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

山口 通智[1]　菊池 浩明[1]

概要：音声型 CAPTCHA は，視覚障害者のウェブアクセシビリティの確保と，オンラインサービスにおける機械の不正防止を実現する重要な技術である．本研究では，多様な話者に対する音声認識の難しさを利用した音声型 CAPTCHA を提案する．提案方式では，音声合成技術を利用し，発話速度の変化や外国人話者の選択により，多様な話者を模擬する．また，単語とランダムな音韻列の識別という意味論的な作問をすることで，人間の正答率とユーザビリティの改善や安全性の向上を図る．本稿では，実験を通して提案方式と既存方式の比較を行い，識別問題と話者の多様性を発話速度で模擬した提案方式が優位であることを示す．

キーワード：CAPTCHA，音声認識，音声合成，複数話者，意味論的識別問題，アクセシビリティ

Study on Audio-CAPTCHA by Differentiating Random Phoneme Sequences and Semantic Words Spoken by Multi-speaker

Michitomo Yamaguchi[1]　Hiroaki Kikuchi[1]

Abstract: Audio-CAPTCHA prevents malicious bots from attacking Web services and provides Web accessibility for visually-impaired persons. In this paper, we utilize the difficulty for recognizing voices spoken by multi-person as a CAPTCHA. Our proposal synthesizes various voices by changing voice speed and employing non-native speakers. Moreover, we employ semantic identification problems between random phoneme sequences and semantic words to improve the accuracy of humans, usability and security. We evaluated our scheme in several experiments.

Keywords: CAPTCHA, Automatic Speech Recognition, Speech Synthesis, Multi-speaker, Semantic Identification Problem, Accessibility

1. はじめに

1.1 背景

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) [1] は，人間と機械 (ソフトウェアによる自動化エージェント) を識別するテストである．CAPTCHA は，機械によるサービスの不正利用を防止するために利用されている．

音声型 CAPTCHA は，画像型 CAPTCHA を解けない視覚障害者らに対して，ウェブアクセシビリティを提供する重要な技術である．音声型 CAPTCHA の仕組みは，何らかの方法で歪めた音声に含まれる単語を利用者に認識させる方式である．しかしながら，音声の歪みが酷く人間にも解けないほど難しいとの指摘 [2], [3] や，機械学習を用いた攻撃 [4], [5], [6] により危殆化が進むなど，その構成の難しさが知られている．

1.2 音声型 CAPTCHA の問題点

本稿では，まず音声型 CAPTCHA の一般的な問題について，以下に整理する．

問題点 *(1)* **記憶作業を要するため，ユーザビリティが悪い:**

[1] 明治大学大学院先端数理科学研究科現象数理学専攻
Department of Mathematical Modeling, Analysis and Simulation, Graduate School of Advanced Mathematical Sciences, Meiji University, Tokyo, #164-8525 Japan

人間が音声型 CAPTCHA に解答するには，認識した単語を記憶する必要がある．複数単語を解答する場合は，人間にとってその負荷は高くなる．音声は画像に比べて一度に認識できる情報が制限されるため，確認にも手間を要する．視覚障害者らは，聞き取りながら解答することが難しい [*1]ため，その記憶に要する負荷はさらに高い．

問題点 *(2)* 単語認識は会話認識と比べて，人間には難しく，機械には易しい*:* 人間は機械に比べて，意味論的な解釈など，より高度な認知能力をもつと考えられる．例えば人間は，会話中に多少の聞き間違いがあっても，文脈を解釈して認識結果を矯正できる．単語認識では，意味論的な解釈が使えないため，人間は軽微なものでも聞き間違いを修正できない．機械については，意味論的な解釈が不要な場合の方が，音声認識の精度が向上することが知られている [4], [6].

1.3 *MGK* 方式の分析と本研究の着想

次に，Meutzner らにより CHI '15 で提案された *MGK* 方式 [7] について前節で示した問題点を考察し，本研究の着想を示す．

MGK 方式では，ランダムな音韻[*2]列と単語を重畳なしに連結した音声を使用する．1 つの音声には，複数の単語やランダムな音韻列が含まれる．利用者は，音声中に含まれる単語すべてを表記で解答する．

MGK 方式には，問題点 (1) の指摘が該当する．問題を体験した被験者からは，CAPTCHA というより記憶力を問われている気がしたという意見があった．

MGK 方式は，ランダムな音韻列の挿入により，問題点 (2) の解決を図ってる．HTK (付録 A.1) に代表される音声認識方式では，単語認識で意味論的な解釈はしないため，音韻を含む音を非言語雑音として無視しない．よって，HTK はランダムな音韻列を単語として認識してしまい，音声の正しい認識ができない．

着想 *(1)* ランダムな音韻列と単語の識別問題の導入*:* 本研究では，問題点 (1) の解決を図るため，ランダムな音韻列が単語として誤認識される点を利用し，単語とランダムな音韻列の意味論的な識別問題を提案する．識別問題では，単語の表記を解答する必要がないため，記憶の負担を大幅に削減できる．また，問題点 (2) についても，識別問題ならば単語の正確な認識が不要なので，軽微な聞き間違いによる人間の正答率低下を防止できる．

注意すべきは，*MGK* 方式では単語自体には音声に歪みを与えていない点である．著者らが *MGK* 方式を実装し HTK による単語のみの認識率を調べると，55–71 % 程度の正答率を維持していた．識別問題の場合，特定の識別対象の検出率が高いと，その知識を使用した攻撃に対して脆弱になってしまう．

着想 *(2)* 音声認識の難易度が高い，多様な発話の音声の生成*:* 音声認識は，特定話者より不特定話者の方が困難である [8], [9]．音声認識では，話者の特徴をモデル化した音響モデルを使用する．音声から得られる特徴量の揺らぎが大きくなれは，特徴と音韻の結びつきが希薄になり，認識率が低下する．一方で人間は，多様な話者の音声を認識できる．例えば人間は，学習中の非母国語話者の発話でも認識できるが，機械でそれは困難である．

1.4 本研究の貢献

本研究では，ランダムな音韻列と単語の意味論的な識別問題を用いた音声型 CAPTCHA を提案する．意味論的な解釈問題は，機械の攻撃に対して頑強になる．識別問題では正確な単語認識が不要なため，人間の記憶に要する負担の削減や，軽微な聞き間違いによる正答率低下を防止できる．

提案方式では，機械による識別問題の解答を困難にするため，音声合成によって，多様な話者を模擬する．音声の生成は，速度やピッチ，話者の母国語の選択により，様々な音声を合成する．

本稿では，提案方式の実装と実験を通して，次のことを明らかにする．

- 多様な話者を模擬するため，音声速度やピッチを変える方式と，外国人話者を選択する方式を実装する．実験を通して，安全性や人間の正答率を調査し，より適した方式を明らかにする．
- 実験を通して，提案方式と *MGK* 方式を比較し，提案方式の優位性を示す．
- 被験者らの主観的意見や CAPTCHA の解答時間から，提案方式のユーザビリティに関する長所・短所を明らかにする．

2. 関連研究

音声型 CAPTCHA は，視覚障害者向けの方式として研究されてきた．

Holman ら [10] は，サイレンや鳥などの身近な事物を画像と音声の両方で提示し，そのいずれによっても解答可能とする方式を提案した．Shahreza ら [11] は，画像や音声の認識に加え，発話を組み合わせた方式を提案した．

Ximenes [12] らは，Knock Knock Jokes の解釈能力を用いた方式を提案した．音声認識の困難性に基づいた方式ではないが，駄洒落という音的な特徴を利用している．

福岡ら [13] は，音声の一部を削除しつつ白色雑音を挿入することで，音声を歪める方式を提案した．音韻修復効果 [14] により，単純な音声の削除に比べて，この方式は人間に認識しやすい．

視覚障害者にも利用可能な CAPTCHA としては，音

[*1] 視覚障害者らが使用するスクリーンリーダでは，キーボードの入力結果を読み上げる．そのため，音声の聞き取り中に解答すると，自らの操作音が音声認識の邪魔になる．

[*2] 意味の弁別をなす最小の音声単位 (phoneme).

声型 CAPTCHA 以外の方式も提案されている．参考文献 [15], [16], [17], [18] では，人間の文脈解釈能力を CAPTCHA として利用した．

音声型 CAPTCHA は，ユーザビリティや安全性に関する研究もなされてきた．

ユーザビリティに関しては，人間にも解けないほど難しい音声型 CAPTCHA の存在が指摘されている [2], [3]．また，NFB (National Federation of the Blind) による CAPTCHA のアクセシビリティに関する指摘 [19] は，音声型 CAPTCHA が人間に難しいことを世界的に周知した．

安全性に関しては，機械学習を利用した方式が示されている [4], [5], [6]．佐野ら [6] は，HMM (Hidden Markov Model) に基づく機械学習を用いた攻撃に対抗する CAPTCHA の設計に関して，白色雑音などの統計的な雑音より，歌などの意味論的な雑音の挿入が有効であるというガイドラインを示した．

3. 安全性定義と評価方法

本稿では，鴨志田ら [16] の安全性定義を利用する．本稿では，*Ham* をある言語に属する単語とし，*Spam* をランダムな音韻列とする．

X を出題音声を表す確率変数，Y を解答を表す確率変数，H を *Ham*，S を *Spam* とする．作問者が正答が *Ham* となる文を出題して，利用者が *Ham* と解答する条件付き確率は，$P(Y = H|X = H)$ と表せる．問題中の *Ham* と *Spam* 数をそれぞれ h, s とし $z = h + s$ とすれば，*Ham* と *Spam* を出題する確率はそれぞれ，$P(X = H) = h/z, P(X = S) = s/z$ となる．CAPTCHA の成功率は，これらの同時確率で，次のように与える．

$$
\begin{aligned}
P(Y = H, X = H) &= P(Y = H|X = H)P(X = H) \\
P(Y = S, X = H) &= P(Y = S|X = H)P(X = H) \\
P(Y = H, X = S) &= P(Y = H|X = S)P(X = S) \\
P(Y = S, X = S) &= P(Y = S|X = S)P(X = S)
\end{aligned}
$$

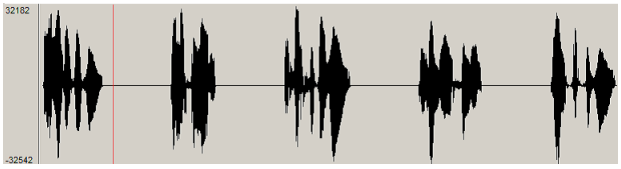
人間による解答 Y_h に対する，その CAPTCHA 1 問あたりの失敗率を，

$$
P_h = P(Y_h = S, X = H) + P(Y_h = H, X = S) \tag{1}
$$

とする．同様に，機械による解答 Y_m に対する，その CAPTCHA 1 問あたりの成功率を，以下のように表す．

$$
P_m = P(Y_m = S, X = S) + P(Y_m = H, X = H) \tag{2}
$$

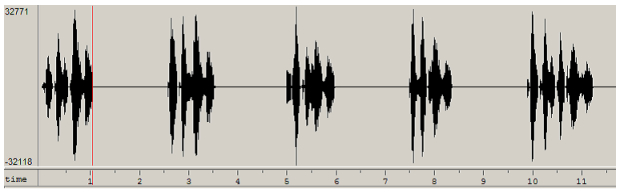
機械による攻撃成功率については，音声認識ツールによる補助を受ける場合を含めて検討する．HTK による音声認識処理を表す確率変数を W とする．ある音声に対して正しい認識結果を得る事象を $W = t$，そうでない事象を $W = f$ とすれば，式 (2) の右辺の項は，



2 番目: “たのしめる”, 5 番目: “がいこくじん”, 1,3,4 番目: ランダムな音韻列

図 1 提案方式 1 の音声ファイルの波形例

Fig. 1 Example of a Waveform regarding our Proposal 1.



1 番目: “ふたりとも”, 4 番目: “ただちに”, 2,3,5 番目: ランダムな音韻列

図 2 提案方式 2 の音声ファイルの波形例

Fig. 2 Example of a Waveform regarding our Proposal 2.

$$
\begin{aligned}
&P(Y_m = S, X = S) \\
&\quad = P(Y_m = S|W = t)P(W = t|X = S) \\
&\qquad + P(Y_m = S|W = f)P(W = f|X = S)
\end{aligned}
$$

$$
\begin{aligned}
&P(Y_m = H, X = H) \\
&\quad = P(Y_m = H|W = t)P(W = t|X = H) \\
&\qquad + P(Y_m = H|W = f)P(W = f|X = H)
\end{aligned}
$$

となる．この導出方法は [16] に示されているので，本稿での記載は省略する．

人間による CAPTCHA の正答数が $k < \theta$ となる確率を，人間拒否率 *FRR* (False human Rejection Rate) と定める．ロボットによる CAPTCHA の正答数が $k \geq \theta$ となる確率を，機械受入率 *FAR* (False machine Acceptance Rate) と定める．このとき，*FRR* および *FAR* を，確率 P_h および P_m の二項分布で，次のように与える．

$$
FRR = \sum_{k=\theta}^{z} \binom{z}{k} P_h^k (1 - P_h)^{z-k}, \quad FAR = \sum_{k=\theta}^{z} \binom{z}{k} P_m^k (1 - P_m)^{z-k}
$$

本稿では，既存方式と提案方式の比較を容易にするため，$\theta = 1, z = 1$ での *FRR* と *FAR* から，次の *F*-値を用いる．

$$
F = \frac{2 \cdot (1 - FAR) \cdot (1 - FRR)}{(1 - FAR) + (1 - FRR)} \tag{3}
$$

4. 提案方式

4.1 提案方式の特徴

多様な話者を模擬した合成音声の利用: 音声認識は，特定話者より不特定話者の方が困難であることが知られている．

参考文献 [8] によれば，発話速度は個人差の 1 要因であり，音声認識率に影響を及ぼす．発話速度は，音声の再生速度やピッチの調整により，多様な変化を得ることができる．

参考文献 [9] からは，共通音響モデルを用いた複数方言の認識結果を知ることができる．これは，不特定話者に対する音声認識の難しさの 1 例である．本研究では，ある言語の単語を非母国語話者に発話させることで，方言よりもさらに多様な発話の種類を利用する．外国語話者を含めた音響モデルの作成には，母国語話者単一のモデルに比べて，音韻やイントネーションなどでより多くの情報を加味せねばならない．よって，適切な音響モデルの作成は，より困難になると期待できる．

ランダムな音韻列と単語の識別問題の利用: Darlene ら [20] のカクテルパーティ効果に関連した研究結果によれば，人間はより親しみのある語句に反応する傾向がある．人間にとって意味を持つ単語は，ランダムな音韻列に比べれば，より親しみがあると考えられる．本研究では，この効果により，ランダムな音韻列と単語の識別が人間には容易になることを期待する．

4.2 方式定義

提案方式のアルゴリズムを以下に示す．

提案方式のアルゴリズム

(1) 言語 $\mathcal{L}$ に属する単語の集合 $\mathcal{S}_h$ と，ランダムな文字列の集合 $\mathcal{S}_s$ を作成する．

(2) $\mathcal{S}_h, \mathcal{S}_s$ の文字列を音声に変換する．

提案方式 1 $\mathcal{L}$ を母国語とする話者で音声合成する．その際，発話速度とピッチを無作為に変更する．

提案方式 2 $\mathcal{L}$ を母国語としない話者で音声合成する．発話速度とピッチの変更は提案方式 2 でも適用可能だが，それぞれの効果を明確にするため，本稿ではこれらを分けて扱う．

(3) $\mathcal{S}_h$ から h 個，$\mathcal{S}_s$ から s 個の音声を取り出し，それぞれ $\{Ham_0, Ham_1, \ldots, Ham_{h-1}\}$, $\{Spam_0, Spam_1, \ldots, Spam_{s-1}\}$ とする．$z = h + s$ とする．これらの和集合から，重複なく無作為に要素を取り出し，重畳なく音声を連結する．各音声同士は，適当な無音区間を挟むようにする．連結した音声ファイルを利用者に提示する．

(4) 利用者は，提示された音声ファイルから，言語として認識可能な音声 (*Ham*) を全て選択 (解答) する．

(5) 正答数 k を求め，$k \geq$ 閾値 θ ならば利用者を受理，そうでなければ拒否する．

図 1 と図 2 に，提案方式 1 と提案方式 2 に従い $(h, s) = (2, 3)$ の条件で作問した音声の波形の例を示す．これらの図の縦軸は振幅であり，横軸は時間である．図 1 は日本語話者の音声で，図 2 はスペイン語話者の音声である．

5. 評価

5.1 評価項目

提案方式の有効性を検証するため，次の実験を行う．

表 1 実験に用いた音声の方式ごとの特徴
Table 1 Features on Sounds Used in Experiments.

Scheme	Speed and Pitch	Speaker
MGK [7]	Fixed	Native (Japanese)
Our Proposal 1	Modified	Native (Japanese)
Our Proposal 2	Fixed	Non-native (Spanish)

実験 1 HTK による単語やランダムな音韻列の認識率の評価

実験 2 人間による単語やランダムな音韻列の認識率の評価

実験 3 人間に対するユーザビリティの評価

5.2 実験方法

5.2.1 共通設定

Ham として用いる日本語単語は，「日本語教育のための基本語彙」 [21] から 157 個の自立語を抽出した．この 157 という数は，以下の条件を満たすようにして，単語を選択した結果に由来する．

- それぞれの単語のローマ字表記における編集距離 (Levenshtein distance) が 4 より大きい
- 50 音すべての文字が，少なくとも 1 つの単語の先頭文字として出現する
- 外来語は選択しない
- *MGK* 方式 [7] で基本単語として抽出した 127 個程度の数

Spam として用いるランダムな文字列は，*Ham* となる日本語単語の仮名文字をコーパスとする階数 1 のマルコフ連鎖を用いて 100 個を生成した．*Spam* に属する文字列についても，ローマ字表記における編集距離が 4 より大きくなるように生成した．

文字列の音声化には，音声合成を利用した．日本語話者を使用する場合は Open JTalk [*3] を，外国語話者を使用する場合は eSpeak [*4] を使用した．今回の実験では，外国語話者としてスペイン語話者を用いた．また，音声速度やピッチを変動させる場合は，1.0 を基準値としてそれぞれ 0.75 – 0.95，−300 – +300 の範囲で無作為に操作した．実験では，表 1 に示された特徴を持つ音声の *Ham*, *Spam* コーパスを使用した．

CAPTCHA として提示する音声ファイルは, *Ham* と *Spam* からそれぞれ $N_{Ham}(> 0)$ 個と $N_{Spam} = 5 - N_{Ham}$ 個の音声を無作為に取り出し，それらを 1.0 – −1.5 秒の無音区間で連結して作成した．利用者は，5 つのそれぞれの音声ついて，*Ham* か *Spam* かを解答する．すなわち，利用者に提示する各音声ファイルは，5 つの識別問題によって構成される．

[*3] http://open-jtalk.sourceforge.net/
[*4] http://espeak.sourceforge.net/

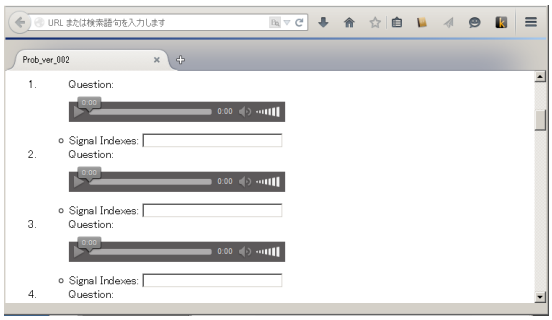


図 3 提案方式の出題画面の例
Fig. 3 Appearance of our CAPTCHA

5.2.2 実験 1

Ham と *Spam* の音声コーパスを用いて，方式ごとに 100 個の音声ファイルを生成した．これらは，HTK に入力するテストデータセットとなる．

HTK に訓練データセットとして入力する音声は，*Ham* として用いる日本語単語を，日本語話者の標準速度とピッチで読み上げたデータを使用した．提案方式 1 と 2 に対しては，それぞれの方式と同様の方法で生成した音声も，訓練データセットに加えた．HMM の訓練方法は，HTK チュートリアルにある標準設定を使用した．なお，ランダムな音韻列については，実際には無数に生成できるため，訓練データセットとして使用しない．

HTK の音声認識は，全ての単語が等確率で出現する設定で行った．*Spam* の認識については，*Ham* 以外の音声を単語と認識した場合に認識失敗として扱った．これは，HTK では挿入エラーとして検出される．*Ham* の認識については，音声ファイルに含まれる単語数を HTK に知識として与え，*Ham* 音声が単語の並び順を含めて正しく認識できたかどうかで評価した．

5.2.3 実験 2

次の手順で生成した問題を用いて，実験を実施した．問題は，html 形式 (図 3) で提示した．

(1) 各提案方式に対して，$N_{Ham} = 1, 2, 3$ ごとに 4 個の音声ファイルを生成する．
(2) 音声ファイルを被験者に問題として提示する．問題の順番は，提案方式ごとに無作為に並び替える．被験者には，問題ごとに少なくとも 1 つの *Ham* 音声があること，音声の再生回数や解答時間に関する制限は無いことを通知する．
(3) 被験者は，問題ごとに 5 つの音声から *Ham* を解答する．解答は，各音声に付けられた番号を用いて行う．
(4) 被験者の解答を採点する．各音声について *Ham* か *Spam* かの認識結果を集計する．

実験は 2 回に分けて実施した．提案方式 1 に関しては，大学生の男女 8 人に参加して頂いた．提案方式 2 に関しては，大学生の男女 4 人と，30 代 3 人，60 代 1 人の社会人の男女に参加して頂いた．参加頂いた被験者は，全員が晴眼者であった．

表 2 HTK による単語やランダムな音韻列の条件付き検出率 [%]
Table 2 Conditional Probabilities of Words and Random Phoneme Sequences to be Detected by HTK.

Scheme	N_{Ham}	$P(W = t\|X = H)$	$P(W = t\|X = S)$
MGK [7]	1	73.00	0.00
	2	55.50	0.00
	3	52.67	0.00
Our Proposal 1	1	27.00	0.00
	2	24.50	0.00
	3	25.33	0.00
Our Proposal 2	1	34.00	0.00
	2	31.00	0.00
	3	25.67	0.00

5.2.4 実験 3

本稿では，人間の要した解答時間を調査し，提案方式のユーザビリティの指標とする．また，被験者からの主観的な意見も参考にする．解答時間のデータや主観的な意見は，実験 2 を実施した際に取得した．

5.3 実験結果

5.3.1 実験 1

表 2 に，HTK による単語やランダムな音韻列の認識結果を示す．

全ての方式において，HTK は *Spam* を *Ham* と誤検出した．この理由は，ランダムな音韻列が単語と同じ音韻で構成されるためである．参考文献 [7] の結果と同様に，HTK はランダムな音韻列を非言語雑音として無視できず，何らかの単語として認識した．

Ham の検出については，発話の多様性の違いが影響していると考えられる．

MGK 方式は，*Ham* 音声自体には特別な処理を施さないので，認識精度の高い特定話者向けの音響モデルを利用できる．そのため，HTK は *Ham* を高い確率で検出した．

提案方式 1 では，単語ごとに複数種類の速度とピッチを持つ音声を無数に作成できる．音響モデルは，各単語ラベルと特徴量の対応に幅ができる不特定話者モデルとなる．

提案方式 2 では，母国語を異にする話者のもつ特徴量を，音響モデルとして学習する必要がある．ただし，提案方式 1 と比べると，単語ラベルに対する特徴量の多様性は小さい．そのため，HTK にとっては，提案方式 1 の方がより認識しづらい結果になったと考えられる．

N_{Ham} が大きいほど *Ham* の検出率が低下する傾向については，コーパスに使用した単語数 157 と，作問された *Ham* 数に理由があると推測している．$N_{Ham} = 1, 2, 3$ に対し，評価した *Ham* 数は，それぞれ 100, 200, 300 となる．$N_{Ham} = 1$ ではすべての単語が使用されたわけではないので，その際 HTK の認識率の低い単語が除外されている可能性がある．

表 3 被験者らによる単語やランダムな音韻列の条件付き認識率 [%]
Table 3 Conditional Probabilities of Words and Random Phoneme Sequences to be Detected by Participants.

Scheme	N_{Ham}	$P(Y = H\|X = H)$	$P(Y = H\|X = S)$
Our Proposal 1	1	87.58	7.03
	2	82.81	7.29
	3	88.54	6.25
Our Proposal 2	1	96.88	4.69
	2	64.06	8.33
	3	56.25	6.25

5.3.2 実験 2

表 3 に，人間による単語とランダムな音韻列の認識結果を示す．

人間による *Spam* の認識率は，提案方式や N_{Ham} に依存しない高い値である．人間は聞き取った音声に対し，意味論的な解釈ができるため，機械に比べると *Spam* の認識が容易なのだと考えらる．また，作問された音声は，速度・ピッチ・話者などが変動するため，日常的な会話に比べは聞き取りづらい．よって，*Spam* よりな認識傾向があることも推測される．

提案方式 1 の *Ham* の認識については，N_{Ham} に依存しない高い値を示している．この結果からは，人間は音声速度やピッチの変動に対して，ある程度頑強な認識ができることがわかる．

提案方式 2 の *Ham* の認識については，$N_{Ham} = 1$ とそれ以外で大きな差がある．この理由として，被験者に与えた $N_{Ham} > 0$ の知識の影響があるという仮説を検討している．被験者は，問題に含まれる音声のすべてを *Spam* だと認識した場合でも $N_{Ham} > 0$ の知識によって，もっとも聞き取りやすい音声を *Ham* と解答できる．$N_{Ham} > 1$ の場合は，同様の判断がなされた場合，他の *Ham* の認識は失敗してしまう．

提案方式 2 における $N_{Ham} > 1$ での *Ham* 認識率が低い他の要因としては，日本人と異なる発話者の影響も考えられる．人間にとっては，普段耳慣れない発話は，慣れ親しんだ発話に比べて認識しにくいと推測できる．

$N_{Ham} > 0$ の知識の影響に関する仮説が正しければ，$N_{Ham} > 1$ の実験結果は，$N_{Ham} = 1$ の結果より，正確な *Ham* の認識率を示していると考えられる．そうであれば，提案方式 1 で生成した音声は提案方式 2 のものに比べて，人間が認識やすいといえる．この推論の確認は，$N_{Ham} > 0$ の知識を与えない場合の実験が必要である．

5.3.3 実験 3

被験者に提示した音声ファイルごとの平均解答時間を，表 4 に示す．表 4 の結果は，各方式ごとに最大・最小の処理時間を要した被験者のデータは含まない．解答時間には，音声の再生や解答記入に要した時間を含む．提案方式によって音声の再生時間の平均が異なる理由は，作問に使用された単語やランダムな音韻列の違いや，単語間に挟んだ無音期間の違いのためである．

表 4 被験者らの 1 音声ファイル †1 あたりの解答時間 [秒]
Table 4 Response Time [sec.] for each audio file†1.

Our Scheme	Response Time	Variance	Duration for a Question
Proposal 1	23.7	18.1	8.8
Proposal 2	25.3	12.2	11.2

†1: Each audio file consists of 5 questions.

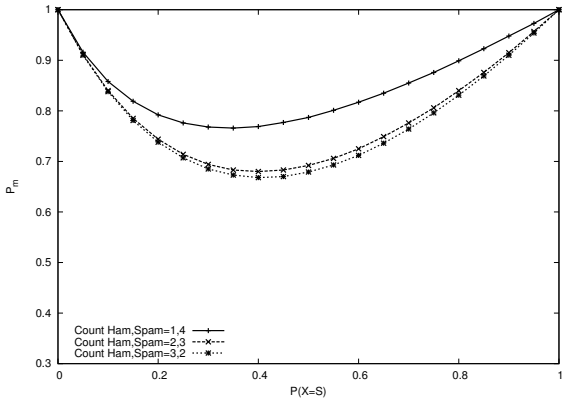


図 4 *MGK* 方式で生成された問題 1 問当たりの機械の攻撃成功率 (P_m)
Fig. 4 Machines' Success Rate for each Question (P_m) regarding *MGK*-scheme.

解答時間と音声の再生時間から，被験者らは，1 音声ファイルあたり 2 回程度の再生を繰り返していると推測できる．音声を繰り返し再生している理由は，提案方式の音声が聞き取りづらいためなので，人間が快適に使用するためには課題があることがわかる．なお，1 つの音声ファイルには 5 つの *Ham* または *Spam* が含まれるため，識別問題 1 問あたりの解答時間は，いずれの提案方式でも約 5 秒であった．

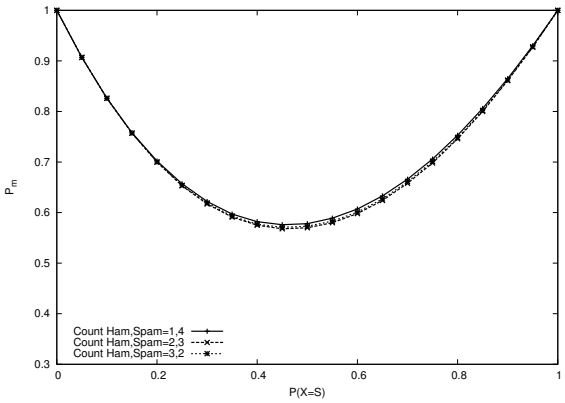
主観的な意見としては，提案方式と *MGK* 方式の双方で，聞き取りづらいとの回答が多かった．*MGK* 方式で指摘のあった記憶作業が負担になるとの意見は，提案方式に関しては出なかった．これは，提案方式が識別問題を採用したことで，単語の表記による解答を不要とした効果だと思われる．

6. 方式の比較

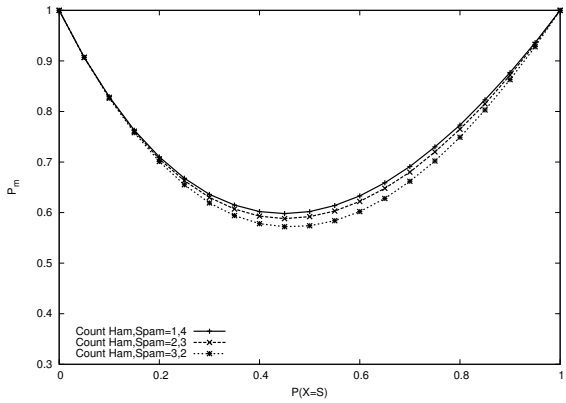
本節では，節 3 で示した安全性定義と評価指標を用いて，提案方式や既存方式の性能を比較する．

音声認識を利用した攻撃: 本稿では，HTK と同等の性能を持つ *Ham* と *Spam* の検出器 $\mathcal{A}$ の存在を仮定する．攻撃者は，$\mathcal{A}$ の出力結果を利用した推定を行うとする．

機械による攻撃の成功率 P_m の計算方法を，$s = 5, h = 15$ の場合を例に挙げて示す．検出器 $\mathcal{A}$ により *Ham* が認識される事象を $W = t$，認識されない事象を $W = f$ とする．$N_{Ham} = 1$ における提案方式 1 の場合を例にとれば，



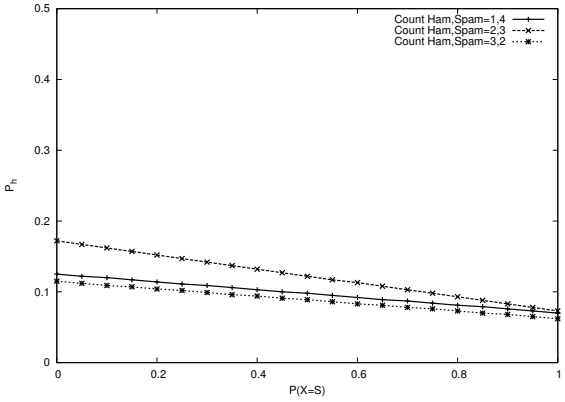
[1] Our Proposal 1.



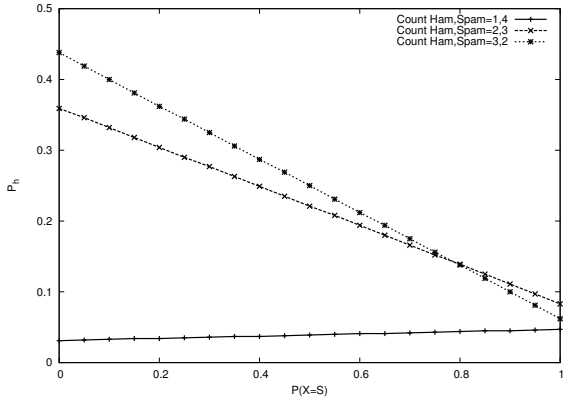
[2] Our Proposal 2.

図 5　提案方式 1，2 で生成された問題 1 問当たりの機械の攻撃成功率 (P_m)

Fig. 5　Machines' Success Rate for each Question (P_m) regarding our Proposal 1 and 2.



[1] Our Proposal 1.



[3] Our Proposal 2.

図 6　提案方式 1，2 で生成された問題 1 問当たりの人間の失敗率 (P_h)

Fig. 6　Humans' Failure Rate for each Question (P_h) regarding our Proposal 1 and 2.

表 2 の結果から $P(W = t|X = H) = 0.27$ となる．同様に，$P(W = t|X = S) = 0$ となる．$P(X = S) = 0.25$, $P(X = H) = 0.75$ なので，$P(W = t) = 0.2025$, $P(W = f) = 0.7975$ となる．この分類器では，$W = t$ であれば *Ham* と解答する．そうでなければ，$P(X = H|W = f) = 0.687$ の確率で *Ham* と，$P(X = S|W = f) = 0.313$ の確率で *Spam* と解答する．したがって，$P(Y_m = H, X = H) = 0.771$, $P(Y_m = S, X = S) = 0.313$ となり，式 (2) から $P_m = 0.657$ となる．

以上の方法で，各方式に対して s, h の組み合わせを考える．図 4 と図 5 に，*Ham* と *Spam* の識別問題 1 問中の *Spam* 含有率に対する P_m の変動を示す．

提案方式は，*MGK* 方式に比べ P_m が低いので，安全性が高いことが分かる．提案方式 1 と 2 では，提案方式 1 の方が若干ではあるが，低い P_m を示す結果となった．

人間による音声認識能力: 表 3 の結果から，式 (1) を用いて P_h を計算する．図 6 に，問題中の *Spam* 含有率に対する P_h の変動を示す．

表 3 が示す通り，人間は *Spam* の認識を得意とするため，*Spam* を多く含む場合に F_h が低下する．

表 5　既存方式と提案方式の比較

Table 5　Comparison between Conventional Schemes and our Proposal.

Scheme	N_{Ham}	FRR	FAR	F-ratio
reCapcha†1	–	0.53	0.586	0.440
MGK [7]†2	1	0.078	0.766	0.373
	2	0.078	0.680	0.475
	3	0.078	0.668	0.488
Our Proposal 1	1	0.098	0.576	0.577
	2	0.122	0.568	0.579
	3	0.089	0.571	0.584
Our Proposal 2	1	0.039	0.598	0.567
	2	0.221	0.588	0.539
	3	0.250	0.572	0.545

†1: The values of *FRR* and *FAR* are referred from [3] and [6], respectively.

†2: The value of *FRR* is referred from [7]. Tte paper shows just the results of $P(Y = H|X = S)$. Hence, we suppose $P(Y = H|X = S) = 0$ that is the best case for *MGK*-scheme.

提案方式と既存方式の比較: 方式ごとに，P_m が最小となる (s, t) 条件を選択し，その *FRR* と *FAR* から式 (3) に従い F 値を計算する．F 値を用いた比較結果を，表 5 に示す．

表 5 より，提案方式 1 が最もよい性能を示すことが分かる．

7. おわりに

本稿では，ランダムな音韻列と単語の識別問題を利用した音声型 CAPTCHA を提案した．識別問題の利用により，既存の音声型 CAPTCHA の問題である人間の記憶作業への負担や，軽微な聞き間違いによる正答率低下を防止した．また，音声を多様な話者の発話を模擬して生成することで，音響モデルの生成を困難にし，機械による音声認識を困難にした．本稿では，実験を通して提案方式と既存方式の比較を行い，発話の多様性を音声速度とピッチで模擬する方式が，最も高い性能を持つことを示した．

参考文献

[1] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Captcha: Using hard ai problems for security. In *Proceedings of EUROCRYPT*, volume 2656, pages 294–311. Springer-Verlag, 2003.

[2] 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. ACM, 2009.

[3] Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell, and Dan Jurafsky. How good are humans at solving captchas? a large scale evaluation. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, pages 399–413. IEEE Computer Society, 2010.

[4] Elie Bursztein, Romain Beauxis, Hristo S. Paskov, Daniele Perito, Celine Fabry, and John C. Mitchell. The failure of noise-based non-continuous audio captchas. In *32nd IEEE Symposium on Security and Privacy, S&P 2011, 22-25 May 2011, Berkeley, California, USA*, pages 19–31, 2011.

[5] Hendrik Meutzner, Viet-Hung Nguyen, Thorsten Holz, and Dorothea Kolossa. Using automatic speech recognition for attacking acoustic captchas: the trade-off between usability and security. In *Proceedings of the 30th Annual Computer Security Applications Conference*, pages 276–285. ACM, 2014.

[6] Shotaro Sano, Takuma Otsuka, Katsutoshi Itoyama, and Hiroshi Okuno. Hmm-based attacks on google's recaptcha with continuous visual and audio symbols (preprint). *IPSJ Journal*, 56(11), nov 2015.

[7] Hendrik Meutzner, Santosh Gupta, and Dorothea Kolossa. 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*, CHI '15, pages 2335–2338. ACM, 2015.

[8] 篠崎 隆宏 and 古井 貞熙. 話し言葉音声認識における話者間の認識率変動要因の解析. Technical Report 123(2001-SLP-039), 東京工業大学大学院情報理工学研究科, 東京工業大学大学院情報理工学研究科, dec 2001.

[9] 平山 直樹, 吉野 幸一郎, 糸山 克寿, 森 信介, and 奥乃 博. 擬似生成した複数方言言語モデル混合による混合方言音声認識. **情報処理学会論文誌**, 55(7):1681–1694, jul 2014.

[10] Jonathan Holman, Jonathan Lazar, Jinjuan Heidi Feng, 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. ACM, 2007.

[11] Sajad Shirali-Shahreza, Gerald Penn, Ravin Balakrishnan, and Yashar Ganjali. Seesay and hearsay captcha for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2147–2156. ACM, 2013.

[12] Pablo Ximenes, Andre Santos, Marcial Fernandez, and Jr. Celestino, Joaquim. A captcha in the text domain. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, volume 4277, pages 605–615. Springer-Verlag, 2006.

[13] 福岡 千尋, 西本 卓也, and 渡辺 隆行. 音韻修復効果を用いた音声 captcha の検討 (高齢者の認知機能保障技術及び一般). **電子情報通信学会技術研究報告**. *WIT*, **福祉情報工学**, 108(332):83–88, nov 2008.

[14] J. C. R. Licklider George A. Miller. The intelligibility of interrupted speech.

[15] Takumi Yamamoto, J.D. Tygar, and Masakatsu Nishigaki. Captcha using strangeness in machine translation. *2013 IEEE 27th International Conference on Advanced Information Networking and Applications*, 0:430–437, 2010.

[16] 鴨志田 芳典 and 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**, 54(9):2156–2166, 9 2013.

[17] 山口 通智. 人間ロボット判別テストのバリアフリー化のためのネット上文章の採取加工技法. **ヒューマンインタフェース学会論文誌**, 15(4):337–352, 2013.

[18] 山口 通智, 岡本 健, and 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**, 56(9):1834–1845, sep 2015.

[19] BBC News. Blind federation criticises captcha security test: `http://www.bbc.com/news/technology-22754006`, 2013.

[20] Miranda Pond Darlene A. Brodeur. The development of selective attention in children with attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology*, 29:229–239, June 2001.

[21] 国立国語研究所. **日本語教育のための基本語彙調査**. 秀英出版, 1984.

[22] S.J. Young and Sj Young. The htk hidden markov model toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44, 1994.

[23] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International Journal on Emerging Technologies*, 2009.

付　　録

A.1 音声認識処理の概要

音声認識の分野で広く利用されている HTK (Hidden Markov Model Toolkit [22]) を例に挙げ，音声認識処理の概略を示す．

教師ありの音響モデル (HMM) の作成　訓練データセットとする音声ファイルから，MFCC [23] などの特徴量による特徴ベクタを抽出し，単語単位で対応するラベルを付ける．音響モデルは，単語の構成要素となる音韻を状態に割り当てた HMMs として，訓練を行う．

音響モデルを用いた音声認識　認識処理では，単語リスト，単語の出現ルール (文法)，音響モデル，そしてテストデータセットの特徴ベクタを入力する．HTK は，入力に対する HMMs の状態遷移から音韻列を作成し，単語リストから，音韻列に対応する最も確からしい単語列を出力する．音韻を含まない雑音については，HMMs が活性化されないので，HTK はこれを無視する．