3W-08

モノラル音声信号における複数話者の同時発話による音声 CAPTCHA の頑健性向上

挾間晃彦† 梅澤猛‡ 大澤範高‡

千葉大学大学院融合科学研究科† 千葉大学大学院工学研究院‡

1. はじめに

悪意のあるプログラムによる不正な Web アクセスを防ぐために広く利用されている文字判読型の CAPTCHA(Completely Automated Public Turing test to tell Computers and Humans Apart)は，視覚障がい者のアクセシビリティを確保できない．そのため，音声を使った CAPTCHA も存在しており，図 1 に示すような聞き取り対象となる音声にランダムノイズを重畳させた音声を使う手法が一般的である．しかし，隠れマルコフモデルを用いて 58%の精度で自動認識可能[1]と報告されるなど，攻撃耐性に課題がある．

そこで，本研究では，音声信号処理においてモノラル音声の音源分離が困難であることに着目し，複数話者による同時発話を聞き分ける音声 CAPTCHA によって，視覚障がい者のアクセシビリティの確保と自動音声認識への耐性向上を両立させることを目指す．

2. 関連研究

2.1. 音声 CAPTCHA

福岡らは，音声 CAPTCHA に対する音韻修復効果の有効性を検討した[2]．音韻修復とは，音声の一部を削っても別の音を挿入すると音声が滑らかに聞こえ，聴取が容易になるという現象である．原音声から多くの音を削ると，人も機械も認識が困難になるが，削除した音声部分に雑音を挿入すると人だけが音韻修復効果により聴取が容易になると報告されている．

Meutzner らは，人が日常的に聞き分けることに慣れている残響に注目し，残響を含んだ CAPTCHA 音声が人と機械の認識差を作り出せるか検討した[3]．実験により，妥当な残響時間を設定することで人と自動音声認識の認識率に差異を作り出せることを示した．

Evaluation on Robustness of Single-Channel Audio CAPTCHA Using Simultaneous Utterance

† Hasama Akihiko, Graduate School of Advanced Integration Science, Chiba University

‡ Umezawa Takeshi, Osawa Noritaka, Graduate School of Engineering, Chiba University

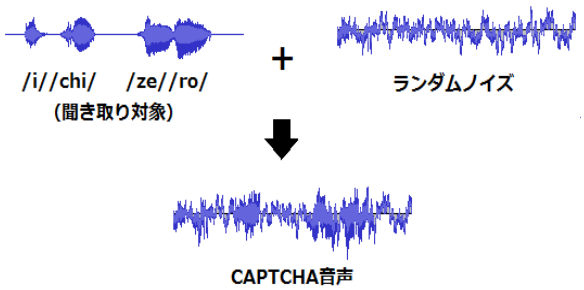

図 1. ランダムノイズを重畳させた音声信号の例

2.2. 自動音声認識

雑音，残響の除去や同時発話の音源分離は，自動音声認識分野の課題とされている．雑音，残響の除去に対しては，ディープニューラルネットワークを利用したデノイジングオートエンコーダが非常に有効であることが示されており[4]，現在一般的なランダムノイズを使ったものや，音韻修復効果，または残響を利用したものは音声 CAPTCHA の手法としては有効ではないと考えられる．

一方で，同時発話の音源分離については，モノラル信号の音源分離は不良設定問題であり，音源に含まれる性質を手がかりに分離する必要がある．そのため，複数のマイクロフォンを使い多チャンネルの信号を入手することで，音の定位を利用して分離する手法が提案されている．本研究では，モノラル信号の音源分離が困難であることを利用する．

3. 提案

図 2 に提案する音声 CAPTCHA の音声信号の例を示す．2 人以上の異なる話者によって発声された短いモーラからなる単語音声をモノラル音声にミキシングすることで，自動音声認識の精度の低下を図ると同時に，一試行当たりの再生時間を短くすることにより人が解答するときの負荷を抑えることを狙う．本稿では，同時発話に対する人の聴取能力を調査し，提案の有効性を検討する．

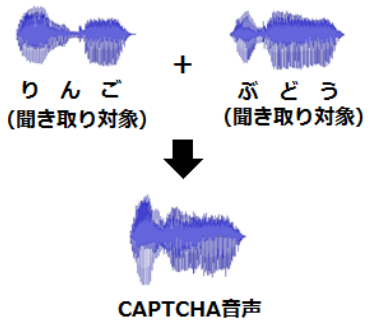

図 2. 同時発話の音声信号の例

4. 実験

単語音声の同時発話に対する人の聴取能力を調査するために，被験者実験を行った．

4.1. 実験条件と手順

男女 2 種類の合成音声を用い，3 モーラの日常的な日本語名詞の単語音声を重複なく各 50 個用意した．単語音声の作成にはフリーの音声合成ソフト「SofTalk」を用い，女声(rm3)と男声(m5)を使用した．2 種類の単語音声から無作為に抽出した各 1 個を同時に再生して音声をターゲット音声とした．実験を通して，ターゲット音声を構成する単語音声の組み合わせには同一のものを選ばないものとした．

健常な 20 代の日本人男性 6 人を被験者とし，ターゲット音声をヘッドフォンにより聞き，2 つの単語をキーボードを使った文字入力によって解答するよう指示した．1 つのターゲット音声につき，最大 5 回の聞き直しを許し，1 回の聴取に対する解答は 1 回とした．2 つの単語を間違えず入力するか，5 回の聴取を行うかで 1 タスク完了とし，被験者 1 人あたり 20 回タスクを繰り返すこととした．

4.2. 実験結果

各被験者の正解率を表 1 に示す．正解率は，全タスクに対して 2 単語正解したタスクの割合，1 単語のみ正解したタスクの割合，女声と男声それぞれについて正解したタスクの割合を示した．2 単語正解率は低かったが，1 単語のみの正解率を含むと全被験者が 50%を超えた．被験者ごとに女声と男声の正解率の優劣は異なり，合成音声の違いによる偏りはみられなかった．また，実験に用いた全ターゲット音声に含まれる単語音声の中で，正解率が高かったものと低かったものそれぞれについて出現数が多い順に表 2 に示す．

5. 考察

表 1 の結果から， 3 モーラの単語の同時発話に対する聴取能力が低く，2 単語を聞き取ることは困難であることが示唆された．また，表 2 について，誤答が多かった単語の解答状況を調べたところ，「ぶどう」に対する「ごぼう」のようにアクセントと母音が似た単語と誤っている事例が多くみられた．このことから，音韻が似た単語が存在する場合は，それらの単語は採用しないほうがよいと考えられる．

表 1.被験者ごとの正解率(%)

	2 単語	1 単語	女声	男声
被験者 A	30	30	45	45
被験者 B	15	55	55	30
被験者 C	15	40	25	45
被験者 D	55	40	80	70
被験者 E	35	55	80	45
被験者 F	20	35	45	30

表 2.単語ごとの出現／正解数

単語	出現数	正解数	単語	出現数	正解数
せかい	5	5	ぶどう	7	1
なまえ	5	5	にほん	7	1
ほうき	5	5	ちらし	5	1
じかん	5	4	まぐろ	5	1
すいか	5	4	すずめ	5	2
つばめ	5	4	ろしあ	5	2
てにす	4	4	りぼん	4	0

6. まとめと今後の課題

音声 CAPTCHA の耐性向上のため，新たな音声 CAPTCHA として，モノラル音声における複数話者による同時発話の利用を提案した．本稿では，人の同時発話に対する聴取能力の調査を行い，音韻が聴取に影響を与えていることが示唆された．今後は，モーラ数の変化や音韻に基づく単語の選択が同時発話の聴取精度に影響を与えるか実験を行い調査する．その後，人の聴取に有効な音声を用い，自動音声認識による攻撃耐性を評価し，提案の有効性を評価する．

参考文献

[1] Shotaro Sano, Takuma Otsuka, Katsutoshi Itoyama, 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.1-13, 2015

[2] 福岡千尋, 西本卓也, 渡辺隆行, "音韻修復効果を用いた音声 CAPTCHA の検討", ヒューマンインタフェース学会研究報告集, Vol.10, No.6, pp.83-88, 2008

[3] Hendrik Meutzner, Santosh Gupta, Viet-Hung Nguyen, Thorsten Holz, Dorothea Kolossa, “Toward Improved Audio CAPTCHAs Based on Auditory Perception and Language Understanding” ACM Transactions on Privacy and Security, Volume 19 Issue 4, Article No. 10, February 2017.
(doi:10.1145/2856820)

[4] 小宮山大樹, 篠崎隆宏, 堀内靖雄, 黒岩眞吾, ”Denoising Autoencoder による残響除去の大語彙音声認識における評価”, 日本音響学会 2013 年秋季講演論文集, 1-P-4d, pp.131-132, September 2013.