing*, 2, 15: 11–15:32.

Schwartz, M.J., 2011. *Audio Captchas easy to defeat* [Online]. InformationWeek Available from: http://www.informationweek.com/news/security/vulnerabilities/229625482 [Accessed 20 May 2012].

Shirali-Shahreza, S. and Shirali-Shahreza, M.H., 2011. Accessibility of CAPTCHA methods. *AISec'11*, 21 October. New York: ACM, 109–110.

Tam, J., *et al.*, 2009. Breaking audio CAPTCHAs. *22nd annual conference on advances in neural information processing systems (NIPS)* 2008, Cambridge, MA: MIT Press, 1625–1632.

Yan, J. and El Ahmad, A.S., 2008. Usability of CAPTCHAs or usability issues in CAPTCHA design. *Symposium on usable privacy and security (SOUPS)*. New York: ACM, 44–52.