On the Viability of CAPTCHAs for Use in Telephony Systems: A Usability Field Study [*]

Niharika Sachdeva[†], Nitesh Saxena[*], Ponnurangam Kumaraguru[†]
[†] IIIT-Delhi, [*] University of Alabama at Birmingham
niharikas@iiitd.ac.in, saxena@cis.uab.edu, pk@iiitd.ac.in

Abstract

Telephony systems are imperative for information exchange offering low cost services and reachability to millions of customers. They have not only benefited legitimate users but have also opened up a convenient communication medium for spammers. Voice spam is often encountered on telephony systems in various forms, such as by means of an automated telemarketing call asking to call a number to win a reward. A large percentage of voice spam is generated through automated system which introduces the classical challenge of distinguishing machines from humans on telephony systems. CAPTCHA is a conventional solution deployed on the web to address this problem. Audio-based CAPTCHAs have been proposed as a solution to curb voice spam. In this paper, we conducted a field study with 90 participants in order to answer two primary research questions: quantifying the amount of inconvenience telephony-based CAPTCHA may cause to users, and how various features of the CAPTCHA, such as duration and size, influence usability of telephony-based CAPTCHA. Our results suggest that currently proposed CAPTCHAs are far from usable, with very low solving accuracies, high solving times and poor overall user experience. We provide certain guidelines that may help improve existing CAPTCHAs for use in telephony systems.

1 Introduction

Telephony is a vital medium for information exchange as it offers services to more than 6 billion mobile users in the world today [19]. In the last decade, telephony systems migrated from Public Switched Telephone Networks (PSTN) to Internet Telephony for communication. Internet telephony also known as Voice over Internet Protocol (VoIP) offers low-cost, and instant communication facilities, e.g., distant calling and video conferencing anywhere around the world. However, VoIP is vulnerable to many attacks such as interception and modification of the transmitting voice packets, and cyber criminal activities, e.g., phishing, and malware [23]. Voice spam, also referred to as Spam over Internet Telephony (SPIT), is an emerging threat to telephony systems causing direct or indirect loss to users. For instance, during the Canadian

[*] The total length of this paper, when put in LNCS format, is at most 16 pages.

spring 2011 federal election, thousands of voters complained about receiving false automated calls, misleading voters about polling locations and discouraging to cast votes [11]. SPIT might originate from the Internet but also affects both PSTN and wireless users. Internet telephony attracts spammers, as SPIT is convenient to generate by reusing the existing botnet infrastructure used for email spam, and can impact masses by sending bulk calls to many users [29]. The prime stakeholders interested in producing SPIT include telemarketers, and malicious bodies intending to fool people forcing them to call expensive numbers (*voice phishing*). The threat raised by SPIT is real and worrisome; for instance, the Telecom Regulatory Authority of India (TRAI) stated that 36,156 subscribers were issued notices, 22,769 subscribers were disconnected, and 94 telemarketers were penalized for spreading spam. Similarly, the Federal Trade Commission (FTC) received more than 200,000 complaints every month in 2012, about automated calls even though 200 million phone users had registered for the "Do-Not-Call" facility. Some of these calls scammed users by offering convenient solutions to save money from their credit card investments [15]. Recent reports show criminals have targeted telephony systems of financial institutions such as banks and emergency helplines, using automated dialing programs and multiple accounts to overwhelm the phone lines [27], [34].

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a mechanism that can differentiate machines (bots) from human users and has successfully reduced the abuse of the Internet resources, particularly reducing spam [36]. Given the success of CAPTCHA (hereafter denoted "captcha" for simplicity) on the Internet, the systems that use *audio captcha* over telephony have been prototyped [24]. FTC also announced a challenge in October 2012 to fight automated calls, asking innovators to propose solutions for stopping robocalls. FTC offered $50,000 to the winners. Many of the solutions proposed during the challenge offered protection through captcha on telephony, e.g., Baton, Telephone Captcha , Call Captcha, and Captcha Calls [2]. The two co-winners of the challenge proposed to filter out unapproved automated calls using a captcha. The winners proposed routing calls to a secondary line, and hung-up the pre-recorded / automated calls using captcha even before the phone rang. This allowed only real callers to connect. However, white lists could be used if a user would like to get genuine automated calls from authenticated sources. Captcha is a classic case for the "human-in-the-loop" paradigm [13] therefore, it is very important to assess how the humans fare with captcha in telephony systems. Telephony communication is lossy in nature, and audio quality is poor in comparison to the Internet which changes the experience of audio captcha on telephony. Audio captcha, like any other captcha, is intended to be easy for humans and hard for machines. However, recent research on telephony captcha largely concentrates on making it hard for machines than making it usable for humans. In the context of the Internet, there has been some recent work on the usability of different types of captcha [7], [10]. Bursztein et al. found that 0.77% (N = 14,000,000) of the time users preferred to answer eBay's audio captcha rather than image captcha on the web [10]. They found this number to be significantly large and felt essential to improve audio captcha on the web [10]. To the best of our knowledge, there has been no comprehensive analysis of the usability of different audio captcha over the phone / mobile through a full-fledged field study. The scope and impact of such an analysis would be quite broad, given that audio captcha will affect every user of the telephony system.

Contributions: In this paper, we analyze the usage pattern and human perspective when faced with captcha on telephony via a comprehensive usability field study. Although captcha could be potentially applied to multiple scenarios, we draw on a banking scenario to transfer money as an initial case study given that users have been affected due to telephony scams [27], [34]. We make the following contributions related to the use of audio captcha in telephony systems:

1. We conducted a field study of various captchas with 90 participants evaluating their performance in the real world telephony environment. Participants remotely dialed our study set-up from 5 cities in India (Delhi, Mumbai, Chennai, Noida, and Vellore).

2. We evaluated and compared the tradeoff among users' accuracy, time spent and keystrokes while solving audio captcha on telephony. We found that the accuracy of captcha decreased multifold on telephony in comparison to web. Also, time taken to solve captcha increased for popular captcha schemes e.g. Recaptcha and slashdot.

3. We propose guiding principles for designing and improving telephony captcha.

Organization: Section 2 elaborates the previous captcha usability studies and telephony captcha studies. Sections 3 and 4 present various hypotheses formulated and the types of captcha used in our study. The implementation and design details, and study methodology are discussed in Section 5. Section 6 elaborates main findings regarding the captcha usage and current user practices. Sections 7 and 8 present the discussion, guidelines proposed from the study and future work.

2 Related Work

SPIT, broadcasting unsolicited bulk calls through VoIP, is one of the emerging threats to the telephony. Various techniques have been suggested to deal with SPIT e.g. device fingerprinting, use of the White, Grey, or Black lists, heuristics, and captcha [3]. Most of the proposed system use combination of one or more above mentioned techniques, to fight SPIT. Captcha is commonly used as a final check for the user. Captcha on the web has been subject to a broad range of research, including design, study of guiding principles, comparative evaluation of different captcha, and attack / threat models. Different types of captcha have been designed for protecting various Internet resources like online accounts, and emails. A few notable design examples include, asking users to identify garbled string of words or characters, identify objects in an image rather than characters [14], identify videos [25], and identify textual captcha that involves clicking on the images rather than typing the response [30]. Other variants include sketcha, which is based on object orientation [31], and Asirra, which uses pictures of cats and dogs [20]. Yan et al. found that contrary to the usual opinion, text based captcha was difficult for foreigners. They showed that the length of captcha interestingly influenced usability and security [21]. Audio captcha, another type of captcha, was especially designed for the visually impaired users on the web [16]. Lazar et al. conducted a study with 40 participants, sighted and blind, to evaluate the radio-clip based captcha. They showed various usability issues related with this captcha and also, proved it to improve the task success rate for sighted users whereas, it was still difficult for blind users [26]. Bursztein et al. showed that captcha especially audio captcha was remarkably difficult for users to solve. They also found that non-native speakers felt English captcha difficult and were slow in solving captcha. [10].

Captcha's effectiveness in controlling machine attacks on the web encouraged its use over telephony. Telephony for years has largely supported verbal / voice based communication. We found that systems proposed for curbing SPIT used the existing web-based audio captcha setup. Tsiakis et al. proposed the principles for understanding the spam economic models, and their analogies to SPIT, which could help evaluate the benefits of audio captcha protection against the costs involved [35]. Polakis et al. developed different attacks, threat models, and solutions for phone captcha [28]. Soupionis et al. proposed and assessed various attributes of an audio captcha, which make it effective against automated test. They also evaluated audio captcha against open source bot implementation [33]. Zhang et al. used out-of-band communication such as the Short Message Service (SMS) to send the captcha text to differentiate bots from humans [37]. In addition to these research efforts, some patents have been developed around audio captcha. These include captcha based on contextual degradation [4] and random personal codes [5]. Johansen et al., developed a VoIP anti-SPIT framework using open source PBX system [24]. A few commercial products aimed at reducing SPIT have also been designed by NEC and Microsoft [1], [18]. However, the usability of audio captcha on telephony in the real world has not been tested. Some preliminary work has been done on evaluating telephony captcha [28]. Soupionis et al. also conceptually evaluated the existing audio captcha solutions [32]. They discouraged use of alphabetic captcha on telephony and suggested a new algorithm for telephony audio captcha. They used soft phones, [1] which were different from real world implementation of telephony captcha. On the contrary, in our work, we conducted our experiment on a real-world system that users could call-in from anywhere in the world.

The study presented in this paper is different from the existing studies in important aspects. Prior studies were geared for the web, using traditional computing devices or for telephony using soft phones, whereas, the target platform for our work is telephony system and the terminal with which the user interacts within our setting is a phone. Furthermore, previous studies were performed in a controlled environment; in contrast, our study is a field evaluation, which involves users interacting with the system in a real world environment. It is also the first study where 90 users participated in evaluating different kind of captchas (web-based and the explicitly designed telephone captcha) in the real world.

3 Hypotheses

We now discuss our hypotheses related to captcha usage and viability on telephony.

H1: Users might be close to the expected / correct answers even though the overall captcha solving accuracy on telephony may be low.

Captcha comprises of a simple challenge not involving much human intelligence, e.g., the addition of 2 numbers. However, the existing literature suggests that users' accuracy to solve audio captcha is low. We hypothesize that the accuracy will be higher, if 1 or 2 mistakes in the user's response are discounted. We calculate the edit distance between the user's response and the correct answer to evaluate response's closeness to the correct answer.

H2: Users' accuracy of answering the captcha correctly on telephony will decrease as the number of key presses required increases.

Audio captcha presents various cognitive challenges to the user. Human brain sequentially processes speech and "short-term" memory handles only $7(+/-2)$ chunks of information [8]. For audio captcha, cognitive load increases with the increase in the size of the captcha, i.e., the number of digits / characters to remember or recognize. This increases the chances of errors. We hypothesize accuracy of users decrease as the size increases. We measure the size (number

[1] Softphone is an application that allows a desktop, laptop or workstation computer to work as a telephone via Voice over IP technology e.g. Skype.

Table 1: Studied audio captchas with their features. Char. set represents the character set. Duration presents average playtime of captcha in seconds. RPC represents Random Personal Code Captcha

Category	**Char. Set**	**Word**	**Repeat**	**Durat-ion**	**Noise**	**Voice**	**Beep**	**Min length**	**Max length**
Google	0-9	No	Yes	34.4	Yes	Male	yes	5	15
Ebay	0-9	No	No	3.7	Yes	Various	No	6	6
Yahoo	0-9	No	No	18.0	Yes	Child	No	6	8
Recaptcha	a-z	Yes	No	10.6	Yes	Female	No	6	6
Slashdot	a-z	Yes	No	2.9	No	Male	No	1	1
CD	1-5	No	No	14	Yes	Male	No	1	1
Math-function	0-9	No	No	6.0	No	Male	No	4	3
RPC	0-9	No	No	20.0	No	Male	No	3	2
C+CD	0-9	No	No	14.0	No	Male	No	4	3

of digits / characters) input by recording the key presses for each captcha. The key presses are recorded in our study setup through the number of DTMF received by the server (as discussed in Section 5).

H3: Users will take more time responding to a captcha that requires more key presses than to the one requiring less key presses.

Size of the captcha, i.e., the number of digits / characters a user has to press also affects the amount of time the user spends on the captcha. We hypothesize that the user will spend more time on a captcha which requires more number of key presses than on a captcha which requires less number of key presses. Analyzing the amount of time a user spends on solving captcha on the telephony is important because more is the time spent, more is the inconvenience caused.

4 Studied Audio CAPTCHAs

In this section, we describe various audio captchas (web and telephony) used in our study and the associated challenges posed to participants. Existing studies proposed to make use of distortions and noise in the existing telephony captcha to improve the security. We found that existing audio captchas on the web offered these features and, therefore, were also included for evaluation. We summarize various features of the web and the telephony audio captchas in Table 1.

Captchas from the Web: We deployed various web based captchas comprising of numeric or alphabetical challenges on telephony in our study. **Yahoo!** offered audio captcha as an alternative to the text captcha on the "create account page" [2] of its e-mail account service. The audio test started with 3 beeps, followed by 6 to 8 digits in various voices (children and females). **Google** captchas obtained from the "Google Account" page were shown when a user requested for an audio captcha. They started with 3 beeps followed by 5 to 15 digits in male voice and offered assistance by repeating the captcha test following the phrase "once again." It took longest to annotate Google captchas (annotation process is discussed in Section 5), as it was difficult to make decisions between noise, e.g. "Oh," "it," and "now," and digits, like "zero," and "eight." **eBay** audio captchas were obtained from the eBay website; these were offered as an alternative to the text captcha. These consisted of the same six digits as shown on

[2] https://login.yahoo.com/config/login?

the website. The tests were consistently 6 digits long in various voices (male and female). The frequency at which digits occurred in the eBay audio was reported to be faster than the other schemes e.g. Google and Yahoo!. **Recaptcha** audios were collected from the "Gmail's login page," and comprised of 6 words in female voice, for example "white, wednesday, two, chef, coach, napkin." **Slashdot** captchas were collected from the "login and join page" of Slashdot website. [3] Each audio presented a word in a male voice. They also offered assistance by spelling the word contained in every test for e.g. "Wrapping w r a p p i n g" where first the complete word was said and then each letter was spelled out.

Captchas for Telephony: We deployed following telephony captchas in our study. **Random Personal Code (RPC)** used random numbers as a menu option instead of a standard menu [5]. For evaluating the scheme, we implemented a 3-digit random menu number using the age details of the participant (discussed in Section 5) e.g. the audio content included, "Press 182, if your age is less than 18." **Contextual Degradation (CD)** [4] suggested adding background noise depending on the context of the call and the individuals associated with the call; this assumes contextual noise is less interfering. For evaluating the scheme, a voice menu was implemented based on the education details of the participant (collected during the pre-study) with mild music from an instrument in the background as contextual noise. **Math-function** captcha [17], recommended using mathematical function, e.g. adding 2 numbers as a test. We implemented this variant requiring participants to add two, 2 or 3 digit numbers chosen randomly, for e.g. "Solve the following, and enter your response; 48 plus 45 to continue." **Math-function with Contextual Degradation (C + CD)** required participants to solve a simple math function with some contextual music playing in the background, e.g. "Solve the following and enter your response; 48 plus 45 to continue". The assumption is that the simple usage of Math function would not offer sufficient security.

5 Usability Evaluation

In this section, we present details of our usability field study, including the experimental setup, study methodology, participant recruitment and demographics.

Experimental Setup: Figure 1 presents the 3 phases of our study: pre-study, participants calling up our setup (the actual study), and post-study. First, participants registered for our study and were provided with a scenario of credit card transaction. They were asked to imagine making a transaction worth 10,000 INR (200 USD) using a bank telephony service. At this time the captcha was shown to the participants. We provided participants with a telephone number where they could call to participate in our study, an application ID, and a PIN / password using which participants to access our setup. In the second phase, participants called up the given phone number and authenticated to the system, after which a series of 9 audio captchas (discussed in Section 4) were presented to them. We developed a counterbalancing schedule [9] to reduce learning effect and avoid the influence of natural responses of participants on captcha solving ability. Participants were divided randomly and equally into groups of 9 each. Every participant in the group was assigned a captcha using Latin Squares of 9 X 9 [9]. In the last phase, an email was sent to participants requesting for their feedback about captchas used in our study. We used System Usability Scale (SUS) in addition to few other questions on the user's

[3] http://slashdot.org/

preference to evaluate system performance.

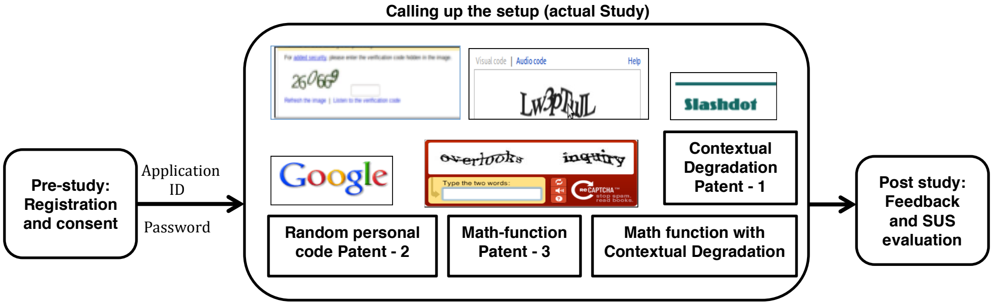

Calling up the setup (actual Study)
Pre-study: Registration and consent
Application ID
Password
Slashdot
Google
overlooks inquiry
Contextual Degradation Patent - 1
Random personal code Patent - 2
Math-function Patent - 3
Math function with Contextual Degradation
Post study: Feedback and SUS evaluation

Figure 1: Three phases of the study: pre-study, calling up the setup, and post-study.

System Design and Implementation: Participants were required to call our system to provide their responses. Our implementation emulated a captcha shield between the Source (caller, which may have malicious intent) and the service being offered (such as making a credit card transaction or checking balance of a bank account). The shield required the caller to correctly answer the presented challenge for accessing the intended service. Participants could call our implementation at anytime and the call could originate from any of the telephony networks, such as PSTN, Cellular Network or VoIP. The input to the captcha shield was DTMFs (Dual-Tone-Multi-Frequency) sent through the key press on the phone. Figure 2 shows the system components included a Linksys gateway SPA 3102 and a Linux Server supporting FreeSWITCH [4]. In order to support calls through FreeSWITCH, we used a Linksys SPA 3102 device as a gateway, which allowed connection to the PSTN network as depicted in Figure 2. The gateway converted analog signals into compatible SIP (Session Initiation Protocol) format, which were then forwarded to FreeSWITCH on the server. We implemented a simple authentication login (using a pre-registered ID and a PIN) for users to participate in the study. For maintaining the privacy of participants, calls were not recorded; only DTMF input to the setup was captured.

Annotators: As a prerequisite to our study, it was essential to create a corpus of audio captchas (discussed in Section 4), and their respective true positive answers. We built the corpus by manual annotation of each captcha; this methodology has been used in the prior studies [32]. Each annotator annotated 50 correct captcha of the assigned service, e.g. eBay, and Google. They entered their response for the captcha challenge presented to them through the respective web service. Once the answer was accepted as correct answer by captcha server of the website, it was added to correct captcha corpus. Annotators were recruited through the mailing list of our institute. They were from Computer Science background; some were undergraduate but the majority were postgraduate students, pursuing Masters or Ph.D degrees at the IIITD, India, e.g. Slashdot annotator visiting Slashdot new user page. [5]

Participants: We recruited participants for evaluating captcha through word of mouth. Participants from different cities of India, Mumbai, Chennai, Delhi, Vellore, and Noida took part in the study. A monetary reward of 75 INR (1.5 USD) was offered to every participant for contributing to the study. One hundred and forty participants registered for the study; of these 90 participants called up our system and completed the study. We were not required to go through

[4] FreeSWITCH is one of the open source telephony platforms which has enabled easy access to telephony often required by various businesses. http://www.http://freeswitch.org//

[5] https://slashdot.org/my/newuser

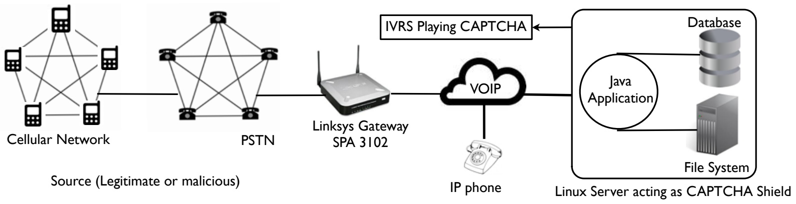


Figure 2: The IVR system setup. The user dials the IVR on phone from any network. The call is received by Gateway, which forwards it to FreeSWITCH. An application written in Java answers the call by playing appropriate voice prompts. The Java application accesses database, file-system for storing and retrieving information.

an Institutional Review Board (IRB) approval process before conducting the study. However, the authors of this paper have previously been involved in studies with U.S. IRB approvals, and have applied similar practices in this study. Participants were shown “consent information” on the registration page, which they agreed to for participating in our study. Participants belonged to different age groups, ranging from 18 years to 65 years. Participants’ profession varied with 45.56% from Computer Science, 8.89% from Designing in various fields, 10% from Finance, and 35.55% were Lawyers, Journalists, and Homemakers. As reported, 25.56% participants had used IVR for 2 to 3 years, 24.44% had used IVR for 3 to 5 years and another 25.56% for 5 years also, and 24.44% users had used IVR for 0 to 2 years. Participants could call in from any type of a phone, may use a headphone or a paper and pen during the call.

6 Study Results

We now discuss the two primary research questions of our study: how much inconvenience does the captcha causes to the users, and how different features of captcha, e.g. duration, size and character set influence captcha’s usability.

6.1 Captcha Inconvenience

In this section, we discuss the first question, i.e., the inconvenience caused to the users in terms of time spent, accuracy and experience with the different audio captchas.

Accuracy: Captchas are meant to be easy for humans to solve however, captchas which can’t be guessed correctly cause considerable inconvenience to the users [10]. We calculate accuracy defined as the percentage of all the captchas solved correctly in each category to measure captcha inconvenience. Table 2 shows that the maximum average accuracy was only 13.71% for telephony captcha. CD telephony captcha performed the best with 18.71% accuracy. It was easier to remember as it involved one time instruction for users like select the correct option from a simple menu [12]. Among the web-based captcha, alphabet captcha performed marginally (Δ = 0.97%) better than number captcha. This marginal improvement could be attributed to the fact that random numbers are more difficult to remember than common English words [12]. Table 2 shows that the percentage of participants who skipped responding to captcha challenge was maximum (46.07%) for Recaptcha. High Skip rate and low accuracy for most captchas show the difficulty caused in solving audio captcha on telephony. We found significant difference (Chi-sq = 81.42, p-value < 0.001) between Accuracy and Error showing considerable inconvenience caused to users while solving audio captcha.

Further comparing the accuracy of audio captcha on the web and telephony, we analyzed our results with existing studies on the web. Bursztein et al. found that Google's audio captchas were hardest to solve, with only 35% *optimistic accuracy* [10], while Slashdot and Yahoo! performed the best with 68% accuracy followed by eBay yielding 63% accuracy on the web. Participants achieved 47% accuracy for Recaptcha on the web. However, in our experiment the accuracy was nil for Google captcha. Slashdot performed the best on telephony as well, but accuracy dropped to 13.73%. eBay performed marginally (1.01%) better than Yahoo! on telephony. We understand that these comparisons may not give an exact picture of the differences between the web and telephony as the two experiments were conducted in different environments and with participants of different ethnicity. However, the large difference in accuracy on the two media, shows the significant increase in the inconvenience caused to participants when shown audio captcha on telephony.

Table 2: Aggregate of error, accuracy values for each of the voice captcha in percentage. Skip gives the percentage of participants who skipped the Captcha. N represents the total number of Captcha presented in each category.

Captcha	**Category**	**Time (s)**	**Accuracy (%)**	**Skip (%)**	Total (N)
CD	Telephony	96.11	18.71	35.67	171
Math-function	Telephony	90.23	17.47	26.51	166
RPC	Telephony	147.44	15.47	40.33	181
C + CD	Telephony	109.59	4.57	40.10	197
Total	**Telephony**	**85.03**	**13.71**	**35.94**	**715**
Ebay	Web-Number	80.25	8.75	13.13	160
Google	Web-Number	123.49	0.00	43.83	162
Yahoo	Web-Number	95.88	7.74	20.24	168
Total	**Web-Number**	**99.87**	**5.51**	**25.71**	**490**
ReCaptcha	Web-Alphabet	120.64	0.00	46.07	171
Slashdot	Web-Alphabet	122.57	13.73	30.06	153
Total	**Web-Alphabet**	**121.60**	**6.48**	**39.51**	**324**

Solving Time: Time is another important aspect which helps in measuring the inconvenience caused to users while solving an audio captcha on telephony. It has been found on the web that a captcha which takes more than 20 seconds causes inconvenience to the users [10]. We found that users took much longer to solve a captcha on telephony than on the web through our study. We measured solving time per captcha as the time elapsed from the instance when the user is presented with a captcha to the instance when the user moves to the next captcha type. Table 2 shows the average time taken to solve different captcha types. Users took the least of 80.25 seconds to solve eBay captcha. There was no significant difference in time taken by users for solving web-based captcha to telephony-based captcha. We found that the most time consuming captcha on telephony was RPC captcha with an average solving time of 147.44 seconds. Among the web-based captcha, Google, ReCaptcha and Slashdot were the most time consuming with mean greater than 120 seconds (min: 120.64 seconds and max: 123.49 seconds). On analyzing the existing studies on the web, we found that Google and ReCaptcha were the most time consuming with a mean value greater than 25 seconds. However, users took on an average 12 seconds to solve Slashdot captcha on the web [10] in comparison to 122.57 seconds on telephony. The increase in the solving time shows that the inconvenience caused to users is more for solving audio captcha on telephony than that on the web.

User Experience: As the ultimate assessment metric to understand the inconvenience caused, we study the feedback provided by the participants for different captcha types. Figure 3(a) shows 50% of the participants found the system (on which users called to answer captcha in our study) extremely complex to use. We found statistical difference (Chi-sq = 12.77, p-value $<$ 0.001) in user's preference for different captcha (numeric, alphabetic or math function). Users' feedback suggests that they did not like alphabet audio captcha as only 14.44% of users preferred alphabet audio captcha. Most participants (52.22%) preferred numeric captcha (Contextual Degradation, Random menu captcha and numeric web based captcha) and 33.33% of the users favored captcha involving math functions (math-function and math-function with contextual degradation). We also analyzed if age has any effect on captcha preference of the participants. Figure 3(b) shows participants in the age group of 36 to 50 did not prefer to use the alphabet audio captcha at all, where as numeric captcha was appreciated among all age groups.

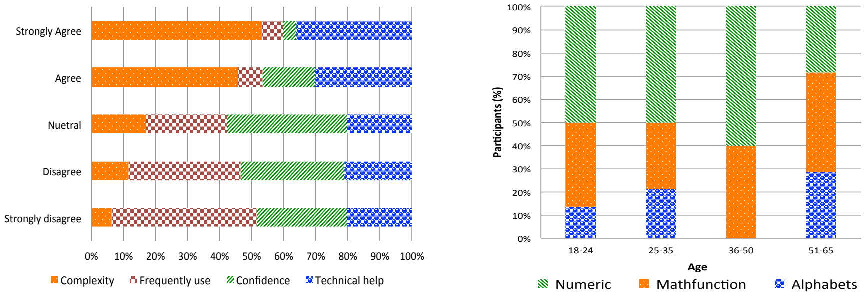
Next, we wanted to analyze if participants felt that using speakerphones or headphones would help them solve the captcha better. We did not find significant difference between participants who agreed / disagreed that use of speakerphones would help (Chi-sq = 0.3846, p-value = 0.5351) but headphones were felt to be helpful by significant number of participants (Chi-sq = 22.70, p-value <0.001). We found that 30% of the users disagreed and 8.89% strongly disagreed, feeling that speakerphones will be of no use. A participant mentioned *"the voice recording was not clear, therefore, any audio accessory would not have helped"*disagreeing with the use of speakerphones. We also asked the participants if they felt the use of headphones would help them solve the captcha better. The results show that 15.56% of the users strongly agreed and 70% agreed; using headphones would help them respond better to the challenge.

In order to understand the user perspective about the mode of input, we asked users if they would prefer to respond verbally to the captcha challenge. We found that statistically significant (Chi-sq = 24.8205, p-value $<$ 0.001) number of participants felt that the verbal input would be helpful; 36.67% of the users agreed with the statement and 28.89% strongly agreed. This suggests that entering responses using a keypad is difficult and causes trouble for respondents; this might be one of the primary reasons for errors in the captcha responses leading to low solving accuracy in the study [22]. We suggest the need for further research to investigate what makes an audio captcha easier to answer – verbal response or keypad touch.We calculated the SUS score for the telephony captcha as 38.42, which is extremely low. [6]. Participants (35.71%) also complained about voice being not clear, a participant commented *"It sounded like ghost voices. I was not able to understand almost any utterance."* Around 8.92% explicitly complained about the accent in the voice captcha stating, *"at least the accent should be better."* Participants (17.85%) also complained about the noise in the audio being very disturbing. A participant commented, *"too many disturbances and the accent was bad and 80% of time I couldn't understand."*

6.2 Hypotheses Validation

The inconvenience parameters discussed above are important for evaluating the usability of captcha. However, these do not provide much guidance on how different captchas feature like duration, and length of captcha (as discussed in Table 2) influence the users' accuracy. In this section, we test the various hypotheses (discussed in Section 3) to study the influence of some of these features on audio captcha usability.

[6] Given that SUS is 68 for average usable system.http://www.measuringusability.com/sus.php



(a) (b)

Figure 3: (a) Users reported the system to be complex, not usable, suggesting the need for technical help for using captcha over telephony. (b) Participants in various age groups found the numeric captcha to be most usable whereas alphabet captcha was least preferred by the participants.

H1– Accuracy vs. Closeness: Section 6.1 shows that accuracy of solving captcha on telephony was low as maximum accuracy achieved was only 13.71% for telephony captcha. However, we found that the users were close to the actual answer of the captcha presented. To analyze closeness, we calculated the most frequent edit distance (also known as Levenshtein distance) between the user's response and the expected answer. Figure 4 shows that the edit distance was 2 for most of the captcha types, followed by 1 (μ = 3.22 and σ^2 = 3.40), implying that users committed 2 errors per captcha response frequently. As an example, for the audio challenge, "824388" (eBay), the user responded "624386". Users were required to remember and input many words for Recaptcha resulting in exceptionally high edit distance. This supports our Hypothesis H1. This also suggests the need for further research on developing fault tolerant captcha, which could distinguish between human natural behavior error variance, and machine attacks on audio captcha. Analyzing human behavior when solving captcha has already been proposed for image based captcha, which can help decrease the inconvenience caused to users. [7]

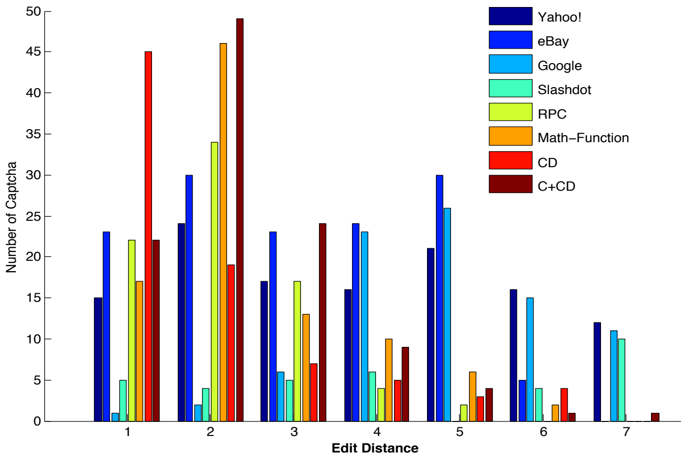


Figure 4: The relation between edit distance and number of captcha for each captcha type.

H2 – Expected key press vs. Accuracy: We found that the length of the answer i.e. size of

[7] Are you human Captcha. Secure from All Angles, http://areyouahuman.com/security/.

captcha did not influence the accuracy of the user's response. The accuracy decreased from 18.71% (CD) to 4.57% (C + CD) irrespective of the fact that users were required to solve the captcha comprising of similar tasks and comparable expected DTMF count. Similar trends were noticed for the web captcha, where users were expected to identify a similar number of spoken digits (eBay, Google, Yahoo) but the accuracy varied from 0% to 8.75%. Table 3 shows the expected DTMF users were required to input for a correct response and the corresponding accuracy for each captcha type. The correlation between expected DTMF count and accuracy was found to be very low (r = −0.52) with a negative fit. Further analysis of the data showed the presence of a positive relationship between Expected key press and accuracy parameters for Math function, but we noticed a negative relationship with correlation coefficient r = − 0.47 for web-based captcha. Finally, we found the significant difference (t-test, t-value = 5.30 p-value < 0.001) between Expected Key Press (Average DTMF) and accuracy in statistical results shows that these two were independent of each other. The results mentioned above do not approve our hypothesis H2.

Table 3: The Average DTMF expected for captcha (Avg. DTMF), accuracy, time and Average DTMF input by users (Avg. User DTMF) of each captcha. N represents the total number of Captcha presented in each category.

Scheme	**Category**	**Avg. DTMF**	**Accuracy**	**Time**	**Avg. User DTMF**	**Total (N)**
CD	Telephony	1.00	18.71	96.11	1.76	171
Math-function	Telephony	2.05	17.47	90.23	2.71	166
RPC	Telephony	3.00	15.47	147.44	3.92	181
C + CD	Telephony	2.06	4.57	109.59	2.65	197
Ebay	Web	6.00	8.75	80.25	3.85	160
Google	Web	6.36	0.00	123.49	4.68	162
Yahoo	Web	7.09	7.74	95.88	4.99	168
Slashdot	Web	15.34	13.73	120.64	6.02	153
ReCaptcha	Web	64.93	0.00	122.57	10.97	171

H3 – Time vs. Number of key press: Table 3 shows that users spent varying amount of time in submitting a comparable number of DTMF responses. For example, the average time spent for Google was 123.49 seconds (min: 17.15 and max: 341.21) whereas for Yahoo, it was 95.88 seconds (min: 25.88 and max: 278.00), although both of them had same average DTMF (5) to input. There was a significant difference between the time taken to solve Google vs. Yahoo! (t-Test, t-value = -12.39, p-value < 0.01). Further, we found a correlation (r =0.85) between time spent and DTMF input for Math-function captcha, suggesting an increase in the time was proportionate to DTMF input. However, this correlation dropped to r = 0.56 for web-based captcha, implying an absence of any strong relativity between time and DTMF input. The results from our study suggest lack of any strong relationship between the time spent by the participants in solving a captcha and the number of DTMF input from them. We found that the correlation between the time spent to answer the captcha and DTMF response from the users was 0.36 for all the captcha used in our study. We found the significant difference (t-test, t-value = 4.33, p-value < 0.0001) between number of key press (Average User DTMF) and accuracy in statistical results suggesting that these two were independent of each other. We further tested, if the duration for which a captcha is played influences the accuracy but found that exposing

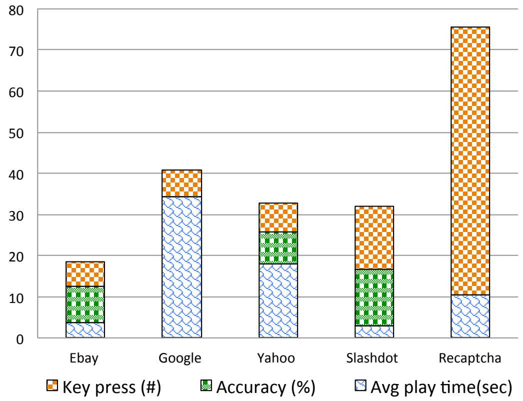


Figure 5: Figure shows Aggregate accuracy (%), Number of DTMF and Average play time (seconds) of web-based captcha. Exposing users longer to Captcha does not help improve accuracy.

users longer to a captcha did not help improve solving accuracy. Figure 5 shows the average playtime of the number web-based captcha (eBay, Yahoo, Google) varied from as low as 3.7 to 34.4 seconds where all these required a similar number of DTMFs to be recognized. Google captcha provided a feature to repeat the challenge in each attempt, without users asking for it explicitly, irrespective of these; the correct response was 0% for Google and 8.75% for eBay.

7 Discussion

In this paper, we explored user's experiences and views on telephony captcha through a real world study. We concentrated on two primary research questions: how much inconvenience does captcha causes to the user, and how different features of the captcha, e.g. duration, size and character set influence captcha's usability. We measured the level of inconvenience caused to users in terms of time spent, accuracy, and perceived difficulty / satisfaction of using various captchas schemes. We found that the accuracy of captcha decreased multifold on telephony in comparison to web, and the time taken to solve captcha increased to as high as 122.57 and 120.64 seconds for popular captcha schemes, e.g. Recaptcha and Slashdot, respectively. However, we found that users were relatively close to the expected correct answers (discussed in Hypotheses Validation of H1), which may suggest the possibility of deploying captcha on telephony platforms in the future. We found that the captcha could be a viable solution for telephony with improved features, such as fault tolerance, better voice and accent as most participants suggested in their feedback. Next, to study the influence of different features, which might improve accuracy, and usability of captcha on telephony, we tested 3 hypotheses. The results from the study supported hypotheses 1, whereas hypothesis 2 and 3 remain tenable. We did not find strong influence of captcha size and duration on solving accuracy.

Contrary to the existing work [28]; we found Math-function captcha performed better than alphabet captcha. Soupionis et al. discarded the use of alphabet captcha assuming that sending letters to answer a captcha could be difficult for an average user. However, we found that alphabet captcha's accuracy was comparable to numeric captcha. Finally, we present our results for recommendations proposed in literature for improving telephony captcha. Polakis et al. suggested adding distortion to speech signals to make it more usable for humans and difficult for machines rather than just adding noise [28]. We used the distortions available on the web-based captchas to test the recommendation. We understand that these distortions might not be the best but as these have been able to provide sufficient security on the web, therefore, for initial testing we used distortions available in web-based captcha. Contrary to the recommendations, we found that it was still difficult for participants to solve these captchas. We found that both forms of noise, intermediate and background, caused inconvenience for the users and captcha solving accuracy was low in the presence of both of them.

We suspect instances during the study, when the DTMF sent by the users may not have completely reached the captcha server. In our setup, the receiver used a PSTN network, which helped in minimizing such errors but in certain real world scenario, where both receiver and caller are wireless users, the lossy nature may lead to total failure of the captcha test [6]. The telephony captcha would probably need to be more loss-tolerant, but this would demand a trade-off between accuracy of captchas and security (i.e., resistance to automated attacks). Telephony network consists of inherent noise, which effects the audio quality. As also observed by the annotators during our annotation phase, the quality of the audio being played through the laptop speaker (often used in the studies so far with soft phones) was much more audible and clear than the one from speakers of the telephone / mobile phone. It would be interesting to study the influence of these factors on the captcha solving performance.

8 Guidelines

Based on our overall results and analysis, we recommend following design conventions for improved telephony captchas.

1. *One time instruction:* All telephony captchas and slashdot captcha in our study presented one time instruction to the users whereas most web-based captchas had random numbers or strings. We found that the accuracy of the telephony captcha, and Slashdot was comparatively better than random strings. This indicates that the captcha involving one time instruction or challenge are appropriate for telephony interfaces (e.g., one word challenge or logic questions) as speech competes with verbal processing [8].

2. *Loss / error tolerant:* The captcha should be error / loss tolerant since the telephony network might drop DTMF carrying user response. The telephone network itself introduces its own noise causing the audio to degrade [6]; hence external noise has to be calibrated accordingly.

3. *Feedback:* Visual / audio feedback for the response should be made available to the users for the response they input, especially required on voice medium because of its inherent lossy nature. We suggest presenting the characters of the captcha for a fraction of a second while the user is solving it (e.g. as in the some of the latest handheld devices for entering the passwords).

4. *Verbal responses:* Users are better trained and prefer voice as input modality instead of keypad touch as telephony networks are primarily voice based. It is advisable to use captcha based on secure verbal inputs (voice recognition) instead of key press. This can help reduce the manual errors in keying in the responses and improve usability.

Finally, we note that our study has some limitations. As the study was conducted in the real world, we had no control on the environmental conditions and assume that all participants had the similar environment. Participants could opt to finish the study in multiple sessions, some participants who called in multiple times to the system were exposed more to the captchas and experienced varied cognitive load.

In the future, we envisage applying other techniques such as illusion effects, earcons to build a captcha system and to evaluate the effectiveness of such new approaches. We plan to conduct a detailed study of the cognitive loads and psychology of the auditory system with captchas to help us design a better mechanism to differentiate between machines and human beings.

Acknowledgements

The research of the first author is supported by TCS (Tata Consultancy Service) Research Scholarship. We would like to thank Dr. Iulia Ion and Paridhi Jain for their input on the study. The authors would also like to thank International Development Research Centre (IRDC) and all members of PreCog research group at IIIT-Delhi. We would like to thank Siddhartha Asthana for helping in deployment of the system and all participants in the study.

References

[1] The Dark Side of Voice. http://content.yudu.com/A1qlhz/CommsDealerJan11/resources/38.htm.

[2] FTC Robocalls Challenge, 2012, http://robocall.challenge.gov/submissions/.

[3] N. K. Andreas U. Schmidt and R. E. Khayari. Spam over internet telephony and how to deal with it. *arXiv preprint arXiv:0806.1610*, 2008.

[4] H. Baird, J. Bentley, D. Lopresti, and S.-Y. Wang. Methods and Apparatus for Defending Against Telephone-Based Robotic Attacks Using Contextual-Based Degradation. 2011. United States Patent.

[5] H. Baird, J. Bentley, D. Lopresti, and S.-Y. Wang. Methods and Apparatus for Defending against Telephone-Based Robotic Attacks using Random Rersonal Codes. 2011. United States Patent.

[6] V. A. Balasubramaniyan, A. Poonawalla, M. Ahamad, M. T. Hunter, and P. Traynor. PinDr0p: Using Single-Ended Audio Features To Determine Call Provenance. Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010.

[7] J. P. Bigham and A. C. Cavender. Evaluating Existing Audio CAPTCHAs and an Interface Optimized for Non-Visual Use. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2009.

[8] D. G. Bonneau and H. E. Blanchard. Human factors and voice interactive systems. *Signals and Communication Technology*, Springer, 2008.

[9] J. Bradley. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528, 1958.

[10] E. Bursztein, S. Bethard, C. Fabry, J. Mitchell, and D. Jurafsky. How Good Are Humans at Solving CAPTCHAs? A Large Scale Evaluation. *Security and Privacy (SP), 2010 IEEE Symposium on.*

[11] Canadian election robocall scan. http://news.nationalpost.com/2012/03/05/robocalls-scandal-likely-the-fault-of-elections-canada-tory-mp/.

[12] G. Cooper. *Research into cognitive load theory and instructional design at UNSW.* http://webmedia.unmc.edu/leis/birk/CooperCogLoad.pdf. 1998.

[13] L. F. Cranor. A framework for reasoning about the human in the loop. In *Usability, Psychology, and Security*, 2008.

[14] R. Datta, J. Li, and J. Z. Wang. Imagination: a robust image-based captcha generation system. In *MULTIMEDIA '05*, pages 331–334.

[15] Federal Trade Commission. Robocalls: All the rage, an FTC summit. http://www.ftc.gov/bcp/workshops/robocalls/docs/RobocallSummitTranscript.pdf, 2012.

[16] G. Sauer and H. Hochheiser and J. Feng and J. Lazar. Towards a Universally Usable CAPTCHA. In *Symposium On Usable Privacy and Security*, 2008.

[17] J. N. Gross. Captcha Using Challenges Optimized for distinguishing between humans and machines. U.S. Patent Application, 2009.

[18] D. Hoffstadt, C. Sorge, and Y. Rebahi. Spam over internet telephony. http://www.tu-chemnitz.de/etit/kn/Zukunft_der_Netze/presentation_hoffstadt.pdf.

[19] International Telecommunication Union. Measuring the information Society, `http://www.itu.int/ITU-D/ict/publications/idi/material/2012/MIS2012_without_Annex_4.pdf`.

[20] J. Elson, J. Douceur and J. Howell and J. Saul. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. In *ACM Conference on Computer and Communications Security*, 2007.

[21] J. Yan, A. Ahmad. Usability of CAPTCHAs Or usability issues in CAPTCHA design. In *Symposium On Usable Privacy and Security*, 2008.

[22] M. Jakobsson and R. Akavipat. Rethinking passwords to adapt to constrained keyboards. In *MoST*, 2012.

[23] M. Jakobsson and Z. Ramzan. *Crimeware: Understanding New Attacks and Defenses*. Symantec Press, 2008.

[24] A. J. Johansen. Improvement of spit prevention technique based on turing test. Master's thesis, Mahanakorn University of Technology, 2010.

[25] K. Kluever and R. Zanibbi. Balancing Usability and Security in a Video CAPTCHA. In *Symposium On Usable Privacy and Security*, pages 1–11, 2009.

[26] Lazar et al. POSTER: Assessing the Usability of the new Radio Clip Based Human Interaction Proofs. Symposium On Usable Privacy and Security, 2010.

[27] S. Martin. Hold the Phone—Will TDoS Be Your Next Big Threat? http://bankinnovation.net/2013/07/hold-the-phone-will-tdos-be-your-next-big-threat/, July 2013.

[28] I. Polakis, G. Kontaxis, and S. Ioannidis. CAPTCHuring Automated (Smart) Phone Attacks. In *SysSec Workshop (SysSec), 2011 First. IEEE, 2011.*

[29] J. Quittek, S. Niccolini, S. Tartarelli, M. Stiemerling, M. Brunner, and T. Ewald. Detecting spit calls by checking human communication patterns. Communications, 2007. ICC'07. IEEE International Conference on. IEEE, 2007.

[30] R. Chow and P. Golle and M. Jakobsson and L. Wang and X. Wang. Making CAPTCHAs Clickable. In *HotMobile*, 2008.

[31] S. Ross and J. Halderman and A. Finkelstein. Sketcha: A CAPTCHA Based on Line Drawings of 3D Models. In *Conference on World Wide Web (WWW)*, 2010.

[32] Y. Soupionis and D. Gritzalis. Audio CAPTCHA: Existing solutions assessment and a new implementation for VoIP telephony. *Computers & Security*, pages 603–618, 2010.

[33] Y. Soupionis, G. Tountas, and D. Gritzalis. Audio CAPTCHA for SIP-Based VoIP. In *SEC 2009*, IFIP, pages 25–38. Springer, 2009.

[34] The Federal Bureau of Investigation. The Latest Phone Scam Targets Your Bank Account. http://www.fbi.gov/news/stories/2010/june/phone-scam, June 2010.

[35] T. Tsiakis, P. Katsaros, and D. Gritzalis. Economic evaluation of interactive audio media for securing internet services. In *ICGS3/e-Democracy*, pages 46–53, 2011.

[36] L. von Ahn, M. Blum, and J. Langford. Telling Humans and Computers Apart (Automatically) or How Lazy Cryptographers Do AI. *Computer Science Department (2002): 149.*

[37] H. Zhang, X. Wen, P. He, and W. Zheng. Dealing with telephone fraud using captcha. In *ICIS*, 2009.