

How Good are Humans at Solving CAPTCHAs? A Large Scale Evaluation

Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell, Dan Jurafsky

elie@cs.stanford.edu, bethard@stanford.edu, celinefabry@gmail.com, mitchell@cs.stanford.edu, jurafsky@stanford.edu
Stanford University

Abstract—Captchas are designed to be easy for humans but hard for machines. However, most recent research has focused only on making them hard for machines. In this paper, we present what is to the best of our knowledge the first large scale evaluation of captchas from the human perspective, with the goal of assessing how much friction captchas present to the average user.

For the purpose of this study we have asked workers from Amazon’s Mechanical Turk and an underground captcha-breaking service to solve more than 318 000 captchas issued from the 21 most popular captcha schemes (13 images schemes and 8 audio scheme).

Analysis of the resulting data reveals that captchas are often difficult for humans, with audio captchas being particularly problematic. We also find some demographic trends indicating, for example, that non-native speakers of English are slower in general and less accurate on English-centric captcha schemes. Evidence from a week’s worth of eBay captchas (14,000,000 samples) suggests that the solving accuracies found in our study are close to real-world values, and that improving audio captchas should become a priority, as nearly 1% of all captchas are delivered as audio rather than images. Finally our study also reveals that it is more effective for an attacker to use Mechanical Turk to solve captchas than an underground service.

I. INTRODUCTION

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) are widely used by websites to distinguish abusive programs from real human users. Captchas typically present a user with a simple test like reading digits or listening to speech and then ask the user to type in what they saw or heard. The image or sound is usually distorted in various ways to make it difficult for a machine to perform the test. When successful, captchas can prevent a wide variety of abuses, such as invalid account creation and spam comments on blogs and forums. Captchas are intended to be easy for humans to perform, and difficult for machines to perform. While there has been much discussion of making captchas difficult for machines (e.g. [2], [4], [13]), to the best of our knowledge there has been no large scale study assessing how well captchas achieve the former goal: making it easy for humans to pass the test.

We address this problem by collecting captcha samples from each of the 13 most used image schemes and 8 most used audio schemes, for a total of over 318,000 captchas.

We then ask humans to solve these and analyze their performance.

Our current annotation efforts have resulted in over 5000 captchas for each image scheme and 3500 captchas for each audio scheme, each annotated by at least three different humans from Amazon’s Mechanical Turk. We also had an additional 1000 captchas from each image scheme annotated three times by an underground service which promises to manually solve captchas in bulk. Based on an analysis of these captchas, we make a number of important findings:

- Despite their goals, captchas are often hard for humans. When we presented image captchas to three different humans, all three agreed only 71% of the time on average.
- Audio captchas are much harder than image captchas. We found perfect agreement by three humans only 31% of the time for audio captchas.
- Some captcha schemes are clearly harder for humans than others. For example, three humans agreed on 93% of *authorize* image captchas, but only 35% of *mail.ru* image captchas.

We obtained statistics from eBay regarding 14 million eBay captchas delivered over a week. This additional data corroborates our Mechanical Turk and underground captcha service statistics and underscores the importance of audio captchas:

- Our Mechanical Turk assessment of eBay image captchas is lower than eBay’s measured success rate: our data shows 93.0% accuracy, compared to eBay’s measured success rate of 98.5% on 14,000,000 eBay site captchas.
- Evaluating the utility of audio captchas is important as they account for almost 1% of all captchas delivered.

We also analyze human variations along a number of demographic lines and find some interesting trends:

- Non-native speakers of English take longer to solve captchas, and are less accurate on captchas that include English words.
- Humans become slightly slower and slightly more accurate with age.
- Ph.D.s are the best at solving audio captchas.

Finally our study shows that for attackers, it is more

efficient to use Mechanical Turk to solve captchas than the underground service, as it is cheaper and more accurate.

All these findings contribute to a better understanding of the human side of captcha tests. The remainder of the paper is organized as follows: In Sec II, we discuss our study methodology. In Sec. III, we introduce the 13 image and 8 audio schemes we analyzed. In Sec. IV, we provide usage statistics for eBay on a 14 million captcha corpus gathered over a 7 day period. In Sec. V, we present the results of our study. In Sec VI, we discuss how the user demographics affect the captcha solving process. In sec. VII, we present some additional related work. Finally, we conclude and give future directions in sec. VIII.

II. STUDY METHODOLOGY

We designed our study for two purposes: to collect information on the speed and accuracy with which humans solve captchas, and to collect information about a broad range of design and demographic factors that could potentially influence these results. To build our captcha corpus, we collected eleven thousand captchas from the **21 most used schemes: 13 image schemes and 8 audio schemes**. In total we scraped more than **90 000** captchas, as discussed in section III below. For human subjects on which to test these captchas, we relied on two sources: *Amazon’s Mechanical Turk* and a underground captcha-breaking service called *Bypass-captcha*.

A. Amazon’s Mechanical Turk

Amazon’s Mechanical Turk (AMT) is an online marketplace from Amazon where requesters can find workers to solve Human Intelligence Tasks (HITs). This service is designed to tackle problems that are difficult for a machine to solve but should be easy for humans. Essentially, the AMT service is intended as a way of crowd-sourcing interesting problems, and has been used extensively to collect annotated data on a variety of tasks, including image classification and filtering of porn for websites. Since AMT provides easy access to human intelligence, it is the perfect service to conduct the task of solving captchas, which is supposed to be easy for humans and hard for computers.

Any task that can be presented as a webpage can be crowd-sourced through AMT, and workers will often perform complicated tasks for relatively small amounts of money (for example, as little as **\$0.05** for solving **one** of our captchas). AMT also has the advantage that the workers (colloquially, “*Turkers*”) are real people and can be asked demographic information such as age, education level, native language, etc., which, as we discuss in section VI, are important for understand how taxing captchas are on different people.

In our experiment, we presented *Turkers* first with a survey asking the following demographic information:

- Age
- Native language (one from the Wikipedia list¹)
- Education (one of: no formal education, high school, bachelors degree, masters degree, Ph.D)
- (If native language is not English) Years studying English
- Industry (one of the United States Bureau of Labor Standard Occupational Classifications)
- Country of birth
- Country of residence
- Years using the internet
- Frequency of internet use (daily, weekly, monthly or yearly)

After filling out this survey, *Turkers* were then presented with **39** image captchas or **24** audio captchas, one at a time, and asked to type in their answers. We built a task scheduler to ensure that three *Turkers* (see Sec. V) saw each captcha even though some *Turkers* gave up on some tasks. As a result, we ended up effectively having more than **318 000** captchas annotated by *Turkers*. In particular we had a very high give up rate, around **50%**, for audio captchas as they are tedious to solve.

Our task scheduler presented the captchas to *Turkers* in two different ways to make sure the order did not influence the results of the study:

- **Random Order:** Fully randomly, where any captcha from any scheme could follow any other.
- **Blocks of Three:** In blocks of three captchas from the same scheme, where the schemes were ordered randomly.

For each captcha, we recorded the time a *Turker* took to solve it and their final response². *Turkers* were then paid anywhere from *\$0.02* to *\$0.50* for solving their full set of 39 image or 24 audio captchas.

B. Underground captcha-Breaking Service

We also investigated using a underground captcha-breaking service, *captcha-bypass.com*. This service promises that captchas submitted to them will be solved by “qualified specialists” for *\$0.005* per captcha. They used to provide an web service that can be accessed via an application programming interface (API) available for .Net, C++. PHP and Java. Since it is a web based service, HTTP call can also be used to interact with it from any language.

For the purpose of this study we used the PHP API to collect data on accuracy and solving time. Of course, the demographic information available through AMT is not available through this service. However, we performed this experiment because to the best of our knowledge, there

¹http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

²We also recorded information about their computing environment, including their browser and operating system, with the hope of using this information in future research

have been no previous studies on how efficient underground services are at solving captchas, and this experiment can therefore shed some light on an unexplored corner of the black market. Overall we submitted **39000** captchas to this service.



Front page of the underground service we used.

III. CAPTCHA COLLECTION

To run our study on humans, we first needed to collect a large number of captchas representative of what people encounter on the web. We consulted the Alexa list of most used websites³ and identified the top sites which presented captchas as part of their account registration process. We also collected captchas from sites which provide captchas to other sites, e.g. recaptcha.net and captchas.net. We looked both for image captchas and for audio captchas, which are sometimes provided as an alternative to image captchas as an accessibility measure. For each scheme, we collected *11,000* captchas. Tables I and II compare some of the features of the captchas from each of these schemes, and the following sections give a little more detail about the sites and their captchas.

A. Image captchas

Authorize.



Authorize.net is a gateway through which other sites can accept credit cards and electronic check payments ala Paypal. Image captchas from authorize.net consist of five black digits and lowercase letters on a white background. To obfuscate the text, the character sequence as a whole is squeezed and tilted to varying degrees, and both the characters and the background are spotted lightly with gray dots.

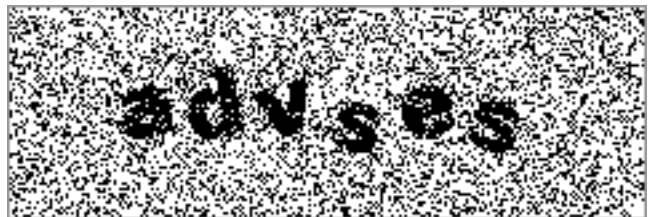
³<http://www.alex.com/topsites>

Baidu.



Baidu.com is the most popular Chinese search engine. Image captchas from baidu.com consist of four black digits and uppercase letters on a white background. To obfuscate the text, a wavy line is drawn across the characters and characters are individually tilted and placed so as to overlap one another.

captchas.net.



captchas.net provides captchas to other sites such as www.blog.de and www.blog.co.uk. Image captchas from captchas.net consist of a white background with six black lowercase letters. To obfuscate the text, individual characters are randomly rotated and shifted up or down and spotted with small white dots, and the background is spotted thickly with small black dots.

digg.com.



Digg.com is a site where users can post web links and vote on the links posted by others that they find most interesting. Image captchas from digg.com consist of five black digits and letters (upper and lowercase) on a white and gray striped background. To obfuscate the text, individual characters are randomly rotated and shifted up or down, and a dense cross-hatch of thin black and gray lines is drawn across the image.

eBay.



Ebay.com is the world's largest online marketplace, where users can buy and sell almost anything. Image captchas from

Scheme	Auth.	Baidu	capt.	Digg	eBay	Ggle	mail.ru	MS	Recap.	Skyrock	Slash.	Blizzard	Y!
Min Len	5	4	6	5	6	5	6	8	5	5	6	6	5
Max Len	5	4	6	5	6	10	6	8	20	6	8	8	8
Char set	a0	0A	a	a	0	a	0aA	0A	0aA-_!	a0	a	a0	0aA
Word	no	no	no	no	no	pseudo	no	no	yes	no	yes	no	no

Table I
IMAGE CAPTCHA FEATURES

ebay.com consist of six letters in a dark color on a white background. To obfuscate the text, individual characters are randomly tilted, shifted up and down and placed so as to overlap one another.

Google.



Google.com is a popular search engine that provides many other services, such as webmail. Image captchas from google.com consist of a white background with red, green or blue lettering that forms a pseudo-word – a sequence of characters that could probably be an English word, but isn't – of four to ten lowercase letters. To obfuscate the text, characters are squeezed, tilted and moved so that they touch each other, and the character sequence is arranged in a wave.

Mail.ru.⁴



Mail.ru is the biggest free Russian webmail provider. Image captchas from mail.ru consist of six blue outlines of letters and numbers on a white background. To obfuscate the text, characters are tilted, bent, moved up and down, and the entire background is covered with hundreds of other outlines of letters and numbers.

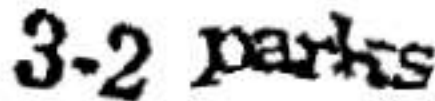
⁴The Mail.ru captcha is scaled at 0.5

Microsoft.



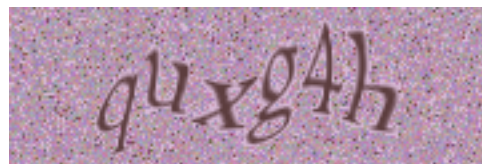
Live.com is the server for Windows Live IDs, the user accounts for Hotmail, MSN Messenger, Xbox LIVE, and other Microsoft sites. Image captchas from live.com consist of eight dark blue digits and uppercase letters on a gray background. To obfuscate the text, characters are squeezed, tilted and moved so that they touch each other, and the sequence is arranged in a wave.

Recaptcha.



Recaptcha.net provides image captchas to a large number of high profile sites, such as Facebook, Ticketmaster, and Craigslist. Captchas from recaptcha.net consist of two words (roughly 5-20 characters, according to the captcha answers we collected) in black on a white background. To obfuscate the text, the words are drawn from scanned books, where optical character recognizers failed on at least one of the words. Additionally, the characters of both words are squeezed into a wave-like shape.

Skyrock.



Skyrock.com is a social network and blogging site that is popular in many French speaking countries. Image captchas from skyrock.com consist of five to six dark colored digits and lowercase letters on a lighter colored background. To obfuscate the text, the characters are squeezed and the sequence is arranged in a wave.

Slashdot.



Slashdot.org is a website with user-submitted and editor-evaluated current affairs news, accompanied by forum-style comments. Image captchas from slashdot.org consist of a single English word of six to eight lower case letters in black on a white background. To obfuscate the text, some characters are filled while others are only outlines, a number of zig-zag lines are drawn over all the letters, and small black dots spot the background.

Blizzard.⁵



WorldOfWarcraft.com is the website for the popular online game World of Warcraft (WoW), run by Blizzard Entertainment. Image captchas from worldofwarcraft.com consist of six to eight bright colored letters on a darker patterned background. To obfuscate the text, characters are slightly tilted, squeezed and shifted up and down.

Yahoo.⁶



Yahoo.com is a popular search engine, portal and webmail provider. Image captchas from yahoo.com consist of five to eight black digits and upper or lowercase letters on a white background. To obfuscate the text, characters are squeezed, bent and moved so that they touch each other, and the character sequence is arranged in a wave.

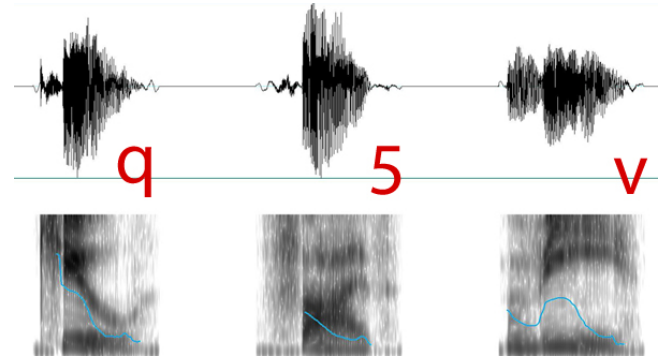
⁵The Blizzard captcha is scaled at 0.75

⁶The Yahoo captcha is scaled at 0.75

B. Audio captchas

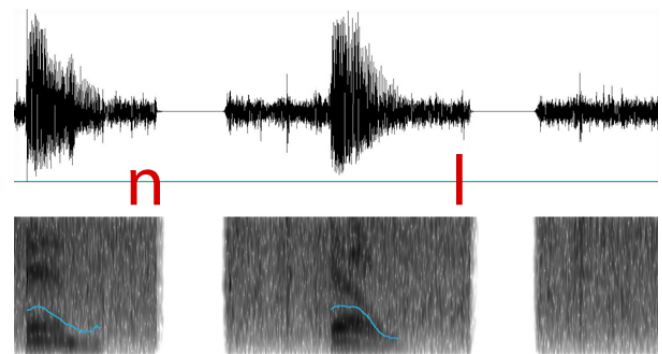
When available for the sites presented above, we also collected their audio captchas. As a result we also studied the following 8 audio captcha schemes.

Authorize.



Audio captchas on authorize.net consist of a female voice speaking aloud each of the five letters or digits of the image captcha. The voice clearly articulates each character, and there is minimal distortion. The example waveform and spectrogram are 3 second clips including the letters Q, 5 and V. These show that a good pause appears between each spoken character, and that the vowel formants (the thick black waves in the middle of the spectrogram, which are good indicators of both vowels and the surrounding consonants) are clearly visible in each word.

Digg.

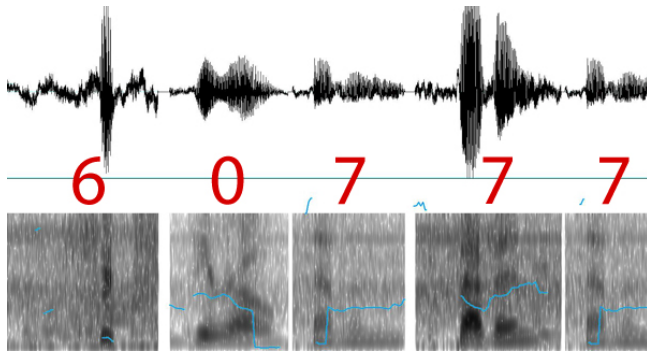


Audio captchas on digg.com consist of a female voice speaking aloud five letters. There is heavy white noise in the background, and sometimes an empty but louder segment is played between letters. The example waveform and spectrogram are 3 second clips including the letters N, L and an empty louder segment. The overall darkness of the spectrogram shows the heavy white noise which somewhat obscures the vowel formants.

Scheme	Authorize	Digg	eBay	Google	Microsoft	Recaptcha	Slashdot	Yahoo
Min len	5	5	6	5*	10	8	6	7
Max len	5	5	6	15*	10	8	8	7
Speaker	Female	Female	Various	Male	Various	Various	Male	Child
Charset	0-9a-z	a-z	0-9	0-9*	0-9	0-9	Word	0-9
Avg. duration	5.0	6.8	4.4	37.1	7.1	25.3	3.4	18.0
Sample rate	8000	8000	8000	8000	8000	8000	22050	22050
Beep	no	no	no	yes	no	no	no	yes
Repeat	no	no	no	yes	no	no	no	no

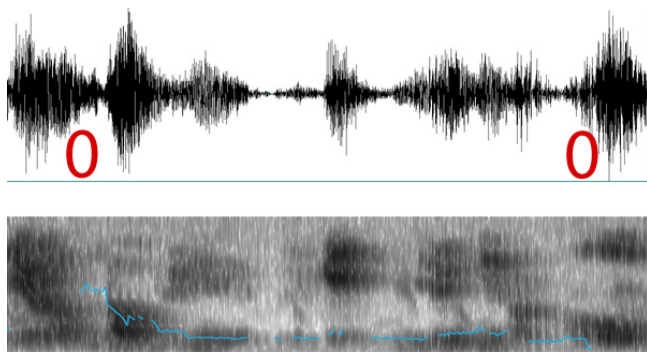
Table II
AUDIO CAPTCHA FEATURES

eBay.



Audio captchas on ebay.com consist of the same six digits from the image captcha being spoken aloud, each by a different speaker in a different setting. The example waveform and spectrogram are 3 second clips containing the digits 6, 0, 7, 7 and 7 - note that the digits in these captchas are delivered much faster than those of authorize.net or digg.com. The waveform shows the variability of the various digits due to different speakers and different background noise levels, and the spectrogram shows that the vowel formants are short and often obscured by the noise.

Google.

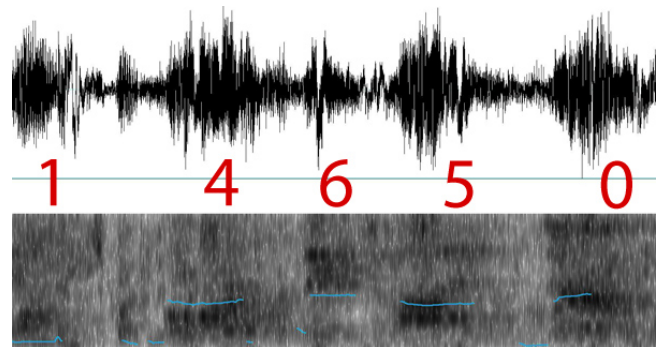


Audio captchas on google.com consist of three beeps, a male voice speaking digits aloud, the phrase “once again”, and a repeat of the male voice speaking the digits. In the background, various voices are playing simultaneously, and confusing decoy words like “now” or “it” are occasionally

interjected. The example waveform and spectrogram are 3 second clips containing the words “zero”, “it” and “oh” at the beginning, middle and end of the segment. The spectrogram shows how similar the formants from the background noise look to the true vowel formants, and of the three largest amplitude wave groups in the waveform, only the first and the last are actual words.

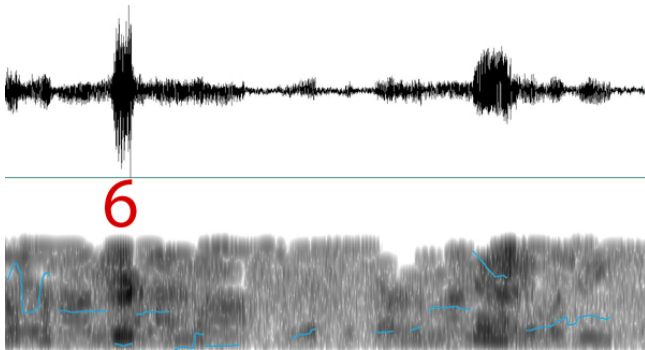
Note that because of the difficulty of this captcha, we were unable to say with confidence how many digits were presented, or even that the captcha was supposed to consist only of digits. Thus, the entries in Table II for Google are estimated from the answers we collected from our human subjects.

Microsoft.



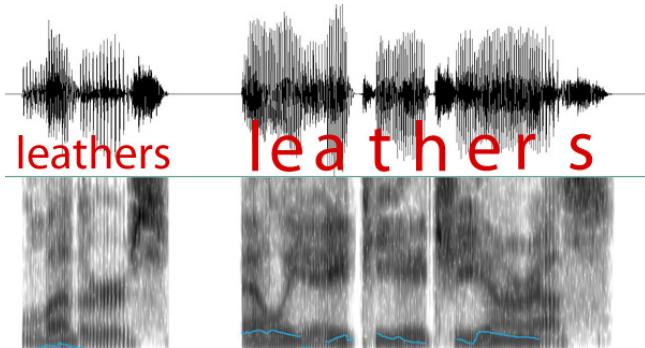
Audio captchas on live.com consist of ten digits being spoken aloud, each by a different speaker over a low quality recording, with various voices playing simultaneously in the background. The example waveform and spectrogram are 3 second clips containing the digits 1, 4, 6, 5 and 0 - like the eBay audio captchas, these digits are delivered quite fast to the user. While all the high amplitude sections of the waveform correspond to the actual digits, the spectrogram shows that the vowel formants are somewhat obscured by the background noise.

Recaptcha.



Audio captchas from recaptcha.net consist of eight digits spoken by different speakers, with voices in the background and occasional confusing words interjected. This is similar to the live.com presentation, but the digits are delivered much more slowly - the 3 second clip in the example waveform and spectrogram includes only the digit 6 and a confusing “eee” sound, with the next actual digit following about a second after the end of this clip⁷.

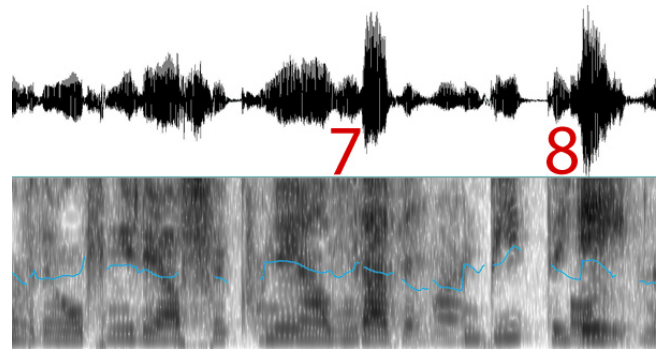
Slashdot.



Audio captchas from slashdot.org consists of a single word spoken aloud, followed by the same word spelled letter by letter. The speech is generated by a text-to-speech system, with a computer-generated male voice. The example waveform and spectrogram are the entire 3 second clip containing *Leathers*, L, E, A, T, H, E, R and S. The spectrogram shows that the vowel formants are very clear (not surprising, as they were computer-generated), though the speed of the letter-spelling speech in these captchas is among the fastest of all the audio captchas we surveyed.

⁷We used the default version supplied by the PHP API, but the recaptcha webpage suggests they also have another audio captcha scheme based on spoken words

Yahoo.



Audio captchas from yahoo.com consist of three beeps and then a child’s voice speaking seven digits with various other child voices in the background. The example waveform and spectrogram are a 3 second clip containing the digits 7 and 8. The digits are the largest amplitude sections on the waveform, though the spectrogram shows that the background voices look very much like the “real” speech.

IV. REAL WORLD USAGE : EBAY DATA

Before testing humans on our corpus of captchas, we first gathered some information about how captchas are used in the wild. With our collaborators at eBay, we gathered statistics on how captchas were used on their site over the course of a week in 2009, as shown in Table IV. Over the course of a week, eBay provided nearly **14 000 000** captchas to its users. Of these, over **200 000** were failed, suggesting that an average eBay user answers their captchas correctly **98.5%** of the time. Thus, in our study, we would expect to see our subjects with captcha solving accuracies in roughly this range.

Another interesting aspect of the eBay statistics is the usage of audio captchas. Of the **14 000 000** captchas, more than **100 000** were audio captchas, indicating that **0.77%** of the time, users prefer to have audio captchas over image captchas. This is actually a surprisingly large percentage compared to our expectations, and shows the importance of studying human understanding of both audio and image captchas.

V. EXPERIMENTAL RESULTS

From our corpus of captchas, we presented **5000** image captchas and **3500** audio captchas to Turkers, and **1000** image captchas to the underground service following our study methodology above. To allow for some analysis of human agreement on captcha answers, we had three subjects annotate each of the captchas, for both Turkers and the underground service. As a result we have annotated **195 000** image captcha and **84 400** audio captcha by the Turker and **39 000** captchas by the underground service. Overall this study is based on more than **318 000** captcha.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Total captcha	2079120	2041526	1902752	1825314	1825314	2178870	2104628
Audio captcha	15592	15476	14370	14482	14482	16412	16578
Failure	31709	30179	28475	28484	28484	33516	32564
Audio ratio	0.75%	0.76%	0.75%	0.79%	0.79%	0.75%	0.79%
Failure ratio	1.52512	1.47826	1.49652	1.56050	1.56050	1.56050	1.54726

Table III
EBAY CAPTCHA STATISTICS FOR A WEEK IN NOVEMBER 2009

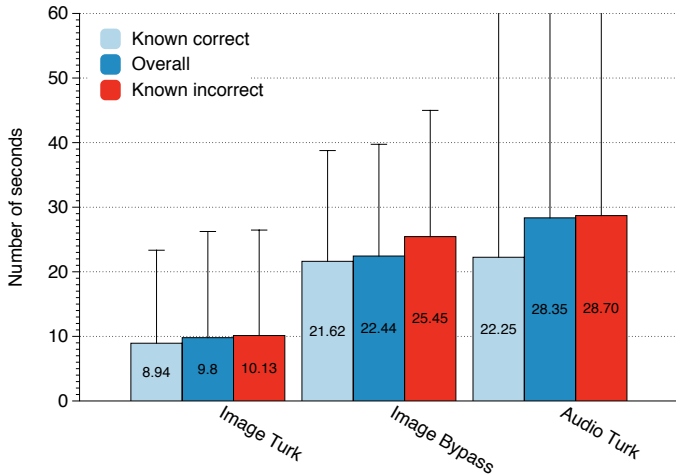


Figure 1. Average solving time for image and audio captchas, for Turkers and the underground service.

We focus on two primary research questions when analyzing the resulting data: how much inconvenience (user friction) does a typical captcha present to a user, and how different captcha schemes affect users with different backgrounds. The following sections explore these two questions in detail.

A. Captcha Friction

Captchas are meant to be a quick and easy verification that a user is in fact a human. Captchas that are time consuming or difficult for humans represent friction or inconvenience for the user. We consider some forms of captcha friction in the following sections.

1) *Solving Time*: One simple measure of friction is the total time a user spends on a given captcha, and Figure 1 shows these statistics.

Overall, we see that image captchas take Turkers about **9.8** seconds to view and solve, while audio captchas take Turkers about **28.4** seconds to hear and solve⁸. While 5-10 seconds is probably an acceptable inconvenience for a user

⁸The underground service takes around 22.44 seconds to solve image captchas, but we can only measure the turnaround time through their API, which may include some overhead for routing, etc.

when captchas are not frequently presented, the 20 or more seconds taken by audio captchas may present a substantial annoyance for users.

In the cases where we believe we know the correct answer for a captcha, we can compare the time it takes a user to give the correct answer with the time it takes them to give an incorrect answer. Across the Turkers, the underground service, and both image and audio captchas, we see that correct answers are given more quickly than incorrect answers, though this difference is most pronounced for audio captchas, where correct answers are **6.5 seconds faster** than incorrect ones. This is another argument for making sure that the captcha scheme used is sufficiently easy for humans, because the more they fail, the longer they'll be spending on the captchas.

It is worth noting that for all these mean solving times, the standard deviations are quite large. This is even true after we remove outliers greater than three standard deviations from the mean, as we do for all of our graphs. This is the standard flaw of taking timing measurements on the internet – people are always free to stop midway through and go do something else for a while if they like (even though the Turkers were explicitly requested not to).

Figures 2 and 3 show histograms of solving times for our users, demonstrating the very long tail of this distribution, but also showing that there is a clear clustering of most users in the 5-15 second range.

Figures 2 and 3 also show how the solving times differ for the different schemes. For image captchas, we see that mail.ru and Microsoft captchas take the most time for Turkers, with means around **13** seconds. The fastest image captchas were from Authorize, Baidu and eBay which take on average around **7** seconds each. In audio captchas, Google, reCaptcha and Yahoo captchas were the most time consuming, with means over **25** seconds, while Authorize, eBay and Slashdot all averaged **12** seconds or less. Note that these timings closely track the average duration of each scheme as shown in Table II, indicating that audio captcha solving time is dominated by the time spent listening to the captcha. These results suggest that careful selection of a captcha scheme can substantially reduce the friction of the captcha system on the user.

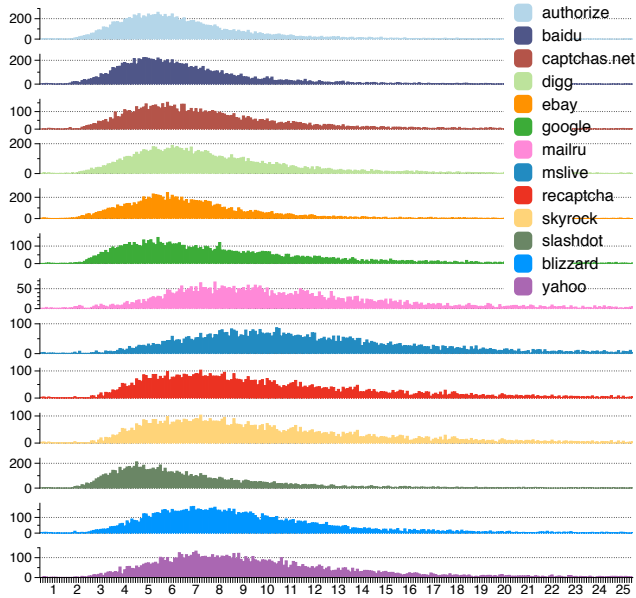


Figure 2. Solving times for image captcha schemes

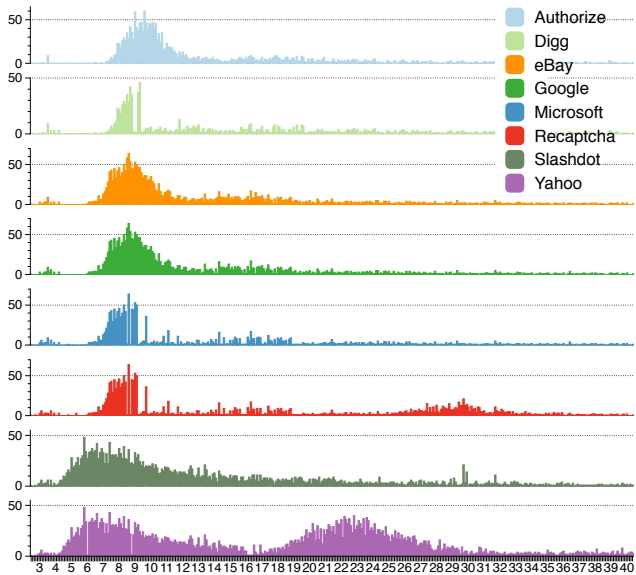


Figure 3. Solving times for audio captcha schemes

2) *Solving Agreement*: Another source of friction worth considering is the inconvenience of having to solve another captcha if a human can't guess the correct answer to the first one. Since we collected the captchas rather than generating them ourselves, we do not know the correct answer for each captcha, so we cannot determine with certainty when a user gets a captcha right or wrong. However, we can get an approximation of this by looking at the number of distinct answers given for each captcha. On a captcha that is easy for

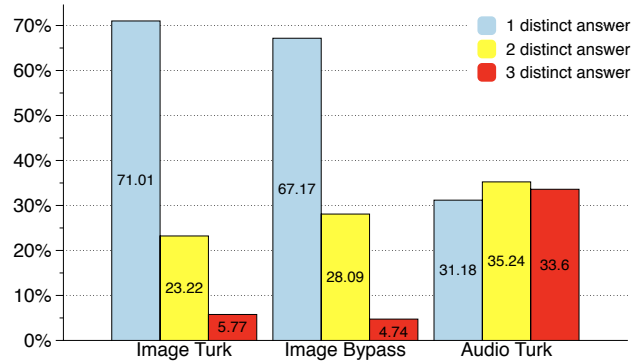


Figure 4. Percents of image and audio captchas given 1, 2 or 3 distinct answers

humans, all three subjects annotating the captcha will agree on the answer, while on a difficult captcha, subjects will disagree and more than one distinct answer will be given. Thus, the greater the number of distinct answers, the greater the disagreement and the harder the captcha.

Figure 4 shows the percent of image and audio captchas that were given 1, 2 or 3 distinct answers by our subjects. Both the Turkers and the underground service reach similar numbers for overall agreement on image captchas, with all three subjects agreeing on a single answer for around **70%** of all captchas. Only about **5%** of the image captchas were so bad that we got a different answer from each of the subjects.

Audio captchas are a total different story: All subjects agree on only **31.2%** of these captchas, and on **33.6%** everyone had a different answer. These results imply that many audio captchas are just too difficult for humans. As with solving time, we also see differences by scheme when looking at number of distinct answers. Figures 5, 6 and 7 show percents of distinct answers for each scheme. Authorize.net has the easiest image captchas for humans, with three subjects agreeing on the answer over **93%** of the time. On the other end of the spectrum is the extremely difficult mail.ru image captchas, which have perfect agreement among subjects less than **35%** of the time. On the audio side, Google, Microsoft, and Digg audio captchas are by far the most difficult, with all subjects producing the same answer only about **1%** of the time. Some of this difficulty with audio captchas may be because we give no instructions on what kinds of words or letters to expect, but this is consistent with for example the Google interface, which at the time of this writing was just a single button beside the image captcha text box that used Javascript to play the recording⁹. Again, we see that captcha friction on users can be substantially reduced by selecting a different captcha scheme.

⁹The alternate reCaptcha audio captchas, not tested in this study, which do not include intentional distortion may better agreement on these.

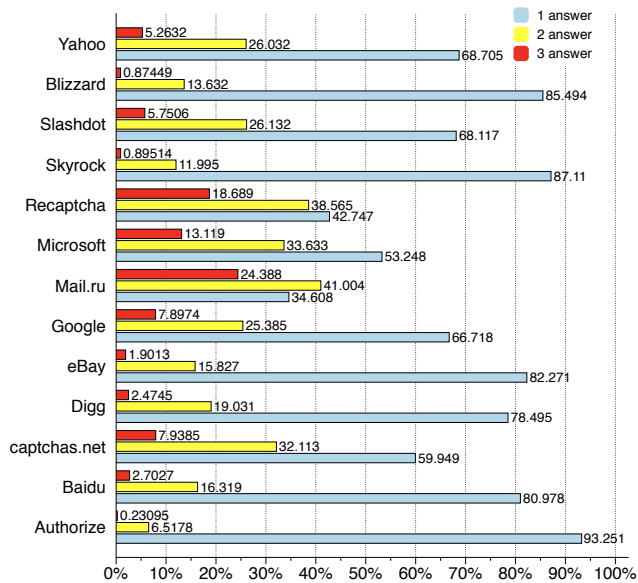


Figure 5. Percents of image captchas given 1, 2 or 3 distinct answers by Turkers for each scheme

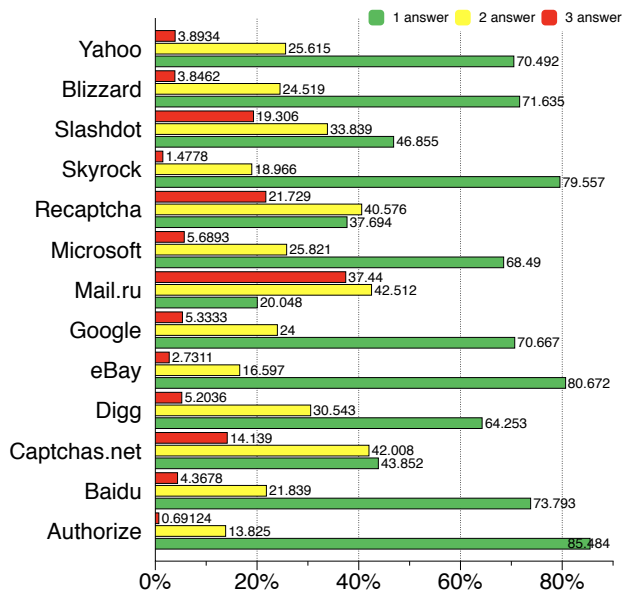


Figure 6. Percents of image captchas given 1, 2 or 3 distinct answers by the underground service for each scheme

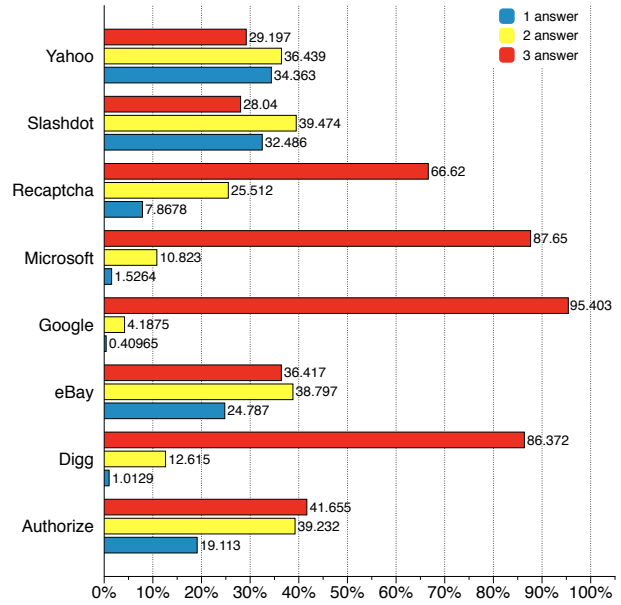


Figure 7. Percents of audio captchas given 1, 2 or 3 distinct answers by Turkers for each scheme

3) *Optimistic Solving Accuracy*: As a final measure of captcha friction, we wanted to calculate how often a human can expect to pass the captcha challenge. Again, we don't know all the answers for the captchas we collected, but we can make an approximation: if the three subjects produced the same answer, we assume that answer is correct. For two distinct answers we make the optimistic assumption that 2 out of 3 answers are correct and when the three users disagree make the optimistic assumption that 1 out of the 3 answers is correct.

Because our accuracy measurement is not perfect, looking at the absolute values may be misleading, but differences in solving accuracy in different scenarios should still reflect real differences.

Figure 8 shows the optimistic solving accuracy for image and audio captchas through both Mechanical Turk and the underground service. Even with our optimistic accuracy, the underground service achieves only **84%** accuracy on image captchas, while Turkers achieve **87%**. We see again that audio captchas are the most difficult, with Turkers solving them correctly only **52%** of the time. Figure 8 also compares the optimistic solving accuracy across different schemes. Among image captchas, both Turkers and the underground service are worst on the mail.ru captchas, achieving **61%** and **70%** accuracy respectively. The easiest captchas are the authorize.net captchas, where Turkers achieve **98%** accuracy and the underground service achieves **95%** accuracy. Note that these results mostly parallel what we saw in the analysis of solving time and solving agreement, where for example,

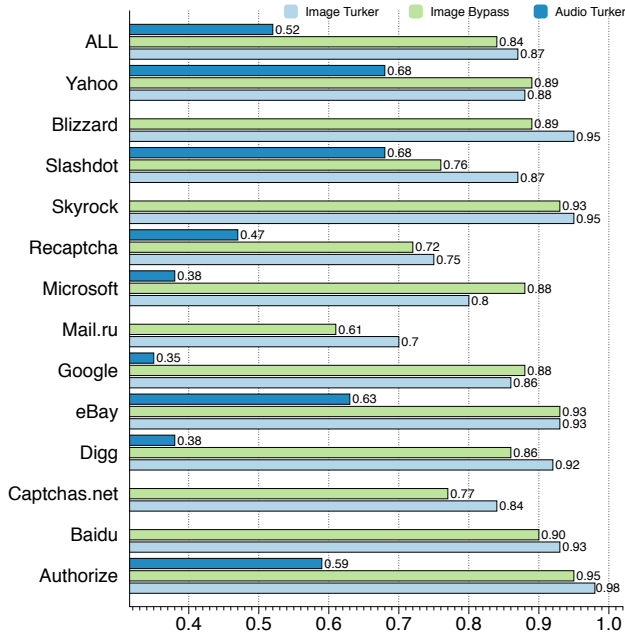


Figure 8. Optimistic solving accuracy for image and audio captchas, overall and for each scheme

live.com and mail.ru are also among the captcha schemes with the slowest responses and the worst agreement.

For eBay image captchas, we see an accuracy of **93%**, which is a little less than the **98.5%** found in the statistics collected by eBay. This is probably because on the eBay site the user may ask for a new captcha before solving it if they think the current one is too hard. That our approximate measurements are at least within a few points of the eBay measurement suggests that our solving accuracies are reasonable.

Figure 8 shows optimistic solving accuracy for all audio the schemes, in blue. Google is the hardest of these schemes, with Turkers achieving only **35%** accuracy, while slashdot.com and yahoo.com are the easiest schemes, with Turkers achieving **68%** accuracy for both of them. Overall, these numbers track closely our earlier results looking at solving agreement, and suggest the need for further research to investigate what makes an audio captcha easy or hard.

4) *Expected Solving Time*: The previous sections have showed a number of ways of measuring captcha friction on users. One way of unifying these all into a single measure is through expected solving time. The expected solving time of a captcha is the total amount of time a user should expect to spend, including not just the time to respond to the initial captcha, but also any time required to respond to additional captchas if the first is failed. Expected time can be measured by the following infinite summation, where t is the time it takes to answer a single captcha, and a is the user’s solving accuracy:

Scheme	Time	Accuracy	Expected time
Authorize	6.8	0.98	6.9
Baidu	7.1	0.93	7.6
Captchas.net	8.2	0.84	9.8
Digg	8.2	0.92	8.9
eBay	7.3	0.93	7.8
Google	9.7	0.86	11.3
mail.ru	12.8	0.7	18.3
Microsoft	13	0.8	16.3
Recaptcha	11.9	0.75	15.8
Skyrock	7.9	0.95	8.3
Slashdot	7.7	0.87	8.8
Blizzard	9.3	0.95	9.8
Yahoo	10.6	0.88	12
Authorize audio	11.9	0.59	20.2
Digg audio	14.8	0.38	39
eBay audio	11.8	0.63	18.8
Google audio	35.2	0.35	100.6
Microsoft audio	16.6	0.38	43.8
Recaptcha audio	30.1	0.47	64.1
Slashdot audio	11.7	0.68	17.2
Yahoo audio	25	0.68	36.8

Table IV
EXPECTED SOLVING TIME FOR EACH SCHEME

$$\begin{aligned}
 est(t, a) &= t + (1 - a)(t + (1 - a)(t + (1 - a)(t + \dots))) \\
 &= t + t(1 - a) + t(1 - a)^2 + \dots \\
 &= \frac{t}{a}
 \end{aligned}$$

Essentially, we always pay the cost of the initial captcha, and then we pay a fractional cost for additional captchas that is weighted by our chance of answering incorrectly. As an example of expected solving time in action, a user solving a Microsoft audio captcha, which takes on average **13** seconds, and on which users have optimistic solving accuracy of **0.8**, the expected solving time is actually **16.3** seconds **25%** longer than the single captcha time would suggest. Table IV shows optimistic solving accuracies for all captcha schemes.

VI. CAPTCHAS AND USER BACKGROUND

Having demonstrated that captchas are often quite difficult for humans, we turn to the question of why. Certainly the various distortion methods used to to increase captcha difficulty for computers play a role, but in this study we focus on the characteristics of the people, not the captchas, that predict captcha difficulty. Thus, we rely primarily on the demographic data we collected from Turkers to investigate these questions.

Overall we had more than **11800** demographic surveys completed by Turkers. Since we authorized Turkers to complete up to five image and audio tasks for us, it is likely that some of these surveys are duplicates. However even in the worst case, we had more than 1100 different people answering our questions.

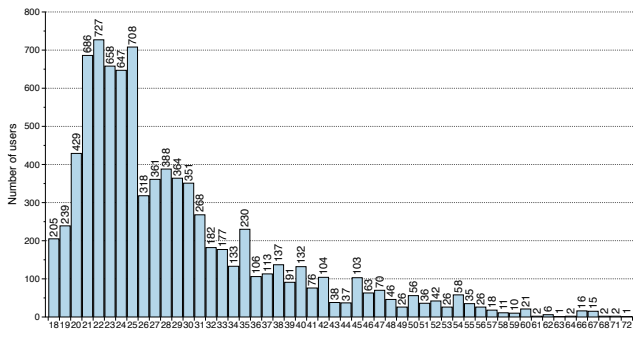


Figure 9. Mechanical Turk worker age distribution

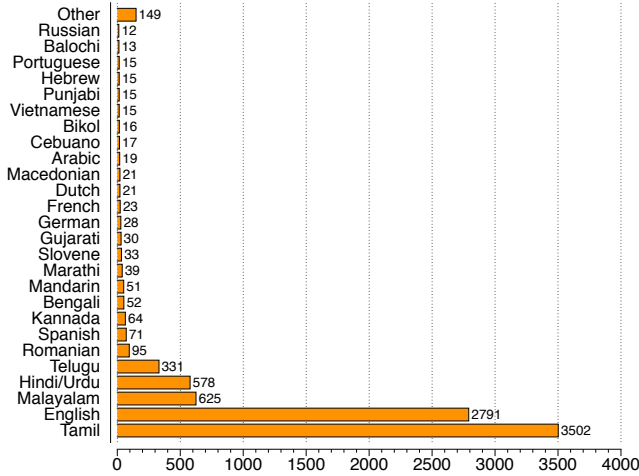


Figure 10. Mechanical Turk worker native language

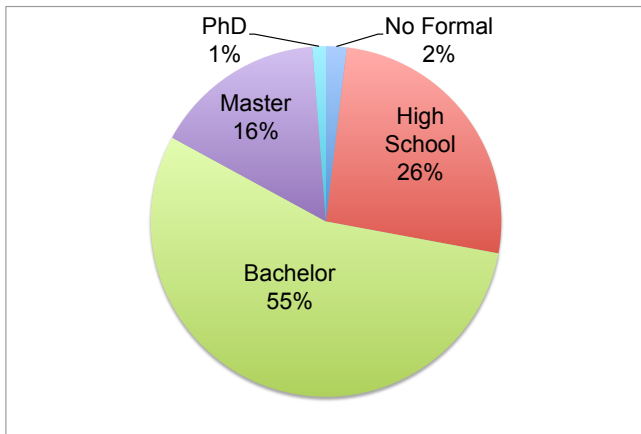


Figure 11. Education repartition.

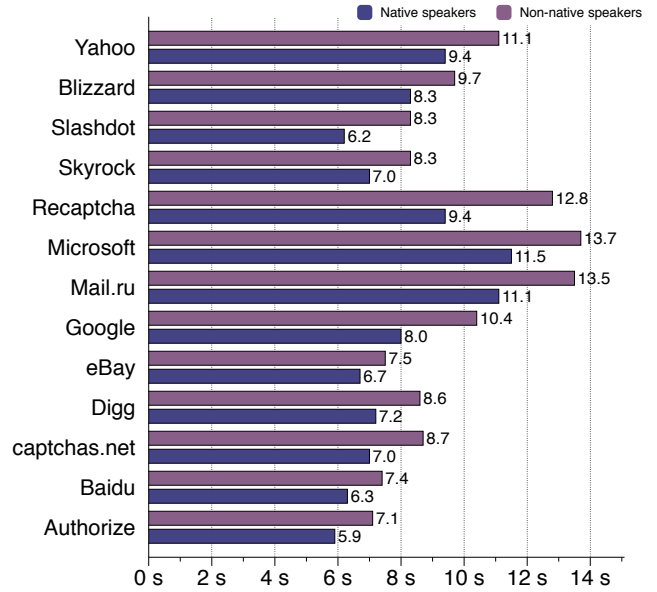


Figure 12. Average image captcha solving time for native and non-native speakers of English.

First, it is useful to get a general picture of the Turkers that worked for us. Figure 9 shows the age distribution of our workers - the average age was around 29 years, and almost half of our workers were 25 or younger. Figure 10 shows the native language distribution - Tamil and English accounted for **40%** and **32%** of the languages, respectively, leaving only **28%** distributed to all other languages. Finally, Table 11 shows the distribution of education levels showing that over **72%** of our workers had a bachelor's degree or higher. Overall, our Turkers represent a young, educated, primarily Tamil and English speaking population.

Our first question was about the effect of native language: can non-native speakers of English perform as well as native speakers on captchas? Figures 12 and 13 show the solving time and solving accuracy for native and non-native speakers, broken down by image captcha scheme. Overall on image captchas, native speakers are substantially faster and slightly more accurate. Looking across the schemes, the schemes that use real words, like recaptcha, and pseudo words, like Google, are solved far faster (up to **30%**) by native English speakers. This last point is important as it suggests that captcha schemes that rely on extensive experience with a single language can be biased against users who are not native speakers of that language.

Considering audio captchas, we still see that non-native speakers of English are usually somewhat slower than native speakers, though for reCaptcha the times are roughly comparable. As for image captchas, solving accuracy is lower across the board for non-native speakers. Moreover, the language bias is again illustrated by the the Slashdot

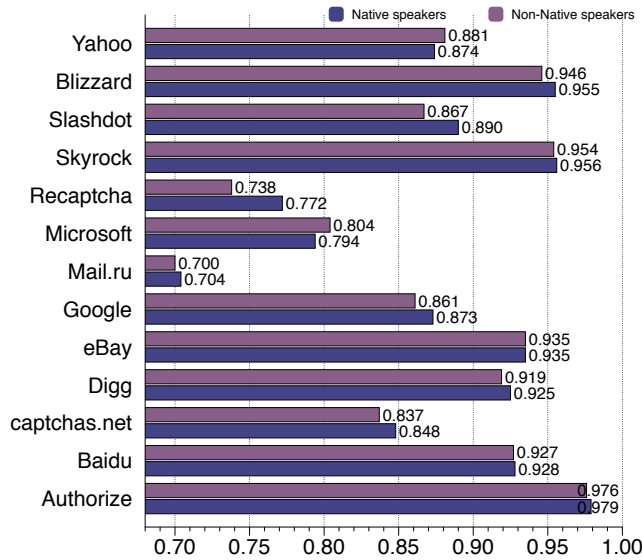


Figure 13. Average image captcha optimistic solving accuracy for native and non-native speakers of English.

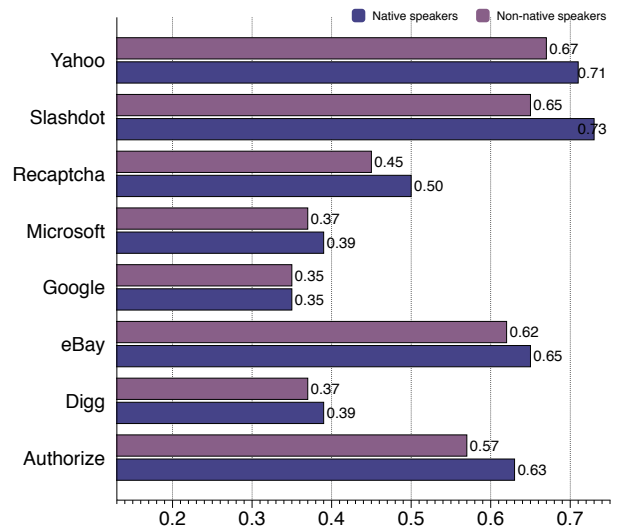


Figure 15. Average audio captcha optimistic solving accuracy for native and non-native speakers of English.

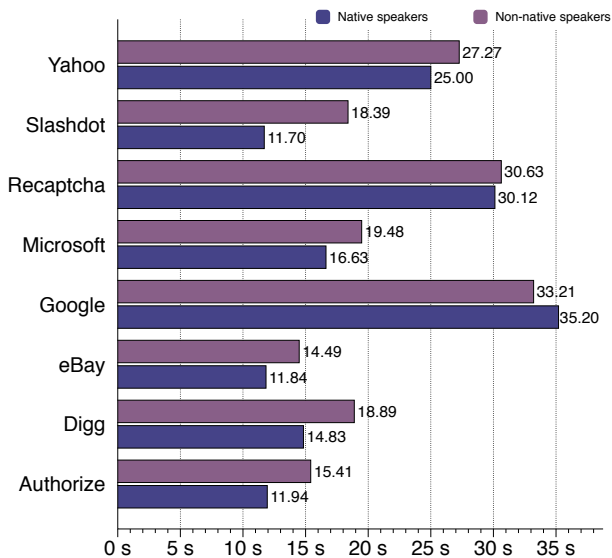


Figure 14. Average audio captcha solving time for native and non-native speakers of English.

audio captchas, which are based on spelling English words: the solving time for non-native speakers is 18.39 seconds, 7.4 seconds longer (57%) than the expected solving time for native speakers. Thus, some significant work needs to be done to make audio captchas reasonable for globally deployed web applications.

We also investigated questions of aging: how does performance differ for old and young users? Figures 16 and 17 show that there is a lot of variation in both solving

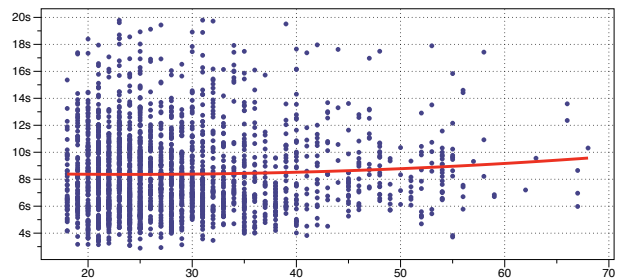


Figure 16. Solving time for Turkers of different ages.

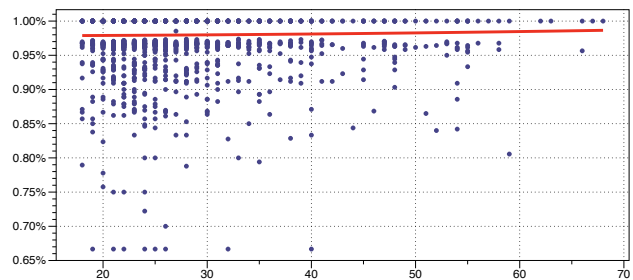


Figure 17. Optimistic solving accuracy for Turkers of different ages.

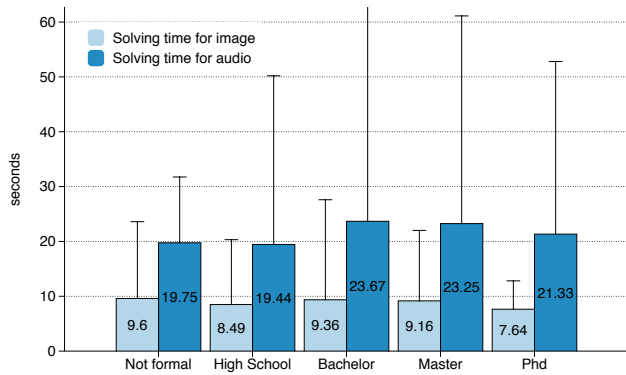


Figure 18. Solving time for Turkers of different education levels.

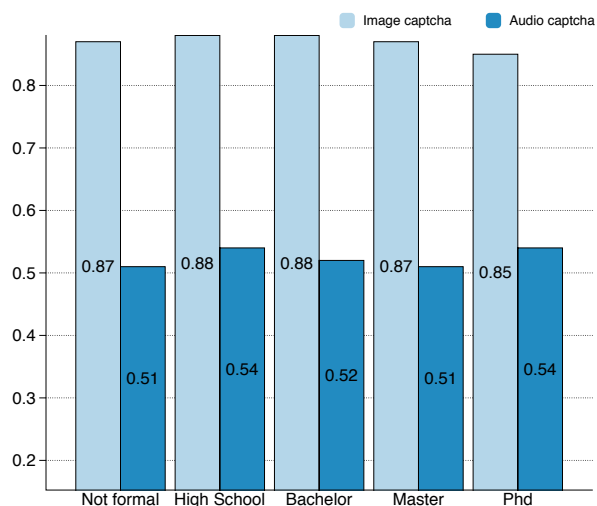


Figure 19. Optimistic solving accuracy for Turkers of different education levels.

time and solving accuracy for users of all ages. However there are small trends visible in these graphs - each year, users slow down by about 0.01 seconds, and become more accurate by about 0.01%. These findings are in line with some psychological research on aging [15] where older people are found to be more accurate than younger subjects who demonstrate greater speed but more errors.

Finally, we looked into the effect of education: do users with more or less education have an advantage in solving captchas? Figure 18 shows solving time for Turkers with different levels of education. There is a small decrease in solving time of image captchas as higher levels of education are obtained, starting at **9.6** seconds for Turkers with no formal education, and dropping to **7.64** seconds for Turkers with a Ph.D. With audio captchas, on the other hand, more education doesn't seem to make people faster: Turkers with only a high school education were faster than Ph.D.s, and Turkers with bachelors or masters degrees were substantially

slower.

Of course, we are interested not just in speed, but also in accuracy, and so Figure 19 shows optimistic solving accuracy for Turkers with different levels of education. Overall the correlation between having a higher education and being more efficient at solving captchas does not seem to hold in our data. However, we are relying on self reported education levels in this work, and it is possible that there is therefore noise in our data as some Turkers may perceive reporting a higher education level as an opportunity to get offered better tasks on Mechanical Turk.

VII. ADDITIONAL RELATED WORK

The closest work to our is that of [3] where they looked at usability issues in presenting audio captchas to humans. Though they only used 10 captchas from each of 10 sites, they also found audio captchas to be more time consuming and difficult than image captchas, and introduced a new user interface to make audio captchas more accessible.

The first discussion of the captcha idea appears in [22], though the term CAPTCHA was coined in [24]. Text/image based captchas have been studied extensively [6], [18], [19] and there is a long record of successful attempts at breaking captchas of popular sites [8]. For example in March 2008, a method to break 60% of MSN visual captchas was disclosed [26]. One of the most famous visual captcha breakers is PWNtcha [16]. In [14], Hernandez-Castro and al use a side channel attack to break labeling captcha. In [12], Golle use learning attack to break the Asirra scheme. Tam and his colleagues [23] built a breaker for audio captchas from google.com, digg.com and an older version of recaptcha.net.

In [4], Cain et. al. studied the relation between captchas and network security. In [25], Yan et. al. used naive pattern recognition to break image captchas. In [1], Athanasopoulos et. al. used animation as captchas. A comparison of human and computer efficiency to recognize single characters was explored in [5]. Many ways of building captchas have been proposed [7], [9]–[11], [17]–[19] Finally, many other techniques have been used to break captchas [20], [21], [27]

VIII. CONCLUSION

We have presented a large scale study of how much trouble captchas present for humans. We collected **5000** captchas from each of **13** most widely used image captcha schemes and **3500** captchas from the **8** most widely used audio captcha schemes, and had them each judged by multiple human subjects from Amazon's Mechanical Turk and an underground captcha-breaking service. Overall, we found that captchas are often harder than they ought to be, with image captchas having an average solving time of **9.8** seconds and three-person agreement of **71.0%**, and audio captchas being much harder, with an average solving time of **28.4** seconds, and three-person agreement of **31.2%**. We also found substantial variation in captcha difficulty across

schemes, with authorize.net image captchas being among the easiest, and google.com audio captchas being among the hardest. We observed that the workers from Mechanical Turk were quicker and more accurate than those from the underground service, and were also willing to solve captchas for smaller amounts of money.

Using the data collected from Amazon's Mechanical Turk, we identified a number of demographic factors that have some influence on the difficulty of a captcha to a user. Non-native speakers of English were slower, though they were generally just as accurate unless the captcha required recognition of English words. We also saw small trends indicating that older users were slower but more accurate. These results invite future research to more deeply investigate how individual differences influence captcha difficulty.

REFERENCES

- [1] E. Athanasopoulos and S. Antonatos. Enhanced captchas: Using animation to tell humans and computers apart. In *IFIP International Federation for Information Processing*, 2006.
- [2] H.S. Baird and T. Riopka. Scattertype: a reading captcha resistant to segmentation attack. In *IS & T/SPIE Document Recognition & Retrieval Conference*, 2005.
- [3] Jeffrey P. Bigham and Anna C. Cavender. Evaluating existing audio captchas and an interface optimized for non-visual use. In *ACM Conference on Human Factors in Computing Systems*, 2009.
- [4] A. Caine and U. Hengartner. *The AI Hardness of CAPTCHAs does not imply Robust Network Security*, volume 238.
- [5] K. Chellapilla, K. Larson, P.Y. Simard, and M. Czerwinski. Computers beat humans at single character recognition in reading based human interaction proofs (hips). In *CEAS*, 2005.
- [6] K Chellapilla and P Simard. Using machine learning to break visual human interaction proofs. In MIT Press, editor, *Neural Information Processing Systems (NIPS)*, 2004.
- [7] M. Chew and H.S. Baird. Baffletext: a human interactive proof. In *10th SPIE/IS&T Doc. Recog. Retr. Conf, DRR 2003*, 2003.
- [8] Dancho Danchev. Microsoft's captcha successfully broken. blog post <http://blogs.zdnet.com/security/?p=1232>, May 2008.
- [9] R. Datta. Imagination: A robust image-based captcha generation system. In *ACM Multimedia Conf.*, 2005.
- [10] Anne Eisenberg. New puzzles that tell humans from machines. http://www.nytimes.com/2009/05/24/business/24novelties.html?_r=1&ref=technology, May 2009.
- [11] J. Elson, J.R. Douceur, J. Howell, and J. Saul. Asirra: A captcha that exploits interest-aligned manual image categorization. In *4th ACM CCS*, 2007.
- [12] P. Golle. Machine learning attacks against the asirra captcha. In *ACM CCS 2008*, 2008.
- [13] C.J. Hernandez-Castro and A.: Ribagorda. Pitfalls in captcha design and implementation: the math captcha, a case study. <http://dx.doi.org/10.1016/j.cose.2009.06.006>, 2009.
- [14] C.J. Hernandez-Castro, A. Ribagorda, and Y. Saez. Side-channel attack on labeling captchas. <http://arxiv.org/abs/0908.1185>, 2009.
- [15] Terence M. Hines and Michael I. Posner. Slow but sure: A chronometric analysis of the process of aging. In *The 84th Annual Convention of the American Psychological Association*, 1976.
- [16] Sam Hocevar. Pwntcha captcha decoder. web site, <http://sam.zoy.org/pwntcha>.
- [17] M.E. Hoque, D.J. Russomanno, and M. Yeasin. 2d captchas from 3d models. In *IEEE SoutheastCon 2006*, 2006.
- [18] P Simard K Chellapilla, K Larson and M Czerwinski. Building segmentation based human- friendly human interaction proofs. In Springer-Verlag, editor, *2nd Int'l Workshop on Human Interaction Proofs*, 2005.
- [19] P Simard K Chellapilla, K Larson and M Czerwinski. Designing human friendly human interaction proofs. In ACM, editor, *CHI05*, 2005.
- [20] G. Mori and J. Malik. Recognizing objects in adversarial clutter: Breaking a visual captcha. In *CVPR 2003*, pages 134–144, 2003.
- [21] G. Moy. Distortion estimation techniques in solving visual captchas. In *CVPR 2004*, 2004.
- [22] Moni Naor. Verification of a human in the loop or identification via the turing test. Available electronically: <http://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.ps>, 1997.
- [23] Simsa Tam, J., S. J., Hyde, and L. Von Ahn. Breaking audio captchas. http://www.captcha.net/Breaking_Audio_CAPTCHAs.pdf.
- [24] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In Springer, editor, *Eurocrypt*, 2003.
- [25] J. Yan and A.S.E. Ahmad. Breaking visual captchas with naive pattern recognition algorithms. In *ACSAC 2007*, 2007.
- [26] Jeff Yan and Ahmad Salah El Ahmad. A low-cost attack on a microsoft captcha. Ex confidential draft http://homepages.cs.ncl.ac.uk/jeff.yan/msn_draft.pdf, 2008.
- [27] H. Yeen. Breaking captchas without using ocr. http://www.puremango.co.uk/2005/11/breaking_captcha_115/.