Blind and Human: Exploring More Usable Audio CAPTCHA Designs

Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das, *Georgia Institute of Technology*

https://www.usenix.org/conference/soups2020/presentation/fanelle

This paper is included in the Proceedings of the Sixteenth Symposium on Usable Privacy and Security.

August 10–11, 2020

978-1-939133-16-8

Open access to the Proceedings of the Sixteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

Blind and Human: Exploring More Usable Audio CAPTCHA Designs

Valerie Fanelle
Georgia Institute of Technology

Sepideh Karimi
Georgia Institute of Technology

Aditi Shah
Georgia Institute of Technology

Bharath Subramanian
Georgia Institute of Technology

Sauvik Das
Georgia Institute of Technology

Abstract

For people with visual impairments (PVIs), audio CAPTCHAs are accessible alternatives to standard visual CAPTCHAs. However, current audio CAPTCHA designs are slower to complete and less accurate than their visual counterparts. We designed and evaluated four novel audio CAPTCHAs that we hypothesized would increase accuracy and speed. To evaluate our designs along these measures, we ran a three-session, within-subjects experiment with 67 PVIs from around the world — the majority being from the U.S. and India. Thirty three participants completed all three sessions, each separated by one week. These participants completed a total of 39 distinct audio CAPTCHA challenges across our prototype designs and the control, all presented in random order. Most importantly, all four of our new designs were significantly more accurate and faster than the control condition, and were rated as preferable over the control. A post-hoc security evaluation suggested that our designs had different strengths and weaknesses vis-a-vis two adversaries: a random guessing adversary and a NLP adversary. Ultimately, our results suggest that the best design to use is dependent on use-context.

1 Introduction

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) are commonly used online to differentiate between human users and non-human bots [22]. In doing so, many CAPTCHAs ask users to engage in visual-processing tasks that are simple for humans, yet difficult for bots [20]. However, these visual-processing tasks are inaccessible to the 285 million people with visual impairments (PVIs) worldwide — 39 million of whom are totally blind, and 246 million who have low vision [35]. Instead, PVIs rely on audio CAPTCHAs, which aim to differentiate humans from bots using acoustic processing tasks.

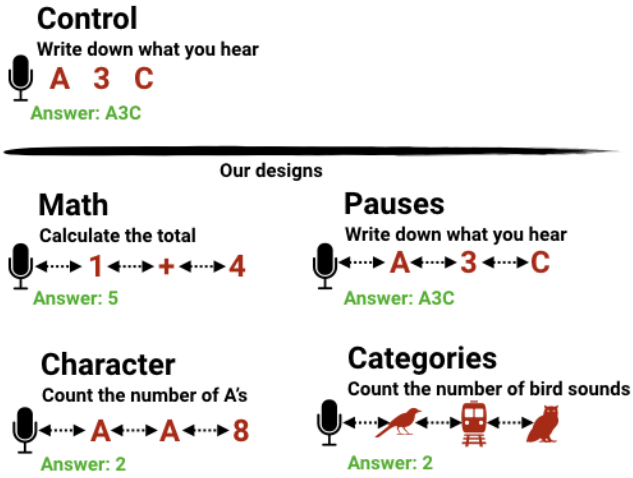

Figure 1: We designed, implemented and evaluated four novel audio CAPTCHAs. The Math prototype asked users to calculate a running total; the Character prototype asked users to count the occurrence of a character in an alphanumeric series; the Pauses prototype asked users to transcribe the alphanumeric characters they heard, but it incorporated longer pauses between characters; and, the Categories prototype, asked users to count the number of sounds, in a series, that belonged to a certain category.

In their current state, audio CAPTCHAs are significantly less usable than their visual counterparts [4, 10, 25, 37]. While visual CAPTCHAs take 9.8 seconds to solve with a 93% success rate, on average, audio CAPTCHAs take 51 seconds to solve with a 50% success rate [5, 10, 21, 33]. This

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

difference in speed and accuracy occurs because existing audio CAPTCHAs are modelled after their visual counterparts rather than using designs specific to the audio medium [5, 10, 11]. As such, existing audio CAPTCHAs require impractical levels of attention and memory-capacity from the users who depend on them [5]. This means that visual CAPTCHAs are not an equally challenging alternative to the audio CAPTCHA design; audio CAPTCHAs are more problematic for PVIs than visual designs are for fully-sighted people [17, 28, 32, 34].

Audio interference is one of the biggest issues that users face with existing audio CAPTCHAs [3]. For example, many PVIs rely on screen readers to help navigate user interfaces. When these users start typing the characters they hear in a CAPTCHA challenge, their screen reader software will read each typed letter out loud while they are simultaneously listening for the next character in the challenge. The audio conflict between the typed letter and the spoken letter thus creates unnecessary user frustration and errors. Owing to these frustrations, in a 2017 global study by WebAIM, of the 1,792 PVIs surveyed, 90% ranked audio CAPTCHAs as somewhat or very difficult [34]. These respondents also ranked CAPTCHAs as the second most problematic daily issue on the web, after Adobe Flash. The goal of our paper is to offer insights and designs that bridge the usability gap between audio and visual CAPTCHAs.

Informed by this prior work, as well as the personal experiences of one of the authors, who is blind, we followed an iterative design process to prototype and refine four new audio CAPTCHAs (see Figure 1). The Math prototype asked users to perform simple addition and subtraction; the Character prototype asked users to count the occurrence of a specific character within a string of alphanumeric characters; the Pauses prototype, which is a variation of existing alphanumeric audio CAPTCHA designs, asked users to transcribe the alphanumeric characters they heard but incorporated longer pauses between characters to minimize screen reader interference; and, the Categories prototype, asked users to count the number of sounds, in a series, that belonged to a certain category (e.g., bird chirps, baby cries).

To evaluate these designs, we were guided by three research questions — How do our novel audio CAPTCHAs compare to existing audio CAPCTHAs in terms of: **(RQ1)** task performance metrics such as accuracy and speed? **(RQ2)** security against common attacks (e.g., random guessing, machine-learning based audio classification)? and, **(RQ3)** self-reported and heuristic measures of usability?

To answer these research questions, we conducted a multi-session, within-subjects online experiment. We recruited 67 PVIs from around the world — 38 of whom live in the USA, 22 in India, 2 in Italy, 2 in Germany, 2 in the Czech Republic, and 1 in South Africa. Of the 67 PVIs, 33 participated in all three study sessions. In total, through three time-separated sessions, we asked participants to complete nine iterations of each of our four new prototypes. We recorded their accuracy and completion times with each challenge. Upon completing each challenge, we also had participants complete a brief questionnaire to gauge their in-the-moment reactions to our designs. Through a series of quantitative, qualitative and heuristic analyses on these data, we found that all of our prototypes showed statistically significant improvements in accuracy and completion time, were rated high on subjective and heuristic measures of usability and satisfaction, and were preferred over standard alphanumeric CAPTCHAs.

We also evaluated the security of our prototypes against two threat models: a random guessing adversary and an NLP adversary that leverages commercially available, state-of-the-art speech-to-text recognition and audio event classification. The control condition and our Pauses prototype offered the greatest security against random guessing attacks, but our Categories and Math prototypes offered the greatest resilience against the NLP adversary.

While all of our prototypes outperformed the control in most measures, no single design stood out as the best. The Math prototype was the most accurate, the second fastest, and provided reasonable security against both adversaries. The Character prototype was rated the most usable and satisfying, but was vulnerable against random guessing attacks. The Categories prototype was the most vulnerable against random guessing attacks, but was the fastest and most globally accessible — an important peripheral consideration, given that there are PVIs from various continents, countries, and cultural backgrounds [8, 9, 26, 35, 36]. Finally, the Pauses prototype was most preferred over the control condition, but was second lowest in accuracy and the slowest of our new designs.

2 Related Work

2.1 Challenges with Audio CAPTCHAs

In 2009, researchers at the University of Washington did a large-scale user study with 162 PVIs and found ten existing audio CAPTCHA designs to be difficult and time-consuming. They reported a 39% to 43% success rate for solving such designs on the first try and asserted that audio interfaces are often not direct translations of visual interfaces [5].

Prior work suggests that there are two types of audio CAPTCHAs: content-based and rule-based [19]. Content-based challenges require users to convert the speech of an audio file to text, an example of which is the existing alphanumeric standard. Alternatively, rule-based challenges ask users to interpret information they are hearing (e.g., 'count the number of times you hear the sound of an animal'). Rule-based CAPTCHAs can reduce the burden on short-term memory, because one only needs to remember a running total [19].

Sauer et al. studied the effects of content-based designs that closely resemble the current design norm for both visual and audio CAPTCHAs. In this study they played eight

numbers in distorted voices and asked users to input these numbers in sequence. However, they found that this technique disproportionately placed too high a cognitive load on PVIs, requiring them to either memorize the CAPTCHA series or use external tools to quickly note the entities they have heard. Due to a success rate of 46% and long average times of task completion (65.64 sec), these content-based designs exhibited low usability [13]. To address these concerns, we created low short term cognitive load CAPTCHAs that ask users to remember only one or two entities at a time. We accomplished this via rule-based designs and eliminating audio interference.

Furthermore, researchers have evaluated CAPTCHAs that employ text-based mathematical methods that ask questions such as, "What is the sum of two and four?" [17]. This text-based design is insecure due to the advancement of Natural Language Processing (NLP)-based bots [19]. Compounded with the open source tools available to adversaries online, there remains a need to create usable audio CAPTCHAs that are at least as strong as standard designs [7, 16, 30].

2.2 Improving Audio CAPTCHAs

Interesting innovation is occurring in this field. For instance, Soupionis et al. [31] implemented an audio CAPTCHA prototype for SIP-based Voice over IP (VoIP). However, their design was only tested with sighted users, raising concern about real-world outcomes for PVI's. [35].

Gao, Haichang, et al. [14] also designed a secure audio CAPTCHA that requires humans to read a sentence aloud in a natural voice and asked participants to differentiate the human voices from bots', but also only tested their design with sighted participants.

In another study [15], researchers proposed two alternative CAPTCHA designs: "auditory CAPTCHAs" and "non-nonsense CAPTCHAs," both of which were evaluated for both usability and security using Google's open-source reCAPTCHA technology. Their results showed that when comparing the accuracy levels of both humans and speech recognition algorithms, human success rates are 2.8 - 3.9 times higher. From their findings, Hendrick et al. concluded that all existing CAPTCHAs will eventually be broken, so future research should focus on incorporating human cognition as best as possible. Examples of this approach includes the use of deductive reasoning, sensory skills, and/or problem solving in order to answer correctly. This result motivated our pursuit of rule-based audio CAPTCHA designs.

Finally, Lazar et al. evaluated audio CAPTCHA designs that test sound category identification: e.g., identifying a sound clip as coming from a trumpet, a lion roaring, or a baby crying [18]. They achieved $\geq 90\%$ accuracy. However, they tested their designs with PVIs in a controlled environment with no baseline condition and with twenty participants all from the same location. This work inspired our Categories prototype, which we evaluate more broadly.

Prototype	Instructions	Sample Challenge	Correct Answer
Control (Content-Based)	Record each letter or number you hear.	8G6JVF	8G6JVF
Math (Rule-Based)	After you press play, please perform all of the calculations and provide one total at the end.	7+4-2-1	8
Character (Rule-Based)	Count the number of times '6' is spoken. Type the sum in the text box.	6R169Y6	3
Pauses (Content-Based)	Record each letter or number that you hear.	010J14	010J14
Categories (Rule-Based)	Count the number of times you hear sounds associated with those made by birds.	robin, train, motor, owl, rooster	3

Table 1: High-level summary of the prototype challenges we tested on our participants.

3 Our Audio CAPTCHA Designs

In exploring the design space for usable audio CAPTCHAs, our high-level design goals were to create challenges that minimize audio interference, reduce cognitive load (e.g., the amount of information in short-term memory), use no more than basic knowledge, and are robust against random guessing and off-the-shelf Natural Language Processing (NLP) tools.

Table 1 provides a high-level overview of each of our prototypes, along with example challenges and their corresponding correct answers. Of the four new CAPTCHAs we designed, three were rule-based in light of the aforementioned design goals. To ensure a baseline level of security in creating challenges for each of these prototypes, we followed advice from prior research to perturb the raw audio of the challenges [37]. The challenges we created contained variations in speed (very slow, slow, normal, fast, and very fast), pitch (male and female), and type of background noise. The background noises varied from public spaces (e.g., cafes), to the sounds of planes and wind. These challenges thus incorporated high randomness to complicate speech-to-text attacks.

3.1 Math Prototype

Our first prototype is a rule-based design that challenged users with basic addition and subtraction problems; some were mixed and others were exclusively focused on addition or subtraction. An example "Math" prototype challenge would be: "8 plus 4 minus 2 subtract 1 add 9" with two second gaps after each element. At each step, the user would only need to cognitively keep track of the running totals: 8, then 12, then 10, then 9, then 18. The need to do on-demand calculations might be challenging, but by keeping the operands within single digits and limiting the total number of operations, we hypothesized that the challenge would be easier and faster than the baseline control, owing to its reduced memory burden and single value entry.

3.2 Character Prototype

Our second prototype asked users to count the number of times they heard one specific character in a string of random letters and numbers. For instance, in one such challenge we asked people to identify the number of times they heard the character "s" within the string "3sjkS49sxo" — the answer being 3. Each character was read aloud with one second gaps in between. Similar to the Math prototype, we hypothesized that this design would result in greater accuracy and faster input completion speeds. This was due to the reduced cognitive demand of the need to keep track of just one running total and entering in only one input at the end of the recording.

3.3 Pauses Prototype

Our third prototype was a slight modification of the standard, content-based, alphanumeric CAPTCHAs that ask users to type, in sequence, all the characters heard in a random string. The key difference is that we included a two second pause in between characters to mitigate the interference between screen reader transcription and the challenge characters screen readers read aloud. This design should be relatively simple to deploy given its similarities to existing audio CAPTCHAs.

3.4 Categories Prototype

Our final prototype asked users to count the number of times they heard a certain "category" of sound (e.g., bird chirps, cars honks) embedded in a string of other sounds. Each sound was separated by a two second gap. For example, a user might have been asked to identify the number of times they heard birds chirping within a stream of sounds like trains and vehicular motors. The user answered with the total number of bird sounds detected throughout the CAPTCHA. Similar to the first two rule-based designs, due to reduced cognitive load, we expected this CAPTCHA to be completed with higher accuracy and speed than the control condition. A peripheral benefit of this design is that it is language-agnostic, though we note that there may sometimes be cultural differences in category membership — e.g., whether a rooster's crow should be counted in the bird chirp category.

4 Evaluation Methodology

We ran a controlled, within-subjects, online experiment with 67 blind and visually impaired users from around the world. Our study was IRB-approved.

4.1 Experiment and Procedure

Our experiment consisted of five conditions: four new designs and one baseline control condition that was used to emulate the industry-wide standard alphanumeric audio CAPTCHAs. To account for novelty and learning effects, we conducted three time-separated sessions, each spaced one week apart. In each session, participants completed three audio CAPTCHA challenges for each of our designs, and one challenge for the control. In total, participants were presented with the same 13 challenges per session in a randomized order.

We used Audacity, an open source audio-editing software, to create each CAPTCHA. Individual clips for each character and word were generated using a text-to-speech program that can synthesize audio in both male and female voices. These audio files included characters like 0-9, a-z, words for add, subtract, plus, and minus. We also accumulated open source audio clips of varying phenomena (i.e. birds chirping, instrument recordings, etc.) and background noises [1, 2]. These clips (apart from the categorical sounds used in the Categories prototype) were then distorted by applying audio effects that changed each clip's pitch, speed, and amplification. We used the same set and number of audio clips to create the 39 challenges and all CAPTCHAs were merged with distorted background noise at the same decibel level. These distortions were used to improve both the security and ecological validity of our designs [1, 2, 19].

The resulting audio CAPTCHAs were 16 to 18 seconds long, with a one second pause at the beginning to enable users to navigate to the edit text box to record their answers. In order to mimic existing audio CAPTCHA designs, the control challenge did not have an initial one second pause.

The web platform we designed to administer our CAPTCHA designs was tested in a preliminary pilot study in early 2019. According to feedback from the pilot, we then adjusted three of our designs, replaced another one entirely, and conducted a 3 week long within-subjects experiment in the summer of 2019.

We developed the online experimental test-bed using jQuery and HTML5 for the front-end, PHP for the back-end, and Heroku for hosting. In consultation with one of our research team members, who is visually impaired, we kept the user interface simple and accessible for PVIs who would need to navigate the interface with screen readers.

Participants first encountered a landing page in which they could see details about the study and provide informed consent. Next, participants were asked to complete a challenge under the standard design (control), followed by batches of three challenges each for our four custom prototypes (treatment). We randomized the order in which the treatment prototypes were presented to each user. An example challenge is illustrated in Appendix 4.

For each participant's first session, we conducted a video conference call on Zoom [38] in order to ensure that participants could use the experiment test-bed and complete the subsequent two sessions independently. We asked participants to share their screens to confirm their use of a screen reader to complete the study. Throughout the duration of the session we guided them between pages and answered their questions. One week after the first and second sessions were completed, we then emailed participants a subsequent link to complete the second and third sessions, respectively.

After completing the batch of 3 challenges for each prototype in each session, participants filled out a questionnaire in which we asked them to rate, on a Likert scale from 1 - 5, the usability of and their satisfaction with each prototype — "1" was coded as very low and "5" was coded as very high. We then asked participants if they preferred the prototype in comparison to the control. We also asked open-ended questions to solicit participants' thoughts and feedback on our designs. For the first session, we asked participants these open-ended questions in real time via Zoom. For the latter two, participants wrote-in their responses manually. Finally, at the end of the study, we collected each participant's age.

The data streams that informed our findings include quantitative and qualitative data, as well as our own facilitator observations, from both this study and the pilot in 2019.

4.2 Recruitment and Compensation

We reached out to a number of global organizations, including the American Foundation for the Blind, the National Federation of the Blind, Braille Works, the American Printing House for the Blind, the Blind Graduate's Forum of India, and VisionAid. We also leveraged blind social and support groups on social networking platforms (Facebook, WhatsApp, etc.) and mailing lists (Access India, Program-L and Voice Vision).

In total, we received 225 responses as a result of this outreach. Due to time and resource constraints we scheduled sessions with 150 participants over the course of six weeks, choosing participants in the order that we received their information. Accounting for those who dropped out or never responded, we interviewed 67 participants for at least one session. Each person was compensated 10 USD per completed session, for a total of 30 USD for completing all three sessions. Compensation was distributed in the form of regional Amazon gift certificates.

Our primary criteria for determining participant eligibility was their use of low-vision assistive technologies (e.g., braille displays, screen readers, and screen magnifiers) to navigate computer screens. Only one of our participants relied on screen magnification software. Although he had some vision, it was not clear enough for him to be able to solve visual CAPTCHAs. All other participants used screen readers and none used braille displays. Thus, we use the term "people with visual impairments" (PVI) to describe all of our participants.

	Accuracy Model (Logistic)	**Time Model (Linear)**
Fixed effect coefficients		
Session Number	0.35*	−0.17***
Math v. Control	2.78***	−0.73***
Character v. Control	2.50***	−0.70***
Pauses v. Control	1.77**	−0.61***
Categories v. Control	1.53*	−0.76***
Character v. Math	−0.27	0.03
Pauses v. Math	−1.00	0.12
Categories v. Math	−1.24*	−0.03
Pauses v. Character	−0.73	0.09
Categories v. Character	−0.97	−0.06
Categories v. Pauses	−0.24	−0.15
Age	0.005	0.002
Intercept	−0.87	−0.89***
Random intercepts variance		
Participant (N=67)	0.50	0.19
Challenge (N=39)	0.61	0.02

p <= 0.05, ** p <= 0.01, *** p <= 0.001

Table 2: Mixed-effects regression modeling both accuracy and completion time against prototype, session number, and age, with each participant and each challenge having its own random intercept. For the accuracy model, we ran a logistic regression and for the completion time model, we ran a linear regression. Broadly, the highlighted rows on the top indicate that all of our prototypes were significantly more accurate and faster than the control, and that participants grew more accurate and faster in subsequent sessions. The variance in random intercepts suggest significant variation across participants and challenges in success but not in completion time.

5 Results

5.1 Participant Demographics

Sixty-seven PVIs participated in the first zoom session; 34 of these continued on to remotely complete the two remaining sessions of our study. All participants were at least 18 years old and were, on average, 33.1 ($\sigma = 15.3$) years old. We did not collect gender data. All participants were able to speak, read, and write proficient English, which was verified in the first session via a face-to-face screen-sharing video chat.

5.2 Data Pre-Processing

Across all participants, we collected data on 2,259 CAPTCHA attempts. Of these attempts, we dropped 11 data points that were corrupted through data collection errors (i.e. with infeasible completion times of over 50 years, which we suspect is due to improperly set browser clocks). We also dropped one extreme outlier with a completion time of 93 minutes, or 46 standard deviations away from the dataset's mean completion time ($\mu = 36.9$ seconds, $\sigma = 120.6$ seconds). We suspect this participant left their browser window open while being away. Thus, we dropped 12 data points in total (0.5%). Our final dataset consisted of 2,247 CAPTCHA attempts from 67 PVIs.

5.3 RQ1: Task Performance Evaluation

We first evaluated how our novel designs compared to the control condition across two important task performance metrics: accuracy and completion time.

5.3.1 Accuracy

Across all our participants, in decreasing order, the accuracy rates for each prototype were: 89.2% for Math, 86.9% for Character, 76.2% for Pauses, 70.3% for Categories, and 42.9% for the control.

To test if these differences in accuracy were statistically significant, we modeled the accuracy of our designs with a random-intercepts logistic regression using the lme4 package in R. Our input data were the 2,247 individual attempts at solving a CAPTCHA challenge. Our dependent variable was a binary measure of whether or not a participant successfully completed the challenge. Our IV was the prototype used in the challenge — a categorical variable encompassing our four treatment designs and the control. As covariates, we included the session number and participant age. We also included a random intercept term for the 67 distinct participants and the 39 distinct challenges to account for and model the effects of repeated observations. We used R's multcomp package to conduct pairwise comparisons between each of the ${}_5C_2 = 10$ combinations of our novel prototype designs and the control. The results are shown in the first column of Table 2, with p-values adjusted using Bonferroni correction.

Most importantly, **we found that each of our prototypes — Math ($b = +2.78$), Character ($b = +2.50$), Pauses ($b = +1.85$), and Categories ($b = +1.50$) — were completed with significantly higher accuracy than the control.** We also found evidence of a learning effect: participants were significantly more accurate in later sessions ($b = +0.35$). After correcting for multiple testing, we did not find many statistically significant differences in accuracy between our four designs, with one exception: the Categories prototype was significantly less accurate than the Math prototype ($b = -1.24$).

The variance in random intercepts across distinct participants ($\sigma^2 = 0.50$) and challenges ($\sigma^2 = 0.61$) suggests that

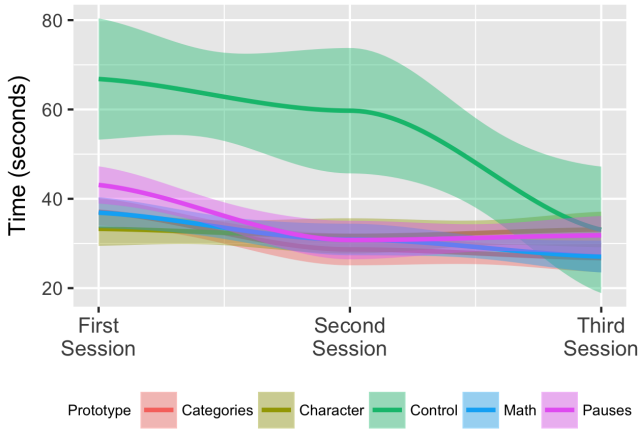

Figure 2: Distribution of average completion times, with 95% confidence intervals, for each prototype across sessions. There is evidence of a significant learning effect in which completion times drop with increased accuracy and repeated exposure. Thus, participants learned and adapted to our novel designs.

performance could vary in non-trivial ways between individual participants and across different challenges. To better illustrate this point, our most successful participant (a 49 year old from the USA) got 100% of their challenges correct, while our least successful participant (a 41 year old from India) got 46% of their challenges correct. Our most successful challenge was solved with 97.5% overall accuracy (the third challenge of the Categories prototype in the second session), while our least successful challenge was solved with 17.2% accuracy (the control challenge in the second session).

5.3.2 Completion Time

We next investigated how our designs varied by completion time. Broadly, the average completion time for a challenge was lowest for the Categories prototype (31.1 s), followed by the Math prototype (31.7s), Character (32.7 s), Pauses (35.4 s) and, finally, the control (53.6 s).

To test if these differences were statistically significant, we ran a second random-intercepts regression. The model parameters were the same as the aforementioned accuracy model, although with two exceptions: the DV was scaled and centered for time taken to complete each design. Because the DV was continuous instead of binary, we employed a linear regression. Once again, we used R's lme4 package to estimate the model, and R's multcomp package to do pairwise comparisons across the prototype designs, with p-values adjusted using Bonferroni correction.

The results can be seen in the second column of Table 2, labeled "Time Model." For a numeric predictor (i.e., Age, Session Number), the model suggests that a positive coef-

ficient of $b = +1.0$, as the predictor increases by one unit, means the estimated completion time will *increase* by one standard deviation. A negative coefficient implies that the estimated completion time would *decrease* by one standard deviation. For a categorical predictor (i.e, Prototype design), a positive coefficient of $b = +1.0$ would suggest that the estimated difference in completion time between two levels of the predictor — a treatment and a control (e.g., Math vs. Control) — is one standard deviation, with the comparison condition taking longer than the control. A negative coefficient implies the opposite — that the control takes one standard deviation longer than the comparison. All of these examples assume that all other predictors (IVs + covariates) are held constant.

We found that all of our novel prototype designs — Math ($b = -0.73$), Character ($b = -0.7$), Pauses ($b = -0.61$), and Categories ($b = -0.76$) — were significantly faster than the control. We did not find a significant difference in completion time between any two of our custom prototypes. We again found a learning effect: participants were significantly faster in later sessions ($b = -0.17$). Figure 2 illustrates the distribution of completion times across all four of our prototypes and the control condition, and shows how those distributions vary across sessions. On average, completion times decreased over the course of all three sessions for every prototype design, most significantly for the control.

The variance in random intercepts across distinct participants ($\sigma^2 = 0.19$) and challenges ($\sigma^2 = 0.02$) was fairly small, suggesting that, accounting for the fixed effects in the model, completion times did not dramatically vary between participants and challenges.

5.4 RQ2: Security Evaluation

While we consider our primary contribution to be a usability assessment of alternative audio CAPTCHA designs, we also evaluated the security of our prototypes relative to the control condition. We considered two threat models.

The first is a *random-guessing adversary*. For content-based CAPTCHAs like the control and Pauses prototypes, this adversary is trivially defeated. Assuming a 32-character alphabet (all English letters along with 0 - 9 digits), a random string of 6 characters would yield a search space of 32^6 possibilities, which would be impractical to randomly guess. For our rule-based audio CAPTCHAs, however, random guessing is more potent. For the Character and Categories prototypes, the space of possible outputs given a 10-character long string is 0 - 10 — i.e, the random guessing adversary would have a $1/11 \approx 9\%$ success rate. For the Math prototype, we assume five single-digit operands connected through either plus or minus operators, while there are $10^5 + 2^4 = 100,016$ possible inputs, the range of possible outputs varies from $[0-9-9-9-9 = -36, 9+9+9+9+9 = 45]$. Thus, there are 72 possible outputs, so a random guessing adversary would have a $1/72 \approx 1\%$ chance at breaking the Math prototype. A smarter adversary might notice that the distribution of outputs is not uniform but a normal distribution centered at 5. Thus, by guessing "5" on every attempt they would increase their chances of defeating the Math prototype to $\approx 3\%$.

The second adversary we considered is one who uses state-of-the-art NLP — either commercially available or easily trainable using public-domain knowledge and data — to deconstruct the audio file and solve the challenge. Motivated by Polakis et al. [27], for the Math, Character, and Pauses prototypes, we tested the robustness of our designs using Google's automated, off-the-shelf speech recognition software. We considered a clip broken if all the entities were successfully parsed from audio to text, because once the content is parsed, the application of rules to get the correct answer is trivial. For the control prototype, 0 out of 3 (0%) designs were successfully parsed. The percentages of challenges within each design that were broken by this threat model include: 2 out of 9 (22%) for Math; 1 out of 9 (11%) for Character; and 6 out of 9 (67%) for Pauses.

The Categories prototype could not be tested using off-the-shelf parsing services because an appropriate parser for real-world sound classification does not exist. Thus, we created our own parser using deep learning. We implemented this parser on Tensorflow and trained it on Google Research's Audioset data, a collection of 632 audio event classes and over 2 million 10 second sound clips labeled by humans [12]. We considered a clip broken if the parser was able to predict 'true positives' and 'true negatives.' For instance, if the CAPTCHA challenge is to count the total number of animal sounds, the 'true positives' include sounds that are identified as animal-related and 'true negatives' are successfully identified as non-animal sounds. Similarly, 'false positives' are identified as sounds that are incorrectly labeled as animal sounds and 'false negatives' are animal sounds that are incorrectly labeled non-animal sounds. We found that of all predictions made for 48 sub-clips across all categories, challenges, and sessions, 13 were true positives and 16 were true negatives. There were 8 false positives and 11 false negatives. The average error of 2.1% per clip contributed to either false positive or false negative predictions, so we can deduce that none of the clips belonging to the Categories prototype were fully parsed.

In summary, the random-guessing adversary would be trivially defeated by the Pauses prototype, but could have a small chance at defeating the Math, Character, and Categories prototypes. This could be made harder by increasing the length of the CAPTCHA, though likely at the cost of usability, speed, and accuracy. The NLP adversary would have good success at breaking the Pauses prototype, and a slightly better chance at breaking the Math and Character prototypes than the random-guessing adversary, but would struggle with Categories.

Next, we considered how these results compared to the control condition. Prior work has shown that standard audio CAPTCHAs are largely insecure against state-of-the-art machine learning. For instance, Bursztein et al. (2011)

showed that 45% of Yahoo, 49% of Microsoft, and 83% of eBay CAPTCHAs can be broken. Also, Sano et al. (2013) and Meutzner et al. (2014) successfully broke Google's reCAPTCHA with a success rate of 52% and 63% while Tam et al. (2009) achieved an accuracy of up to 71% using machine learning techniques like AdaBoost, SVM, and k-NN. With the deep learning evolution, audio CAPTCHA attacks continue to succeed. In 2017, Saumya et al. [27] developed a low cost 'AudioBreaker' system using off-the-shelf speech recognition services. It successfully broke seven popular audio CAPTCHA schemes along with 98% accuracy in breaking Google's reCaptcha. In 2019, Heemany et al. [29] even demonstrated 85% accuracy in breaking designs with higher background noise levels.

In short, all of the CAPTCHAs we considered, including the control, could be broken by motivated adversaries. Thus, we must consider the use-context. While our designs should not be used for security-critical applications, they should provide sufficient security for low-risk contexts in day-to-day web browsing (e.g., comment form submission).

5.5 RQ3: Usability Evaluation

Finally, we conducted a series of quantitative, qualitative and heuristic usability evaluations based on our observations of participants in the initial Zoom session as well as participants' survey responses.

5.5.1 Usability, Satisfaction, and Preference

After each challenge, participants rated the usability and satisfaction of our designs on a 5-point Likert scale and also answered whether or not they preferred our design over the control. This gave us 2,247 usability, satisfaction and preference data points. Due to the Hawthorne effect, the absolute values of these ratings were not as important as their relative ordering, which helped illuminate participant preferences.

The distributions for usability and satisfaction were highly skewed, with participants rating 1,916 challenges a "5" on usability and 1,865 a "5" on satisfaction. To simplify analysis, we converted these scales into binary values: "5" or not "5." We then conducted three random intercepts logistic regressions, using R's lme4 package, correlating usability, satisfaction and preference-over-control to prototype design. We included a random-intercepts term for participant and challenge to control for repeated observations. We ran pairwise comparisons using R's multcomp package, adjusting p-values with Bonferroni correction. The results are shown in Table 3.

Usability. The regression results in Table 3 suggest that, controlling for the effects of individual preference, challenge variance, and the session number, the Character and Categories prototypes were rated significantly more usable than Math and Pauses; the Math prototype was rated more usable than Pauses; and the Pauses prototype was rated less usable than the other three. We also found a significant positive effect of Session number, suggesting that participants found all prototypes more usable in later sessions.

	Satisfaction	Usability	Pref. Over Control
Character vs. Math	1.31***	0.96**	1.11***
Pauses vs. Math	−0.47	−0.57*	0.36
Categories vs. Math	0.32*	0.74***	0.34***
Pauses vs. Character	−1.79***	−1.52***	0.76***
Categories vs. Character	−0.99**	−0.22	−0.77***
Categories vs. Pauses	0.79**	1.31***	−0.02
Session	0.32**	0.74***	0.34***
Intercept	3.05***	2.65***	0.42

* p <= 0.05, ** p <= 0.01, *** p <= 0.001

Table 3: Random-intercepts logistic regression results modeling satisfaction, usability and preference as a function of prototype design and session number. The Character prototype was rated the most usable and satisfying. The Pauses prototype was most preferred over the control.

Satisfaction. Table 3 also shows that the Character prototype had a significantly higher satisfaction rating than all other prototypes; the Categories prototype out-performed Math and Pauses; and there was no significant difference found between Pauses and Math. We also found a significant positive effect of Session number, again suggesting that participants' satisfaction increased in later sessions.

Preference. Overall, participants reported preference for our prototypes over the control: 73% preferred Pauses to control, 67% preferred Character, 61% preferred Categories and 52% preferred Math. Controlling for repeated observations, challenge exposure, and session number, Table 3 shows which pairwise differences are statistically significant. In short, Pauses was preferred more than Character; Character was preferred more often than Categories and Math; and Categories was preferred more often than Math. Participants' overall preference for our prototypes over control also increased in later sessions.

In sum, the Character prototype had the highest usability and satisfaction ratings and was the second most preferred over control after Pauses. The Math prototype was generally rated the least usable, least satisfying, and least preferred. This result presents an unfortunate dilemma — the prototype that provided the highest accuracy and second highest speed was also the least subjectively "usable."

5.5.2 Heuristic Analysis: Quantitative

Beyond individual perceptions of usability, we next performed a quantitative heuristic analysis to assess the usability of our prototypes. Jacob Nielson's heuristics for designing usable systems span five core components to ensure design quality:

learnability, or the ease of correctly completing a CAPTCHA upon initial exposure; *efficiency*, or the rate at which users can learn how to complete new designs; *memorability*, or the ability to re-learn the correct use of a CAPTCHA design after a period of inactivity; *errors*, or the extent and nature of errors users make; and finally, *satisfaction*, or users' level of enjoyment in completing the CAPTCHA [24].

In order to *quantitatively* measure these criteria, we used the following data to address Nielson's heuristics:

1. *Learnability*: the session 1 accuracy rates for our prototypes against the control design.
2. *Efficiency*: whether the average number of replays decreased across sessions 1 through 3.
3. *Memorability*: whether the accuracy rates for our prototypes increased between sessions 1 through 3.
4. *Errors*: the average time it took for users to complete each challenge accurately.
5. *Satisfaction*: the self-reported user satisfaction scores (1 - 5) for each prototype.

Table 4 illustrates that the initial *learnability* of our designs, as measured by average accuracy in session 1, is comparatively higher than that of the control CAPTCHA. These numbers indicate that our designs were more learnable than the control, despite the fact that users had the most exposure to the control from day-to-day web browsing. Additionally, the significantly higher initial accuracy of the Pauses prototype, which was identical to the control apart from small time gaps in-between characters, signal that users need slower-paced CAPTCHAs to answer them correctly.

In terms of *efficiency*, two of our designs showed a steadily decreasing number of replays in subsequent sessions. The Math and Categories prototypes displayed clear improvements — users required fewer replays in session 3 than in session 1. The control also required fewer replays in subsequent sessions, but still had the highest average number of replays per session compared to all other designs.

Similarly, in terms of *memorability*, only the Math and Categories prototypes had improved accuracy scores in subsequent sessions. Users' subjective ratings of these prototypes — usability, satisfaction, and preference over the control — also improved over time (see Figure 3).

In practice, all of our designs and control, if answered correctly the first time, should take about the same amount of time to complete. However, users spent the longest time completing the control CAPTCHA correctly, suggesting that it was the most prone to *errors*. Generally, the Categories prototype was fastest: it took 9.4 fewer seconds to accurately complete than the control. The Math prototype was second fastest, followed by Character and then Pauses.

As we did not collect subjective feedback for the control, we are unable to contrast the *satisfaction* scores of our novel designs vis-a-vis the control. However, as we saw in the previous section, Character and Pauses had the highest satisfaction scores of 4.93 and 4.85, respectively.

5.5.3 Heuristic Analysis: Qualitative

Next, we used a two-dimensional subset of Yan et al.'s [37] qualitative usability assessment framework to analyze participants' open-ended feedback, in order to better understand their perceptions and difficulties with each of our new designs. Specifically, the two dimensions we qualitatively assessed were:

1. *Distortion*: level and type of distortion, use of confusing characters, and design difficulty for native and non-native speakers.
2. *Content*: language specificity of the characters used, the length of each CAPTCHA challenge, length of answers, and predictability of the design.

We start with a broad assessment of these dimensions, and then discuss individual participant feedback pertaining to these dimensions for each of our prototypes.

In terms of *distortion*, the audio files we used in the control, Math, Character, and Pauses prototypes were drawn from the same set of letters, numbers, and operators that varied in terms of voice, speed and pitch. As a compromise between usability and security, we picked sounds that were dynamic and not overwhelmingly loud. Because of the distortion of background noises and characters, participants noted that certain letter groupings like "2," "q," and "u" were hard to distinguish. Other participants with hearing problems had difficulty understanding some of the deeper voices that resulted from very slow audio speeds.

Additionally, all prototypes except Categories relied on the user's ability to understand letters, numbers, and mathematical operators spoken in English. However, knowledge of these fundamentals is certainly attainable for non-native speakers since challenge instructions can be translated by web pages into nearly any language. The Categories prototype, in its use of more universal audio events, had the fewest language-specific constraints [6] [37].

In terms of *content*, as portrayed in Table 1, the six-to-eight-character length strings of our CAPTCHA challenges were comparable to existing designs. While challenge length and instruction sets for all our prototypes were predictable, their content varied in predictability. Of our novel designs, Pauses was the most familiar, while the other designs were based on less predictable rule-based methods. However, the higher average accuracy of Math and Character prototypes over the Pauses prototype and control suggests that content predictability is not necessary for success in usage.

	Learnability	Efficiency	Memorability	Errors	Satisfaction
	Avg. session 1 accuracy	**Num. replays decrease across sessions?**	**Improved accuracy across sessions?**	**Avg. time to answer correctly (seconds)**	**Avg. satisfaction score (1-5)**
Control	32%	Y	N	39.0	N/A
Math	83%	Y	Y	30.2	4.6
Character	89%	N	N	30.1	4.8
Pauses	72%	N	N	34.6	4.8
Categories	56%	Y	Y	29.6	4.5

Table 4: This table illustrates the quantitative usability assessment based on Nielson's criteria. We found that all of our designs exemplified higher usability than the control, with the exception of the efficiency and satisfaction categories, which are discussed in more detail below. Notably, Categories scored well in the Efficiency, Memorability, Errors, and Satisfaction heuristics.

Math Prototype:

Front-heavy errors. Many of the mistakes participants made with this prototype occurred in earlier sessions, owing at least partially due to confusion with the *content* of instructions. For instance, a few users who had misunderstood the challenge submitted a string of the entire equation rather than providing a single sum. This content-specific error, which we also observed for the Character prototype, was likely due to the conflation of instructions with existing content-based designs that require users to repeat exactly what they hear.

Since average accuracy rates improved over time, from 84.8% in session 1 to 92.8% in session 3, we suspect that accuracy may continue to increase as familiarity with the design grows. Participants reported feeling better prepared to use the Math prototype in later sessions. For example, one 38 year old participant from the Czech Republic said, *"It is simply usable if people remember the last result. I had problems in previous runs but in this I learned how to concentrate."*

Accessibility concerns. Some participants were concerned about the accessibility of a math-based design, namely regarding the *distortion* and *content*. For instance, a 46 year old from the USA said, *"It wasn't difficult for me but sighted individuals do not have to do math like this and I don't feel I should have to be challenged in a way that others are not. Having said this I found it easy to use. I am concerned if this model were used with persons who had cognitive challenges or if it were used with children the task could be too complex."* Another participant noted that we failed to consider users with multiple physical impairments, such as loss of vision *and* hearing. This tendency to speak for less cognitively-able members of the PVI community was quite common, as research shows there is a trade-off in that advanced cognitive abilities allow audio CAPTCHA users to complete challenges faster [23]. These users were hinting at a fairness divide between visual and audio CAPTCHAs that emerged from real usability differences between the two authentication methods.

Character Prototype:

Confusing instructions. Similar to the Math prototype, many of the errors that users made with the Character prototype were due to confusion with the *content* of instructions we provided. Recall an example of the provided instructions which were: "You must count the number of times '6' is spoken throughout the audio clip. Type the sum in the text box at the end." In evaluating that challenge, a 55 year old user from India said, *"The term 'sum' is confusing particularly when we are asked to count the number of times '6' is spoken throughout the clip and to write the sum in the box. If '6' is spoken three times then we should write '3' or '18'? Hence more clarity is needed on this type of CAPTCHA."* This signals that the exact wording of instructions must be carefully considered before implementation.

Similar sounding characters. Participants also pointed out areas of difficulty related to *distortion*, such as the length of the alphanumeric string being too long or that letters with similar phonics sounded too similar to each other. Examples of confusing letter groupings were: ("2", "q", "u") and ("b", "e", "z", "d", "v", "p", "c", "t", "g"). This issue is inherent to all alphanumeric audio CAPTCHA designs, further suggesting the need for exploring non-language based designs.

Ease. Despite the aforementioned challenges, users generally found this prototype easy to use, thus suggesting positive outcomes related to Yan et al.'s heuristic usability criteria [37]. Several participants reported that its difficulty level was comparable to their perceptions of visual CAPTCHAs, which was one of our design goals. A 49 year old from the USA stated, *"I was able to understand all the letters and numbers even with the distortion of the sounds. Counting letters and numbers is easier than trying to remember or type in the whole set which sometimes requires listening to it 2-3 times."*

Pauses Prototype:

Accounting for hearing loss. Participants noted the impor-

tance of accounting for hearing loss, particularly in content-based CAPTCHAs like the control and Pauses. For instance, a 34 year old from the USA stated, *"Qs are spoken rather deep and some people have trouble hearing very deep voices such as myself. I have slight hearing loss in my left ear that makes it almost impossible to hear deep [male] voices so I would raise the pitch on Qs."* Participants indicated that higher quality audio samples with a consistent volume level could have improved the accessibility and *distortion* of this design.

Longer pauses were helpful. Participants found the longer gap between characters helped improve accuracy and deal with interruptions. A 58 year old participant from India reported: *"The pauses are helpful to solve the CAPTCHA and hence need to be implemented."* Similarly, a 27 year old from the USA explained how the extra gaps between characters afford greater flexibility and ease of use: *"This time I realized that this layout is much better than I thought. I was interrupted by someone asking me a question and I was able to record the last few characters and play it again to get the first ones. The gaps are so long that I believe people will also be able to find where they left off and keep going."* Overall, this feedback suggests that the *content* of Pauses was usable.

Categories Prototype:

Ambiguous category membership. Category membership can be culturally-specific. When we asked users to identify the number of times a bird sound was played, a few questioned whether a rooster is a bird. In fact, two participants noted that they associate the sound of roosters with their alarm clock, which led them to disassociate the sound of a rooster with a bird. Participants also thought animal categories can be too culturally-dependent and thus should not be used. In other words, in order to overcome barriers related to both *content* and *distortion*, sound categories should be specific and tailored to certain locales or universally recognizable.

Instrumental sounds could also be ambiguous at times and so participants needed instructions to identify a specific type of instrument such as a guitar. One 20 year old participant from the USA said, *"I strongly disagree with this design because it's asking people to make associations. Almost any sound can be associated with a musical instrument."*

Non-linguistic CAPTCHAs may be more universally appropriate. Other participants appreciated that Categories did not require knowledge of the English language. For example, a 21 year old participant from India stated, *"I was thinking these CAPTCHAs might be excellent for people whose their main language is not English and would be a great help for them. For example in many websites the CAPTCHAs are in English and I've talked with some friends who don't speak English at all. They used to tell me that due to these types of CAPTCHAs they needed to find help from their families."*

Fun and ease. Several users expressed that this design was more "interesting", "fun", and easier than alphanumeric alternatives. For instance, a 41-year-old from Italy stated, *"It is far more interesting than typing in numbers or letters because it is language independent, more pleasant and less cognitive demanding."* Likewise, a 27-year-old participant from the USA said, *"It is very easy to use because people can easily identify these common noises. It also requires less brain power than math or memorizing a long string of characters."*.

	Avg Accuracy	Avg Speed	Pref. to control	Security Random	Security NLP
Control	43%	53.6s	2.7%	++	-
Math	89%	31.7s	52%	+	+
Character	87%	32.7s	67%	-	+
Pauses	76%	35.4s	73%	++	-
Categories	70%	31.1s	61%	-	+

Table 5: Summary of key results. The Math prototype had the highest accuracy; the Categories prototype had the highest speed; the Pauses protootype was most preferred; the control and Pauses prototype were most resilient to random guessing; and, the Math, Character and Categories prototypes were most resilient to NLP.

5.5.4 Ad-hoc Usability Observations

Through our observations, user feedback and trial-and-error, we uncovered a number of one-off design attributes for audio CAPTCHAs to improve usability and accessibility. First, participants found 'auto-play' features to be a nuisance that rushed them through the task before they were ready if they, e.g., accidentally skipped through text instructions. Additionally, we found that placing a one-second gap, between hitting the play button and hearing the first character, helped both usability and accessibility. Finally, from our experience with conducting both the pilot and the full study, we found 1.25 seconds to be the optimal time gap between audio clips, regardless of prototype.

6 Discussion

Table 5 summarizes our designs, relative to the control, across key dimensions of interest. Our high-level goal was to design audio CAPTCHAs that were faster, more accurate, and that provided reasonable security. Our key results suggest that all four designs were significantly more accurate and faster (RQ1) than the control. The Math and Character prototypes showed average accuracy rates of 89% and 87%, respectively, which are on par with traditional visual CAPTCHAs. The Categories prototype was the fastest to complete (31.1 s), with the Math prototype being a close second (31.7 s). In terms of security (RQ2), the Math prototype provided decent resilience against both of the adversaries we tested. The Categories and Character prototypes were more vulnerable to random

guessing, while the Control and Pauses prototype were more vulnerable to NLP. Finally, through a series of quantitative, qualitative and heuristic usability analyses (RQ3), we found that the Pauses prototype was most preferred; the Character prototype was most satisfying; the Categories prototype was most globally accessible; and, the Math prototype was least usable, satisfying and preferred.

Based on the diversity of these results, the best design to use is dependent on use-context. The Math prototype provides the best balance of task performance and security against both types of adversaries, but was perceived to be least usable. The Characters prototype was vulnerable to random guessing but was the highest rated in usability and satisfaction. The Pauses prototype was most preferred over the control condition and would be the easiest to deploy in its similarity to existing audio CAPTCHAs. Finally, the Categories prototype was the fastest, inspired the most positive qualitative feedback, and utilized language-agnostic challenges.

6.1 Practical Design Recommendations

Through a combination of trial-and-error, along with open-ended feedback from participants in our pilot study, we have distilled a number of practical recommendations for designers or researchers who might use or improve upon our prototypes:

- Provide ≥ 1 second of silence after the user presses play;
- Place 'play' button beside the answer box without 'auto-play' functionality;
- Place 1.25 second gaps between audio clips;
- Avoid language or cultural-based challenges in favor of ones with universal sounds (i.e. running water);
- Choose specific sound categories when asking users to count non-alphanumeric sounds;
- Consider the loss of various physical abilities in users;
- Code instructions as audio elements to prevent skippage;
- Use high quality audio samples;
- Maintain a consistent volume level for all audio clips.

6.2 Limitations and Future Work

Participant Retention. Participant retention is a common limitation in multi-session studies. Our study took place over three time-separated sessions and required participants to complete the final two sessions independently. While we wanted these to be completed one week apart, some participants procrastinated for a week or more. Additionally, 34 of the initial 67 participants did not complete the final two sessions at all.
One-second pause. We incorporated a one-second pause at the beginning of our new designs so that screen reader users could navigate to the answer box before the challenge began. However, we did not incorporate this pause to the beginning of the control because we wanted it to emulate a real-world baseline, and existing audio CAPTCHAs do not have an initial one-second pause. This discrepancy could have increased the performance of our designs relative to the control.

Ecological validity. Our experiment test-bed was uniquely accessible. In practice, our designs may be embedded within otherwise inaccessible websites, which could impact PVI's performance. While the relative results between the conditions we tested should hold, in practice, accuracy and speed may be different. A field study of our novel audio CAPTCHA designs may help address these concerns in future work.

Intersectional accessibility. Our designs were for PVIs, but we did not consider other disadvantages and impairments. For example, our designs assumed participants did not have hearing impairments. Our Math prototype assumed it was simple for people to do mental arithmetic. Our Categories prototype did not consider cultural influences on category membership. A fruitful area of inquiry for future work may be designing human-intelligence proofs that don't rely on the acuity of human senses, that don't pre-suppose educational background and cognitive abilities, and that are more culturally inclusive.

7 Conclusion

Motivated by the usability shortcomings of existing audio CAPTCHAs, we designed, implemented and evaluated four alternatives that we hypothesized would improve the audio CAPTCHA user experience for people with visual impairments. We experimentally tested this hypothesis in a controlled, randomized within-subjects experiment with 67 PVIs and found that all of our designs significantly outperformed the control condition in both performance measures (accuracy, completion time) and perceptions of usability. None of our designs stood out as a clear winner. Rather, each of them boasted complementary improvements to the user experience — the Math Prototype was the most accurate and second fastest, the Character prototype was rated the most usable and satisfying, the Pauses prototype was the most familiar and preferred over the control, and the Categories prototype was the fastest and most globally accessible. These improvements, however, came at the expense of increased vulnerability against random guessing attacks for three of our four designs. For use-cases where high security is not critical — e.g., form submissions in everyday web browsing — this trade-off may be worth considering to improve the day-to-day browsing experiences of PVIs. In short, our findings help extend the state-of-the-art in usable audio CAPTCHAs and should strengthen the foundation for researchers and practitioners to explore the design space of more usable audio CAPTCHAs.

Acknowledgments

We thank all participants who were generous with their time, plus all the organizational staff members for their support in distributing our recruitment message. We could not have reached so far around the world had it not been for this assistance. We are also indebted to Dr. Yang Wang at the University of Illinois at Urbana Champaign for his extensive critiques during the editing process, Dr. Elissa Redmiles for her thoughtful suggestions on refining our study design, and the support of Youngwook Do from the Georgia Tech SPUD Lab. This work was generously supported by seed funds from Georgia Tech's School of Interactive Computing.

References

[1] Big sound bank. https://bigsoundbank.com/, Accessed: 2019-09-19.

[2] Zapsplat. https://www.zapsplat.com/sound-effect-categories, Accessed: 2019-09-19.

[3] K. Aiswarya and K. S. Kuppusamy. A study of audio captcha and their limitations. 2015.

[4] Sacha Brostoff Angela Sasse and Dirk Weirich. Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19:122–130, 08 2001.

[5] Jeffrey P. Bigham and Anna C. Cavender. Evaluating existing audio captchas and an interface optimized for non-visual use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1829–1838, New York, NY, USA, 2009. ACM.

[6] Stacey Burling. Captcha: The story behind those squiggly computer letters, 2012. https://phys.org/news/2012-06-captcha-story-squiggly-letters.html.

[7] Elie Bursztein and Steven Bethard. Decaptcha: breaking 75% of ebay audio captchas. 01 2009.

[8] Artemios Voyiatzis Christos Fidas. On users' preference on localized vs. latin-based captcha challenges. pages 358–365, 09 2013.

[9] HT Correspondent. Number of blind to come down by 4m as india set to change blindness definition, 2017. https://tinyurl.com/y8se7h4d.

[10] Celine Fabry John C. Mitchell Elie Bursztein, Steven Bethard and Dan Jurafsky. How good are humans at solving captchas? a large scale evaluation. In *2010 IEEE Symposium on Security and Privacy*, pages 399–413, New York, NY, USA, 2010. IEEE.

[11] Bowei Du Melissa Densmore Matthew Kam Sergiu Nedevschi Joyojeet Pal Rabin Patra Sonesh Surana Eric Brewer, Michael Demmer and Kevin Fall. The case for technology in developing regions. *Computer*, 38:25–38, 06 2005.

[12] Google. Audioset. https://research.google.com/audioset/, Accessed: 2019-09-21.

[13] Jinjuan Feng Graig Sauer, Harry Hochheiser and Jonathan Lazar. Towards a universally usable captcha, 2008. https://cups.cs.cmu.edu/soups/2008/SOAPS/sauer.pdf.

[14] Dan Yao Xiyang Liu Haichang Gao, Honggang Liu and Uwe Aickelin. An audio captcha to distinguish humans from computers. *SSRN Electronic Journal*, 01 2010.

[15] Dorothea Kolossa Hendrik Meutzner, Santosh Gupta. Constructing secure audio captchas by exploiting differences between humans and machines. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2335–2338, New York, NY, USA, 2015. ACM.

[16] Sean Hyde Jennifer Tam, Jirí Simsa and Luis Ahn. Breaking audio captchas. pages 1625–1632, 01 2008.

[17] Jinjuan Heidi Feng Jonathan Holman, Jonathan Lazar and John D'Arcy. Developing usable captchas for blind users. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 245–246, New York, NY, USA, 2007. ACM.

[18] Tim Brooks Genna Melamed Brian Wentz Jon Holman Abiodun Olalere Jonathan Lazar, Jinjuan Feng and Nnanna Ekedebe. The soundsright captcha: An improved approach to audio human interaction proofs for blind users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2267–2276, New York, NY, USA, 2012. ACM.

[19] George Hughey Dave Levin Kevin Bock, Daven Patel. uncaptcha: A low-resource defeat of recaptcha's audio challenge, 2017. https://uncaptcha.cs.umd.edu/papers/uncaptcha_woot17.pdf.

[20] Sunny Behal Kiranjot Kaur. Captcha and its techniques: A review. *International Journal of Computer Science and Information Technologies,*, 5, 01 2014.

[21] Patrice Simard Kumar Chellapilla, Kevin Larson and Mary Czerwinski. Designing human friendly human interaction proofs (hips). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 711–720, New York, NY, USA, 2005. ACM.

[22] Manuel Blum Luis von Ahn and John Langford. Telling humans and computers apart (automatically). *Communications of the ACM*, 47, 05 2002.

[23] Panagiotis Germanakos Marios Belk, Christos Fidas and George Samaras. Do human cognitive differences in information processing affect preference and performance of captcha? *International Journal of Human-Computer Studies*, 84, 07 2015.

[24] Jakob Nielsen. Usability 101: Introduction to usability, 2012. https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

[25] A. Bujari Ombretta Gaggi, Giacomo Quadrio. Accessibility for the visually impaired: State of the art and open issues. pages 1–6, 01 2019.

[26] T. Braithwaite M.V. Cicinelli A. Das J.B. Jonas Y. Zheng R.R. Bourne, S.R. Flaxman. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis, 2017. https://doi.org/10.1016/S2214-109X(17)30293-0.

[27] Varshini Sampath Saumya Solanki, Gautam Krishnan and Jason Polakis. In (cyber)space bots can hear you speak: Breaking audio captchas using ots speech recognition. pages 69–80, 11 2017.

[28] Andy Schlaikjer. A dual-use speech captcha: Aiding visually impaired web users while providing transcriptions of audio streams. 01 2007.

[29] Heemany Shekhar. Breaking audio captcha using machine learning/deep learning and related defense mechanism. page 741, 2019.

[30] Takuma Otsuka Shotaro Sano and Hiroshi Okuno. Solving google's continuous audio captcha with hmm-based automatic speech recognition. 11 2013.

[31] Yannis Soupionis and Dimitris Gritzalis. Audio captcha: Existing solutions assessment and a new implementation for voip telephony. *Computers Security*, 29:603–618, 07 2010.

[32] Noshina Tariq and Farrukh Khan. *Match-the-Sound CAPTCHA*, pages 803–808. 07 2018.

[33] Stephen Goglin Travis Schluessler and Erik Johnson. Is a bot at control? detecting input data attacks. In *Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games*, pages 1–6, New York, NY, USA, 2007. ACM.

[34] WebAIM. Screen reader user survey 7 results, 2017. https://webaim.org/projects/screenreadersurvey7.

[35] WHO. Global data on visual impairments 2010, 2010. https://www.who.int/blindness/GLOBALDATAFINALforweb.pdf.

[36] WHO. Visual impairment and blindness 2010, 2010. https://www.who.int/blindness/data_maps/VIFACTSHEETGLODAT2012_2.pdf.

[37] Jeff Yan and Ahmad Ahmad. Usability of captchas or usability issues in captcha design. pages 44–52, 01 2008.

[38] Zoom. Video conferencing, web conferencing, webinars, screen sharing. https://zoom.us/, Accessed: 2019-09-19.

A Appendix: Screenshot of Experimental Test-Bed and Questionnaires

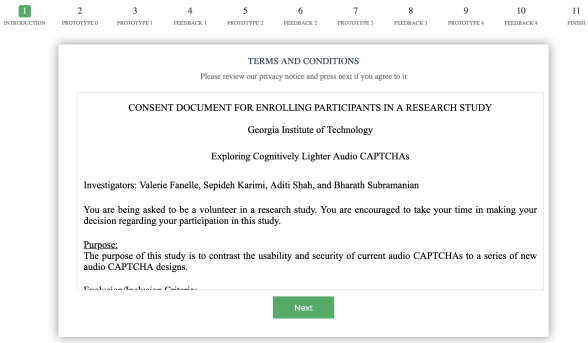

Figure 3: The landing page for the study website on which users clicked the 'Next' button if they decided to consent to the study.

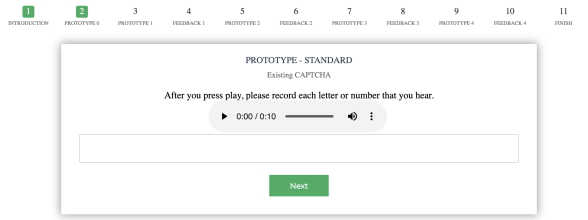

Figure 4: We used a webpage format as a test-bed for each of our CAPTCHA designs. Participants could replay the CAPTCHA as many times as they needed until they felt confident of their answers.

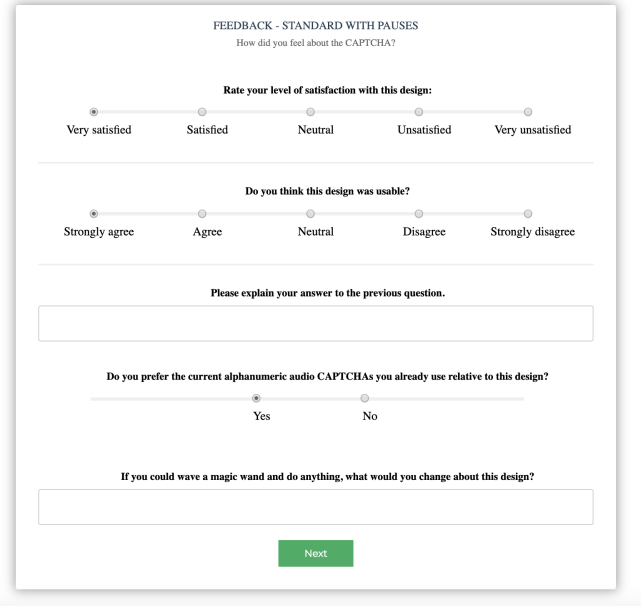

Figure 5: We remotely obtained participant feedback on the usability and satisfaction of each design via the above feedback form.

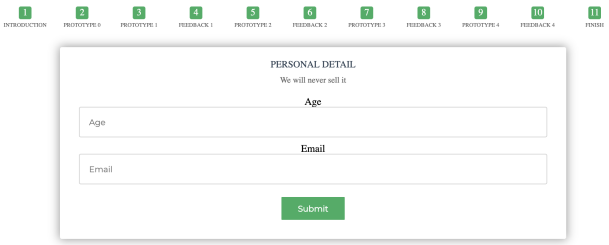

Figure 6: After all 13 CAPTCHAs for that session were presented, participants were asked to enter their email and age for demographics and transcription purposes.