Computer Security Symposium 2017
23 - 25 October 2017

混合された環境音の聞き取りに基づく認証方式

古賀 千裕[1,a)]　佐藤 敬[1,b)]

概要：視覚障がい者でも利用できる CAPTCHA として，音声を利用した音声型 CAPTCHA が存在する．音声型 CAPTCHA では，単語の識別を利用した英数字識別型 CAPTCHA が主流であるが，この方式には 1) 音声の聞き取りが難しい，2) 解答に時間がかかるといった問題点が存在する．本論文では，“混合された環境音の聞き取り”を利用した新しい CAPTCHA 方式を提案する．この方式では，混合された環境音の関係の有無を決定する問題を利用する．環境音の識別は人間には容易だがボットには難しいことが期待される．また，Google の reCAPTCHA v2 と比較し，提案方式の性能，ユーザビリティと安全性を議論する．

キーワード：CAPTCHA, 環境音, アクセシビリティ

An authentication scheme based on recognition of mixed environmental sounds

Chihiro Koga[1,a)]　Takashi Satoh[1,b)]

Abstract: Audio-based CAPTCHAs can be used by visually impaired persons. A typical audio-based CAPTCHA method is based on speech recognition of spoken digits in noisy environment. But, this method has disadvantages; 1) It is difficult to recognize spoken digits, 2) It takes much time to answer. In this paper, we propose a novel CAPTCHA scheme based on “recognition of mixed environmental sounds.” Proposed scheme is using a decision problem with finding relationship between environmental sounds. It is expected to be easy for humans to solve the decision problem, but difficult for bots. We make a comparison our proposed scheme with Google’s reCAPTCHA v2 and discuss its performance, usability and security.

Keywords: CAPTCHA, Environmental sounds, Accessibility

1. はじめに

1.1 研究背景

今日，オンラインサービスにおいてボットが不正利用目的でアカウントを大量取得し，乱用するという問題が横行している．この防止策として，CAPTCHA(Completely Automated Public Turing Test To Tell Comouters and Humans Apart) [1] という技術が広く用いられている．CAPTCHA の代表例として歪んだ文字列画像の解釈を利用した方式や複数の画像から問題に沿った画像を選択する方式 [2]（以下，画像型 CAPTCHA と呼ぶ）が挙げられるが，これらの画像を使用した方式には視覚障がい者には解読困難である．そのため，画像の代わりに音声を利用した，音声型 CAPTCHA がある．音声型 CAPTCHA の中でも，英数字識別型 CAPTCHA [3,4] が主流な方式であるが，この方式には 1) 音声の聞き取りが難しい，2) 解答に時間がかかる，3) ボットによる攻撃に耐性がないといった問題点がある．

1.2 本研究の貢献

本研究では，“混合された環境音の聞き取り”を利用した CAPTCHA を提案する．雑音が加わった音声の識別に比べると，環境音の識別は人間には容易だが BOT には難しいことが期待される．また，環境音は音声に比べて感覚的に識別できるため，提案方式は既存方式に比べ解答容易性

[1] 北九州市立大学
The University of Kitakyushu, Kitakyushu, Fukuoka 808-0135, Japan
[a)] y7mca004@eng.kitakyu-u.ac.jp
[b)] tsatoh@kitakyu-u.ac.jp

が向上すると考えられる．

本論文では，以下の 3 点を明らかにする．

- 提案方式を実装し，人間による評価実験を行う．実験により解答時間，正答率，ユーザビリティの 3 点を評価し，結果を示す．
- 以上で挙げた評価項目を従来方式と比較し，提案方式の優位性を示す．
- 実験結果から提案方式における問題点を洗い出し，今後の課題を示す．

2. 関連研究

2.1 従来の音声型 CAPTCHA

音声型 CAPTCHA の中で広く用いられている方式として，英数字識別型 CAPTCHA が存在する．英数字識別型 CAPTCHA は，ランダムな英数字を音声で流し，それを解答させる方式である．音声に歪みを加えるなどして，コン ピュータによる識別を困難にする工夫がなされている．この方式は Yahoo! JAPAN [3] や Google [4] など，様々なウェブサービスで利用されている．

従来の音声型 CAPTCHA に代わる記憶作業を有しない文意文脈解釈問題を利用した方式として鴨志田らにより提案されたワードサラダ識別型 CAPTCHA [5] がある．ワードサラダ識別型 CAPTCHA は，人間が普段使用している自然言語の文（自然文と呼ぶ．）とコンピュータにより合成された文法は間違っていないが文章としての意味が破綻している文（ワードサラダ）のうち，どちらが自然か（もしくは不自然か）利用者に解答させる．この方式はコンピュータに解読困難とされる “文としての違和感”を人とコンピュータの識別に利用しているが，自然文をウェブ上にある文章から利用しているため，検索エンジンを使用した攻撃に弱い．この改善案として自然文に相当する文にもワードサラダを使用した方式 [6] が山口らにより提案されている．さらに，ワードサラダ識別型 CAPTCHA の改良として，ワードサラダを単語に応用した，単語とランダムな音韻列の識別を利用した音声型 CAPTCHA [7] が提案されている．

2.2 従来方式の問題点

従来の音声型 CAPTCHA には以下のような問題点が存在する．

(1) 聞き取りが難しい

音声認識技術の向上により，コンピュータは音声に多少歪みが加えられていても英数字を識別できるようになった．それに対抗しさらに音声の歪みを大きくしたために，英数字識別型 CAPTCHA の中には，音声の歪みが酷く人間でも解けないほどに難しいと指摘されるものも存在する．

(2) 解答に時間がかかる

音声型 CAPTCHA は FRR(False Recognition Rate) が高い方式であるため，認証されるまでに複数の問題を解く場合がある．また，聞き取った音声を一度覚える必要があり，聞き取る単語数が多いほど利用者に負担がかかってしまう．視覚障がい者は音声を聞き取りながらそれを入力する作業が困難であるため，記憶作業に対する負担はより大きなものとなる．そのため，画像型 CAPTCHA と比べて解答時間が増大しやすい．表 1 のとおり，音声型 CAPTCHA は画像型 CAPTCHA に比べて解答に時間がかかることが分かる．山口らにより提案されたワードサラダ識別型 CAPTCHA においては一回の試行に 3 分以上も時間がかかっている．

表 1 各方式の平均解答時間

	解答時間 (秒)
画像型 [9]	11
英数字識別型 [9]	43
ワードサラダ識別型 [6]	200
単語とランダムな音韻列 1 [7]	23.7
単語とランダムな音韻列 2 [7]	25.3

(3) ボットによる攻撃に脆弱商用で利用されている英数字識別型 CAPTCHA に対して，機械学習による攻撃成功例が報告されている．Sano ら [8] は，HMM（隠れマルコフモデル）を利用することにより，Google reCAPTCHA の攻撃に 58.75%の確率で成功したと報告している．Bursztein ら [10] は，Yahoo!，Microsoft，eBay の 3 つの方式に対して，45%，49%，83%の確率で攻撃に成功している．このように，機械学習により歪みが含まれた音声でも識別できるため，ボットによる攻撃に耐性が弱い．

3. 提案方式

本論文では，解答が人間には容易だがコンピュータには難しい音の素材として環境音を使用する．

3.1 概要

提案方式では，ある一つの環境音に対して，関係のある音を組み合わせた混合音 A と関係のない音を組み合わせた混合音 B の二つのうち，どちらが自然であるか（状況としてふさわしいか）利用者に解答させることによりコンピュータと人を識別する．以下に問題の例を示す．

A. 強風の音+荒波の音
B. 強風の音+ウグイスの鳴き声

ベースとなる環境音，強風に対し，関係のある荒波の音を加えた音 A と関係のないウグイスの鳴き声を加えた音 B のうち，利用者はどちらが自然か判断を行う．この例の場合，強風の中ウグイスの鳴き声が聞こえるのは不自然であるた め，“A が自然である”と解答すれば正解となる．以降より，問題のベースとなる音をベース音，ベース音に加

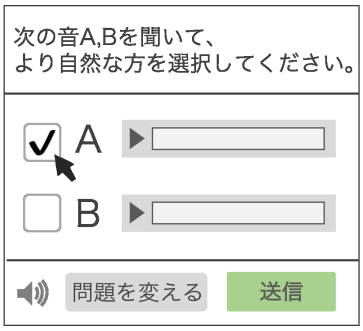

図 1 提案方式で想定される出題画面

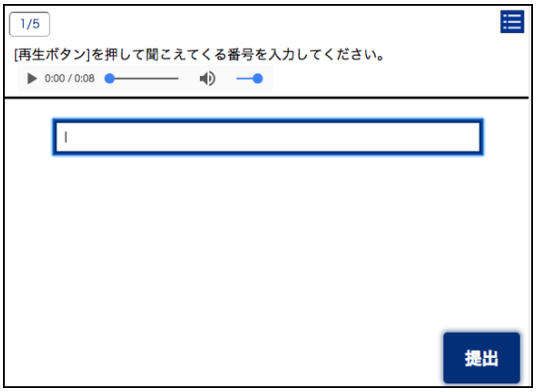

図 2 実験で使用した出題画面 (reCAPTCHAv2)

える音を 追加音，ベース音と追加音を合成した音を混合音と呼ぶ．また，提案方式は文化により誤答の可能性があるため，対象は日本人のみとする．

ここで，提案方式において想定される出題画面を図 1 に提示する．利用者は再生ボタンを押して混合音 A，B を聞き，設問に沿って解答を選択・送信する．送信された解答が正解であれば人間として認証される．また，必要に応じて問題やボリュームの変更ができるようボタンを設定しておく．

提案方式は 2 択の選択問題であるため，適当に解答しても 2 分の 1 の確率でコンピュータによる攻撃に成功する．通常 1 問あたりの攻撃確率が高い方式では利用者に複数の問題を与えるが，今回は 1 回の試行で解く問題数を 1 問と設定する．また，1 問あたりの平均解答時間 30 秒，人間の正答率 80%を目標として設定する．

3.2 期待される効果

(1) 解答時間の短縮

提案方式では，混合音 A，B の長さをそれぞれ 5 秒に設定しても，1 問あたりの音再生時間は 10 秒ほどである．また，解答方式は 2 択の選択問題であるため，音を聞いてから解答を送信するまでに時間がかからない．そのため，本方式における解答時間は目標の 30 秒以内に抑えることができると考えられる．

(2) 攻撃耐性の向上

雑音が加わった音声の識別に比べると，環境音の識別は人間には容易だがボットには難しいことが期待される．2.2 節で述べたように，雑音が加わっていても音声の識別はボットにとって容易であることが分かる．英数字識別型 CAPTCHA は 0 から 9 の数字 10 個の組み合わせからなっている．それに対し，環境音の数は無数に存在する．提案方式に対する攻撃に機械学習を用いる場合，英数字識別型 CAPTCHA と比べると，用意する学習用データは膨大な量が必要となる．このことから，提案方式は英数字識別型 CAPTCHA よりもボットによる攻撃に耐性があると期待できる．

4. 人間による評価実験

提案方式が認証方式として優れているか確認するため，Google reCAPTCHAv2 と提案方式を比較する以下の評価実験を行なった．

実験 1 人間による解答時間と正答率の評価
実験 2 人間によるユーザビリティの評価

4.1 実験方法

今回は，男性 5 名，女性 6 名の計 11 人に実験を行なってもらった．年齢構成は 19〜58 歳であり，視力の状態は 11 件全て晴眼であった．実験前に，被験者に実験の趣旨について説明を行った．説明内容は，音声型 CAPTCHA の概要，Google reCAPTCHAv2，提案方式についてである．被験者には各方式を公平に評価してもらうため，reCAPTCHAv2 を方式 1，提案方式を方式 2 として説明し，各方式の説明についても解答方法を伝えるのみとした．

4.1.1 実験 1

方式 1,2 の問題を用意し，被験者に解答してもらった．実験手順は以下の通りである．

(1) 練習

各方式 1 回ずつ練習問題を提示し，練習を行ってもらった．

(2) 実験

実験は，静かな環境でパソコン (イヤホンなし) を用いて行った．ここで，実験で使用した出題画面を図 2 と図 3，提案方式で利用した音サンプルの組み合わせを表 2 に示す．被験者に「解答が分かるまで音声は何回聞いても良い」ということを伝えた上で，各方式ともに 5 回の試行を行ってもらった．ここでは解答時間と問題の正答率を測った．

4.1.2 実験 2

実験を通して方式 1 と方式 2 のどちらが有効であるか 3 つの項目，1) 方式の分かりやすさ,2) 被験者が主観的に感じた解答時間，3) 解答の容易さに関して 5 段階で相対的に評価を行ってもらった．また，各問の評価の理由について

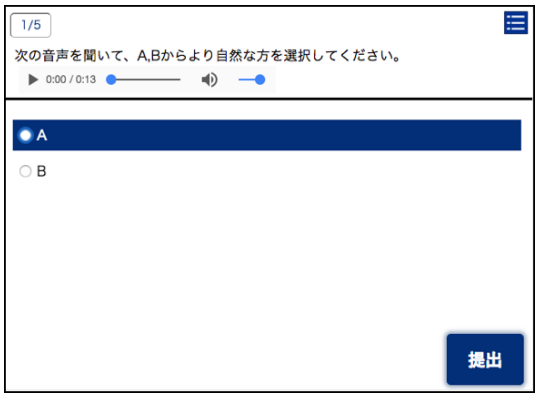

図 3　実験で使用した出題画面 (提案方式)

表 2　実験で使用した混合音の組み合わせ (提案方式)

問題	ベース音	追加音	
練習	ホラー BGM	自然	悲鳴
		不自然	笑い声
1 問目	船の汽笛	自然	波
		不自然	山の渓流
2 問目	花火	自然	太鼓
		不自然	雨
3 問目	交差点	自然	車
		不自然	馬の足音
4 問目	火事	自然	救急車
		不自然	焼き芋屋台の BGM
5 問目	強風	自然	荒波
		不自然	ウグイス

具体的に回答してもらった．

4.2　実験結果

4.2.1　実験 1

方式 1 の平均解答時間と正答率を表 3 に，方式 2 の平均解答時間と正答率を表 4 に示す．解答時間に関しては目標としていた 30 秒を切り，reCAPTCHAv2 より 3.4 秒短縮することに成功した．しかし，正答率は 74.55%と reCAPTCHAv2 に及ばない結果となった．

4.2.2　実験 2

アンケート結果を図 4 に示す．問 1 の “どちらの方式の方が分かりやすいですか”といった質問に対しておよそ 90%もの人が reCAPTCHAv2 の方が分かりやすいと回答した．reCAPTCHAv2 の方が分かりやすいと回答した被験者の中では「reCAPTCHAv2 は設問で言われたことをそのまま実行すればよかった」,「提案方式はどのような音声が流れるかわからなかった」といった意見が多く見受けられた．問 2 の “解答に時間がかからなかったのはどちらの方式ですか (主観的に)”という質問に対しては,「提案方式はどちらが自然であるか迷う問題があった」といった提案方式における問題の難易度に対する指摘があった．しかし,「提案方式は感覚的に判断できたが，reCAPTCHAv2 は数字を一度覚える必要があった.」という提案方式に対

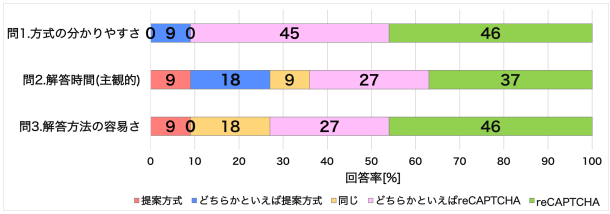

図 4　アンケート結果

して肯定的な意見も上がっている．問 3 の解答方法の容易さを問う質問に対しても，問 1,2 と同じように提案方式の分かりづらさを指摘する意見が見受けられた．ユーザビリティ評価に関しては，提案方式に対する肯定的な意見はありながらも reCAPTCHAv2 に及ばない結果となった．

4.3　考察

4.3.1　解答時間について

実験で得られた結果から考察を行う．CAPTCHA の評価を行う際に，注目すべきもう一つの指標としてリトライ率 (一問あたり何回音声を聞いたか) が挙げられる．reCAPTCHAv2 は音声から流れるすべての数字を正確に聞き取らなければならないため，数字を一つでも聞き逃すとまた最初から音声を聞かなければいけない．そのためリトライ率が高い方式であると言える．今回の実験結果からもそのことが分かる．今回実験に使用した各方式の問題の音再生時間の平均は，reCAPTCHAv2 9.2 秒, 提案方式 13.6 秒である．表 3 の解答時間と比較すると，提案方式の解答時間は音声再生時間の約 1.5 倍であるのに対し，reCAPTCHAv2 は音声再生時間の 2 倍以上も解答に時間がかかっている．このことから reCAPTCHAv2 は提案方式と比べリトライ率が高いことが分かる．この理由から提案方式において reCAPTCHAv2 より解答時間を短縮することが出来たとも考えられる．

4.3.2　正答率について

提案方式において正答率が低迷した原因は，難易度の設定にあると考えられる。表 4 を見ると 1 問目と 2 問目の正答率がそれぞれ 63.64%,45.45%と全体の正答率を下げる要因になっていることが分かる．被験者からも「提案方式の問題は 1,2 問目の難易度が高かった.」という声が上がっており，一番正答率の低かった 2 問目では、花火と組み合わせると雨の音が設置型の花火にも聞こえる．このことから音を合成する際には，ベース音と追加音の関係性だけでなく，音としての相性も加味する必要がある．また，正答率低迷のもう一つの要因として考えられるのは，提案方式の設問の曖昧さである．ほとんどの被験者が設問を読んだだけでは提案方式を理解することが出来ず，ルールが分からないまま実験に挑んでいた．問題の難易度が下がった 3 問目あたりでルールを理解したため，3〜5 問目の正答率は高くなっている．提案方式における設問の問題点については

表 3　解答時間と正答率 (reCAPTCHAv2)

	1 問目	2 問目	3 問目	4 問目	5 問目	全体平均
解答時間 (秒)	31.31	31.87	22.01	20.11	17.97	24.65
正答率 (%)	90.91	90.91	81.82	72.73	100	87.27

表 4　解答時間と正答率 (提案方式)

	1 問目	2 問目	3 問目	4 問目	5 問目	全体平均
解答時間 (秒)	23.48	21.35	19.87	22.25	19.3	21.25
正答率 (%)	63.64	45.45	90.91	81.82	90.91	74.55

以下の段落で述べる．

4.3.3　個人差について

reCAPTCHAv2 で利用される英語の発話は非常にネイティブであるため，英語に触れる機会の少ない人には聞き取りが難しい．今回実験に協力してもらった 50 代男性, 女性は普段英語に触れる機会が無いため，reCAPTCHAv2 における解答時間は 40.69 秒と全体の 2 倍近くの時間がかかっていた．正答率も 50%と reCATCHAv2 の英語は彼らにとって聞き取りが難しかったことが分かる．それに対し，提案方式における解答時間と正答率は 20.61 秒,80%と平均よりも高い値であった．このことから，提案方式は普段から誰もが聞き慣れた環境音を使用するため，年代や学歴などが解答時間, 正答率に影響することが少ないと考えられる．

4.3.4　ユーザビリティについて

被験者が提案方式に対してあげた否定的な意見の中で最も多く見受けられたのは，提案方式の設問における「自然」という曖昧なワードに対する指摘である．提案方式の設問では，流れる音の内容とどのような行動を取れば良いか具体的に示すものが無かった．それに対し，reCAPTCHAv2 では流れる音声と, その音声を聞いた後に取るべき行動を明確に示している．そこで，提案方式における設問を以下のように変更することを考えた．

訂正前: 次の音声を聞いて，A,B からより自然な方を選択してください．

訂正後: 次の音 A,B は, ある環境音に別の音を加えたものです．この 2 つのうち状況としてふさわしい方を選択してください．

流れる音の内容を「環境音に別の音を加えたもの」，音を聞いた後に解答を選択する基準を「状況としてふさわしい方」と明確に示すことで，提案方式における方式の分かりにくさは改善されると考えられる．

5.　提案方式の安全性について

ボットによる攻撃に対する提案方式の安全性について議論する．提案方式に対する攻撃手法として，以下の手順が考えられる．

(1) 混合音を分離する：混合音を合成前の 2 つの音に分離する．
(2) 分離した音を解析する：分離された音がそれぞれ何を表す音なのか解析する．
(3) 音の関係性を判定する：混合音 A，B のどちらが関係が深いのか判定する．

以上の手順 (1)〜(3) に対して，それぞれ独立成分分析，機械学習，検索エンジンを用いた攻撃が考えられる．この中で，機械学習を用いた音響のシーン識別が最も困難であると言える．音響によるシーン識別などの精度を競うコンテスト DCASE2016 Challenge [11] では，音のデータセットを 15 個のクラスタに分類した時の平均精度は 72.5%であった．しかし，公園や電車など特定のシーンにおいては 13.9%，33.6%とあまり高い精度が得られていない．このように，環境音のシーン識別はまだ難しいということが分かる．混合された環境音のシーン識別と混合前の環境音同士の関係の有無の決定は，単なる環境音のシーン識別に比べてより難しい問題である．さらに，(1)〜(3) の攻撃手順全てに成功しなければ，提案方式の攻撃に成功しないと考えられる．そのため，提案方式はボットによる攻撃に耐性があると期待できる．

6.　おわりに

本論文では，視覚障がい者でも利用できる認証方式として英数字識別型 CAPTCHA を例にあげ，この方式が抱える問題点を指摘した．その上で解答が人間に易しくボットに難しいと考えられる “混合された環境音の聞き取り” を利用した新しい認証方式を提案し、人間による評価を行なった．その結果，提案方式における解答時間が reCAPTCHAv2 より 3.4 秒短縮できることを確認した．一方，設問や音の構成を改善するなどして提案方式におけるユーザビリティの向上をはかる必要性が明らかとなった．前述した攻撃手法を実装し提案方式の安全性を評価することが今後の課題である．

参考文献

[1] Luis Von Ahn, Manuel Blum, Nicholas J. Hopper, John Langford. using hard AI problems for security, In Proceedings of EUROCRYPT'03, pp. 294-311, Springer-Verlag Berlin, 2003.

[2] Jeremy Elson, John R. Douceur, Jon Howell and Jared Saul. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization, In Proceedings of 14th ACM Conference on Computer and Communications Security (CCS), ACM, 2007.

[3] Yahoo!JAPAN, "画像や音声による認証に関するヘルプ," https://captcha.yahoo.co.jp/help/captcha_help.php, (参照 2017-08-26).

[4] Google, "ReCAPTCHA を使ったサンプルフォーム," https://www.google.com/recaptcha/api2/demo, (参照 2017-01-31).

[5] 鴨志田芳典, 菊池浩明. マルコフ連鎖による合成文章の不自然さを用いた CAPTCHA の提案と安全性評価. 情報処理学会論文誌, Vol. 54, no. 9, pp. 2156–2166, 2013.

[6] 山口通智, 岡本健, 菊池浩明. 機械合成文の不自然度相対識別問題に基づく CAPTCHA の提案. 情報処理学会論文誌, Vol. 56, no. 9, pp. 1834–1845, 2015.

[7] 山口通智, 菊池浩明. 多様な話者により発話されたランダムな音韻列と単語の識別問題を用いた音声型 CAPTCHA の研究. Computer Security Symposium 2016, pp. 363–370, 2016.

[8] Shotaro Sano, Takuma Otsuka, Katsutoshi Itoyama and Hiroshi G. Okuno. HMM-based Attacks on Google's ReCAPTCHA with Continuous Visual and Audio Symbols. Journal of information processing, Vol. 23, no. 6, pp. 814–826, 2015.

[9] Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell and Dan Jurafsky. How Good are Humans at Solving CAPTCHAs? A Large Scale Evaluation. IEEE Computer Society Washington, pp. 399–413, 2010.

[10] Elie Bursztein, Romain Beauxis, Hristo Paskov, Daniele Perito, Celine Fabry and John Mitchell. The Failure of Noise-Based Non-Continuous Audio Captchas. 2011 IEEE Symposium on Security and Privacy, pp. 19–31, 2011.

[11] IEEE, "DCASE 2016 ," http://www.cs.tut.fi/sgn/arg/dcase2016/, (参照 2017-02-06)