AHFE International

Usability Study of Auditory CAPTCHA

Chia-Hung Lee and Ying-Lien Lee

Department of Industrial Engineering and Management
Chaoyang University of Technology
Wufeng District, Taichung City, 41349, Taiwan

ABSTRACT

CAPTCHA is a security system to distinguish whether a user is a human being or an automated program by asking questions that are hard for artificial intelligence yet relatively easy for human to answer. Two most popular forms of CAPTCHAs are text and audio; this study attempts to explore the latter one, which is common in situation where visual interaction is not applicable, such as in voice-based interaction or for visually challenged users. Auditory CAPTCHAs can be breached by content analysis and guessing through Automatic Speech Recognition (ASR), it is then necessary to blend certain level of interference to counterattack. However, by doing so, auditory CAPTCHAs have became too hard to human being to solve. Solving auditory CAPTCHAs is akin to Cocktail Party Effect, which refers to our ability to process main audio signals preferentially and ignore other irrelevant ones in noisy environments. This study explores the current designs of auditory CAPTCHAs to see how well our "cocktail party ability" performs when interacting with different CAPTCHA designs. An experiment with repeated measurement factorial design is conducted; thirty-six participants take part. The main signals, or the signals to be processed, are pronounced either by random male speaker (RMS), random female speaker (RFS), or mixed speaker (MS); while the interference signals, or the signals to be ignored, are pronounced either by random male (RMN), random female (RFN), or mixed noise (MN). Fifty percent of the interference contents sound similar to the main contents, while the other fifty percent are normal conversation noises. Error rates and subjective preferences are collected during the experiments. Results show that sound similarity is problematic; the error rates are significantly higher than its counterpart. The combination of RMS and RFN has significantly lower error rate due to greatest pitch difference; our participants also prefer this one for its relative easiness. On the other hand, for combination of RMS and RMN, the error rates are significantly higher and the preference scores lower. The results have important implications for auditory CAPTCHA design.

Keywords: auditory CAPTCHA, cocktail party effect, pitch difference

INTRODUCTION

While Internet-based services are penetrating into our lives, the security of these systems becomes an inevitable issue that merits our attention. To prevent abusers from exploiting these resources via automated programs, most service providers add a kind of security mechanism called "CAPTCHA" to ensure that the users about to access the services are genuine human users instead of automated programs. CAPTCHA is the acronym of "Completely Automated Public Turing test to tell Computers and Humans Apart". Its capability of telling computers and humans apart derives from the fact that it can propose questions that state-of-the-art technology can not solve perfectly while human can, and ask the potential user to answer the questions. If the replied answers are correct, the users in question will be deemed as human users; if the answers are incorrect, the users will be deemed as automated programs and be denied from accessing the systems. Currently, text CAPTCHAs and audio CAPTCHAs are the most common ones in the market. This research focuses on the auditory type.

https://openaccess.cms-conferences.org/#/publications/book/978-1-4951-2103-6

Audio CAPTCHAs present audio signals containing text as questions, and the potential users have to supply the text as answers to identify themselves as human users. Since text can be recognized by computer programs, the signals usually undergo certain distortion and background noise to prevent computer programs from successfully recognizing the text. Our ability to separate the distorted speech text from the background noise found in audio CAPTCHA is akin to a cognitive psychology effect called “cocktail party effect”, which refers to our ability to tune in to a specific conversation even in a noisy party (Cherry, 1953); several theories have been proposed to explain the phenomenon (Bronkhorst, 2000). We focus on the comparison of current audio CAPTCHA designs and use the cognitive psychology theories to interpret our findings.

EXPERIMENT

Thirty-six student participants, eighteen of them are male and the other eighteen female, take part in a two-factor within-subject experiment. The average age of the participants is twenty-one. The factor Speaker has three levels, which are random male speaker (RMS), random female speaker (RFS), and mixed speaker (MS) with male and female speakers. The factor Background Noise also has three levels, which are random male noise (RMN), random female noise (RFN), and mixed noise (MN). Participants are briefed about the experiment and sign consent forms before the experiment process begins. Each participant has to finish the nine combinations, each of which has five repetitions, to conclude the experiment. The CAPTCHA questions are rendered to a headphone wore by blindfolded participants to minimize visual distraction; answers to the questions are reported verbally to the experimenter and logged thereby. Dependent variables include error rates and preference scores of the nine combinations. Error rate is defined as the ratio of the numbers of error answers and total questions, while the preference score one denotes the least preferred and score nine denotes the most preferred.

RESULTS

Descriptive statistics of the error rate are shown in Table 1 Male voice for the signal and the background noise turns out to be the worst one in this table. Analysis of variance shows that both factors are significant in terms of error rates, as shown in Table 2. For the preference scores, RMS-RFN and RFS-RMN have high scores, while RMS-RMN and RFS-RFN have low scores, as shown in Table 3. The non-parametric analysis of the preference scores is shown in Table 4, from which we can see that there is significant difference.

Table 1: Descriptive statistics of the error rate

Combination	A (RMS-RMN)	B (RMS-RFN)	C (RMS-MN)	D (RFS-RMN)	E (RFS-RFN)	F (RFS-MN)	G (MS-RMN)	H (MS-RFN)	I (MS-MN)
Mean (SD)	0.268 (0.176)	0.013 (0.040)	0.113 (0.108)	0.020 (0.357)	0.139 (0.106)	0.074 (0.072)	0.112 (0.092)	0.111 (0.095)	0.070 (0.097)

Table 2: ANOVA table of the error rate

Source	**dof**	**F**	**Sig.**
Speaker	1.817	7.534	0.002*
Background noise	1.650	9.064	0.001*
Speaker * Background noise	2.928	46.871	0.000*

Table 3: Descriptive statistics of the preference scores

Combination	A (RMS-RMN)	B (RMS-RFN)	C (RMS-MN)	D (RFS-RMN)	E (RFS-RFN)	F (RFS-MN)	G (MS-RMN)	H (MS-RFN)	I (MS-MN)
Mean (SD)	1.944 (1.567)	8.083 (1.052)	4.833 (2.261)	7.944 (1.013)	2.889 (1.864)	4.917 (2.103)	4.889 (1.848)	4.889 (1.939)	4.611 (1.961)

Table 4: Non-parametric test of preference scores

Chi-square	**dof**	**Asym. Sig.**
154.467	8	0.000*
		* p<0.05

CONCLUSIONS

The results have practical implication for auditory CAPTCHA design. Greater pitch difference between the voices of signal speaker and the background noise has lower error rates and higher preference scores. Such finding is coherent with those in the field of human sensory research (Brokx & Nooteboom, 1981; Cainer, James, & Rajan, 2008; Helenius & Hongisto, 2004; Stevens, Lees, Vonwiller, & Burnham, 2005). Signals pronounced by male speakers seem to be more error prone than those by female speakers. Even the MS-MN combination has lower error rate than the all male one. If user friendliness is more concerned, RFS-RMN and RMS-RFN are good candidates; if security is of top priority, MS-MN is recommended.

REFERENCES

Brokx, J., & Nooteboom, S. (1981). Intonation and the perceptual separation of simultaneous voices: Institute for Perception Research.

Bronkhorst, A. W. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. Acta Acustica united with Acustica, 86(1), 117-128.

Cainer, K. E., James, C., & Rajan, R. (2008). Learning speech-in-noise discrimination in adult humans. Hearing research, 238(1), 155-164.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, 25(5), 975-979.

Helenius, R., & Hongisto, V. (2004). The effect of acoustical improvement of an open-plan office on workers. Paper presented at the Proceedings of Inter-Noise.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. Computer Speech and Language, 19(2), 129-146.