明治大学大学院先端数理科学研究科

2016年度

博士学位請求論文

視覚障害者向けの CAPTCHA に関する研究

A Study on CAPTCHAs Accessible to People with Visual Impairment

学位請求者　現象数理学専攻

山口　通智

あらまし

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) は，人間とロボット (ソフトウェアによる自動化エージェント) を判別し，ロボットによるオンラインサービスの不正利用を防止する技術である．

現在最も普及している視覚型 CAPTCHA は，利用者に歪曲や雑音を付加した画像の解釈をさせる方式だが，視覚障害者にその利用は困難である．代替案として，利用者に難聴化した音声の聞き取りをさせる方式があるが，このような聴覚型 CAPTCHA は人間にも解けない程に難しい．利用者に文章の文意や文脈を解釈させる言語型 CAPTCHA も提案されているが，その安全性については検討すべき余地がある．

既存の言語型 CAPTCHA は，人間の記述した自然文をテストの一部として提示するため，その収集方法に課題がある．インターネット上の文章は豊富な分量を持つが，検索エンジンによりコーパスが特定され，その情報を攻撃者に利用されてしまう．分量の少ない秘匿された文章で作問した場合は，同じ自然文が使いまわされる可能性が高く，その出題履歴を狙った攻撃に対して脆弱となる．

この課題を解決するため本論文では，視覚障害者にも利用可能であり，かつロボットの攻撃に対しても頑強な CAPTCHA として，新しい言語型と聴覚型の 2 方式を提案する．

言語型の第 1 の方式として，異なる「文の自然さ」をもつ機械合成文の相対比較問題を提案する．提案方式では，自然文をテストに含まないため，自然文の利用に起因する脆弱性を克服できる．また，相対比較問題により，人間の認知バイアスによる正答率の低下を抑制する．

第 2 の方式としては，文字置換による自然文の秘匿を検討する．提案方式では，文の一部を改変することで，ロボットによる構文解析や検索を妨害する．また提案方式は，方言などにみられる子音交替の規則に準じた文字置換により，人間に認識しやすい改変を行う．

聴覚型 CAPTCHA については，強い雑音の挿入に起因する人間の正答率の低下が課題となる．既存方式の多くが白色雑音などの統計的雑音を使用するが，ASR (Automatic Speech Recognition) により雑音として除去されやすい．ASR に対抗するために強い雑音を用いた方式は，人間にも解答が困難になる．そこで本論文では，多様な話者に対する音声認識の難しさに着目する．提案方式では，非母国語話者による発話や発話速度やピッチの変動によって，多様な話者を模擬した音声を合成する．

本論文では，実験によって既存方式と提案方式の安全性やユーザビリティを評価し，提案方式の有用性を示す．

目 次

図 目 次

表 目 次

略語の一覧

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CMU	Carnegie Mellon University
DCG	Dynamic Cognitive Games
FAR	False machine Acceptance Rate
FRR	False human Rejection Rate
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
MFCC	Mel-Frequency Cepstrum Coefficients
MIT	Massachusetts Institute of Technology
NFB	National Federation of the Blind
NLP	Natural Language Processing
OCR	Optical Character Recognition
PLP	Perceptual Linear Predictive
RMS	Root Mean Square
RPS	Random Phoneme Sequence
SNR	Signal to Noise Rate
SVM	Support Vector Machine
VoIP	Voice over IP
WAI	Web Accessibility Initiative
WCAG	Web Content Accessibility Guidelines
WHO	World Health Organization
W3C	World Wide Web Consortium

第1章　序論

1.1　本研究の背景

1.1.1　オンラインサービスの普及と不正なソフトウエアの脅威

インターネットの普及により，我々はオンライン上で，多数のサービスを享受できるようになった．オンラインショッピングを始めとするさまざまなオンラインサービスは，サービス提供者と利用者の双方が時間や場所の制約を受けずに利用できるため，爆発的な普及が進んでいる．

その一方で，サービスの提供者と利用者が対面するオフライン方式では想定する必要がなかった，オンライン方式特有の不正が問題になっている．すなわち,「サービスの利用者が本当に人間なのか？」である．

オンライン上の不正として有名な事例に，1999 年の *slash.com* 上での多重投票問題がある [1]．*Slash.com* は,「コンピュータサイエンスの分野で最も優れた大学はどこか？」というオンライン投票を実施した．投票システムは，不正利用者の多重投票を防ぐため，投票者の IP アドレスを記録していた．それにもかかわらず，CMU (Carnegie Mellon University) と MIT (Massachusetts Institute of Technology) に所属する学生の作成した「自動されたソフトウエアエージェント」(以降，ロボットと称す) は，IP アドレスの履歴によるファイアウォールを突破し，不正な多重投票を行った．この例以外にもオンライン上の不正は数多く存在し，2013 年の調査 [2] によれば，ウェブトラフィックの 61% は「不正なロボット」(以降，ボットと称す) により生成されている．

1.1.2　CAPTCHA とその機能

ボットによるオンラインサービス上の不正が問題になると,「サービス利用者が人間であることを，どのようにして保証するか?」という課題が研究者らによって議論されるようになった．Ahn らは，ボットの活動を妨害する試みとして CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) のアイディアを 2000 年に提案し，その後文献 [3] にまとめた．

CAPTCHA の概念図を図 1.1 に示す．CAPTCHA は，AI 問題を用いて人間とロボットを識別する自動化されたチューリングテストであり，その目的はボットによる不正の妨害である．AI 問題とは， Ahn らの定義によれば「人間には容易に解けるが，ロボットには

解答困難な問題[1]」である．AI 問題には，難読化された画像の認識問題や難聴化された音声の認識問題などがあり，それらを利用した CAPTCHA は，それぞれ視覚型と聴覚型に分類される．図 1.2 に，視覚型 CAPTCHA の例を示す．

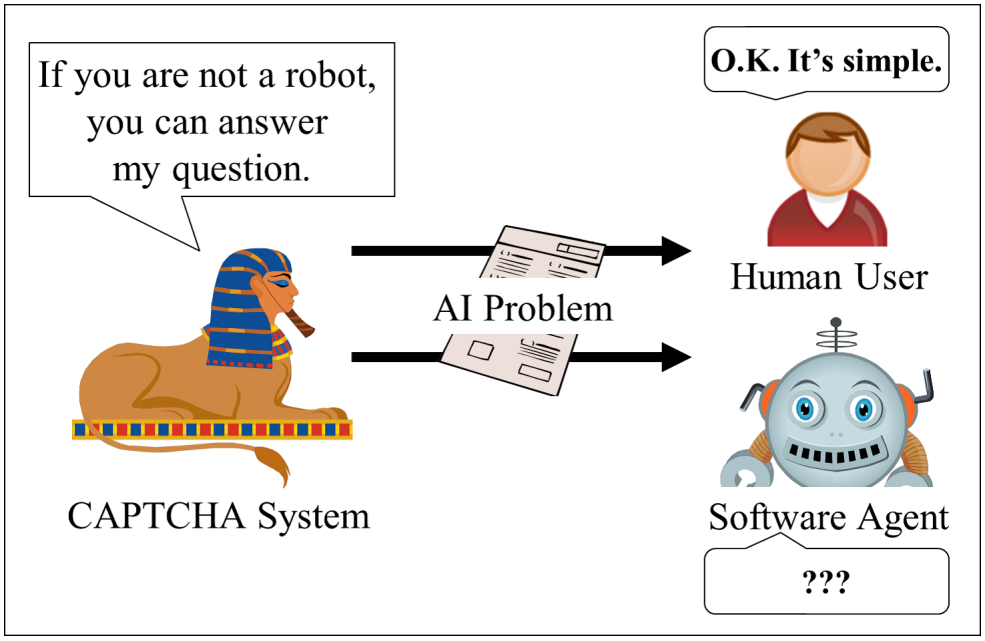


図 1.1: CAPTCHA の概念図
Fig. 1.1: Conceptual Diagram of CAPTCHA.

図 1.2: CAPTCHA の例
Fig. 1.2: Example of CAPTCHA.

Ahn らによる提案後，CAPTCHA はウェブセキュリティにおける基盤技術の 1 つとして急速に普及している．CAPTCHA の導入が効果的な事例を，次に示す [1].

ブログなどにおけるコメントスパムの防止 コメントスパムとは，ボットによるコメント

[1]CAPTCHA の目的は，ボットを駆除することにある．よって，ロボットには解けるが人間には解答困難な問題は，AI 問題にはならない．

欄の荒らしや広告挿入行為，サーチエンジンのランキング向上を狙った偽のコメント投稿を指す．CAPTCHA により，ボットによる不正なコメントの投稿を抑制できる．

オンライン登録の保護 代表例は，フリーメールサービスのアカウント作成の保護である，CAPTCHA の導入以前のフリーメールサービスは，ボットにより毎分数千ものアカウントが不正に取得され，スパム業者に利用されていた．CAPTCHA により，ボットによるサービスの不正登録や利用を抑制できる。

スパム業者からの e-mail アドレスの保護 スパム業者は，ウェブをクローリングすることで e-mail アドレスを収集する．CAPTCHA のようにロボットに解釈が困難な形式で e-mail アドレスを提示することで，クローラによる e-mail アドレスの収集を避ける事ができる．

オンライン投票の保護 ボットによる投票が可能ならば，多重投票による不正な意見操作の恐れがある．CAPTCHA により，人間によってなされた投票結果であると信頼性を与えることができる．

パスワードの保護 パスワードの探索には，総当たり攻撃や単語の組み合わせを試す辞書攻撃などがある．これらの攻撃が成立するには，制限なしに何度もパスワードの入力を試行できなければならない．CAPTCHA により，パスワードの入力を人間のみに制限することで，総当たり攻撃や辞書攻撃を抑制できる．

検索エンジンボットによるインデックス化の防止 ウェブコンテンツの作成者は，検索エンジンによる作成したページのインデックス化を HTML タグの指定で抑制できる．しかし，そのタグ指定をどう処理するかは検索エンジン側にゆだねられている．ページへのアクセスに CAPTCHA を課すことで，検索エンジンによるページのインデックス化を防止できる．

特に，Yahoo! や Google などのフリーメールサービス業者は，独自の CAPTCHA の開発に乗り出すなど，積極的に CAPTCHA の導入を進めている．近年の研究では，CAPTCHA を暗号理論における構成要素として利用する例 [4] や，他のセキュリティ技術と組み合わせることでボットの活動を効率的に阻害する例 [5] もある．また，商用サービスの利用例として，オンラインゲーム上でのボットの不正行為への対策がある [6].

1.1.3 既存の CAPTCHA の問題点

現在利用されている CAPTCHA には，安全性，アクセシビリティ，ユーザビリティに関する問題が指摘されている．

安全性

近年では，計算機の性能向上や機械学習アルゴリズムの改良により，ロボットによるCAPTCHA への攻撃成功例が多数報告されている．CAPTCHA は継続的に攻撃にさらされているが，2011 年の Bursztein らによる報告が特に有名である [7, 8]．Bursztein らは，当時の商用サイトで利用されていた視覚型と聴覚型の大部分の CAPTCHA に対して，機械学習を用いたロボットによる攻撃に成功した．また，最近の研究 [9, 10, 11] では，Bursztein らの報告では比較的頑強であった Google CAPTCHA に対しても，高い確率で攻撃に成功したとの報告がされている．

CAPTCHA の開発者は，CAPTCHA へ新たな攻撃方法が提案されると，それに対して 1) より高度な人間の認知能力の利用と，2) より強い難読化／難聴化で対抗してきた．しかしながら，これらの対策は，アクセシビリティやユーザビリティを犠牲にしたものが多い．

アクセシビリティ

アクセシビリティとは,「障害のある人でも情報にアクセスできるか?」という指標である．アクセシビリティに関する研究は継続的に行われている [12, 13, 14, 15].

インターネットは障害者の間でも広く普及している [16, 17] ため，ウェブサイトのアクセシビリティは重要な課題である．しかしながら，CAPTCHA は，ウェブアクセシビリティを阻害する可能性があるとして，W3C (World Wide Web Consortium) ガイドラインなどで，その方式や利用方法に関しての注意が促されている [18, 19].

CAPTCHA のアクセシビリティの問題は，現在利用されている方式の大部分が，視覚型に属するためである．初期の CAPTCHA は，図 1.2 のような視覚型である．また，より高度な人間の認知能力の利用する CAPTCHA の大部分は，画像を利用する視覚型である [20, 21, 22, 23]．視覚障害者は，画像の判別が困難であるため，視覚型 CAPTCHA を備えたウェブページを利用できない．

代替方式として聴覚型 CAPTCHA も提案されてはいるが，表 1.1 に示されるように，十分に普及しているとは言い難い．また，既存の聴覚型 CAPTCHA には，ユーザビリティにおける欠陥が指摘されている．

ユーザビリティ

ユーザビリティとは,「アクセスした先の情報が使いやすい形で提供されているか?」という指標である．

聴覚型 CAPTCHA は，人間にも解くことが難しいと指摘されている [24, 25, 26]．これらの研究では，人間の聴覚型 CAPTCHA の正答率は，画像型の半分程度しかないと報告されている．正答率が低いと CAPTCHA の認証を受けるために複数回の試行が必要になるため，ユーザビリティは低下する．

表 1.1: CAPTCHA のサービス提供状況 (2013 年 11 月)
Table 1.1: Service Appearance of CAPTCHAs in November 2013.

Service Name	Visual CAPTCHA	Audio CAPTCHA	Other Type of CAPTCHAs	Telephone Dialogue
Microsoft	x	x		
Google	x	x		x
Yahoo! Japan	x	x		
Amazon	x	x		x
WordPress	x			
Ameba Blog	x			
F2C Blog	x			
White House Petition			x†	

†: The type is quiz.

さらに聴覚型 CAPTCHA は，ロボットによる攻撃への対抗策としてより強い難聴化を適用した結果，もはや人間にもほとんど解けないほどに難化してしまった．2013 年には，NFB (National Federation of the Blind) やオンライン請求サイト *Change.org* において，聴覚型 CAPTCHA は事実上人間には解けない方式だと非難されている [27].

1.2 本研究の目的

WCAG 2.0 (Web Content Accessibility Guidelines 2.0) と WAI-ARIA (Web Accessibility Initiative - Accessible Rich Internet Applications) の策定 [18, 28] や，障害者基本法 [29] 第 4 条 2 項の社会的障壁除去への言及にも関わらず，現在普及している CAPTCHA の多くは，視覚障害者がウェブアクセスをする際の障壁になっている．本研究の目的は，視覚障害者にも利用できる CAPTCHA を提案し，このウェブアクセスにおける障壁を除去することである．具体的には，文章の文意や文脈解釈の難しさに基づく AI 問題 (文意文脈解釈問題と称する) を使用する言語型と聴覚型 CAPTCHA の提案を行う．

言語型と聴覚型の両方の CAPTCHA を提案する理由は，より多くの視覚障害者に対応するためである．言語型 CAPTCHA は，言語に関する習熟がある程度利用者に要求されるが，視覚と聴覚といった知覚に制限のないアクセシブルな方式である．聴覚型 CAPTCHA は，聴覚を介した問題の認識が必要になるが，幼年層や言語の学習障害を持つ視覚障害者にも使用できる方式である．

1.3 視覚障害者向け CAPTCHA 構成の困難性

1.3.1 言語型 CAPTCHA における困難性

言語型 CAPTCHA の先行研究には，鴨志田らにより提案された方式 [30] がある．本論文では，以後これを *KK* 方式と称す.

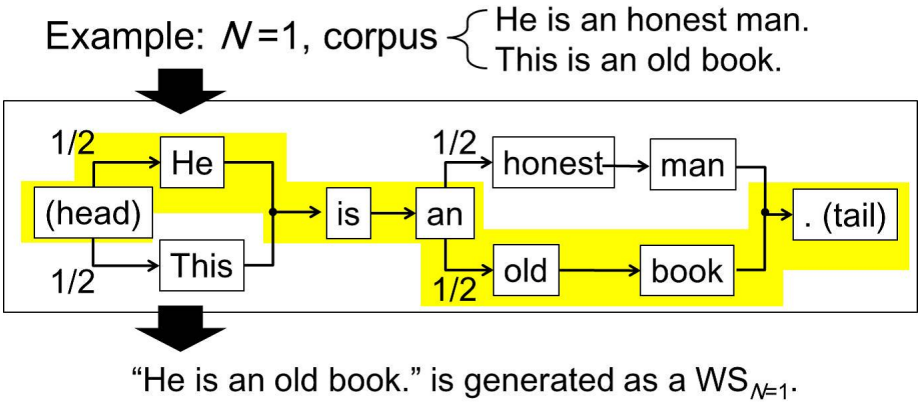


図 1.3: Markov Chain によるワードサラダの生成例
Fig. 1.3: Example of Generating Word Salads with Markov Chain.

KK 方式では，認証の際にワードサラダと呼ばれるマルコフ連鎖を用いて生成された合成文 (図 1.3) と人間の生成した自然文のいずれかを提示し，利用者にそれらを識別させる．ワードサラダと自然文の識別は，それらの間にある「違和感」や「自然さ」というはっきりと定義できない (“ill-defined”) 尺度を用いるため，ロボットに解くことは難しいと期待できる．さらに *KK* 方式は，テストをテキストのみで提示可能であり，利用者に要求される知識もその使用言語と一般常識のみであるため，視覚障害者の利用に適した興味深い試みのひとつである．

しかしながら，本論文では，*KK* 方式の問題新規性と識別性に関して，3 つの問題点を指摘する．

問題点 (L1): 自然文の収集困難性

鴨志田らが論文中で述べているように，ワードサラダは優れた多様性を持つので，自動作問に適している．しかし，人間の記述した多くの文章を必要とするため，多様な自然言語データベース (コーパス) を準備しなければならない．

小さなコーパスを用いた場合は，テストとして提示される自然文が使いまわされてしまう．これは，テストに用いた文章を収集し，使いまわされた文を自然文として解答する攻撃に脆弱である．

問題点 (L2): 検索エンジンを用いた攻撃への脆弱性

新規な問題文を生成するためには，分量の豊富な公開文章をコーパスにする方法が考えられる．しかしこの場合は，検索エンジンを用いた攻撃に対する安全性が問題となる．

自然文は，コーパスとして用いた文章の切り抜きであるため，図 1.4-(a) のように，一般の検索エンジンの検索結果に基づいて容易に識別できる．

問題点 (L3): プライミング効果による人間による正答率の低下

プライミング効果とは，先行する学習によって，後の学習が影響を受けることである．人間が行う事物の評価は，認知バイアスの影響を受けることが知られている．*KK* 方式のような単一文ごとに評価は，プライミング [31] ／アンカリング／確証／追認／保守性などの認知バイアスにより， 過去の解答結果の影響が付きまとう．同じ事物群を提示した場合でも，その順番によって解答結果が変化するのは，十分にありえる事象だと予測できる．図 1.4-(b) にその一例を示す．この場合，Q1 と Q2 は双方とも不自然な文章であるが，Q1 と比べれば人間は Q2 を自然に感じてしまう．

問題点 (L1)，(L2) は，人間の生成した文を自然文として作問に用いることに起因する．すなわち，これらの問題は *KK* 方式に限られたものではなく，[32, 33, 34] [2]などの， 自然文をテストの一部として提示する方式に共通した問題である．

言語型 CAPTCHA には，クイズ CAPTCHA [35, 36] もある．クイズ CAPTCHA は，ホワイトハウスのオンライン請願 [35] でも利用実績のあるアクセシビリティに優れた方式の 1 つである．本論文では，以後これを *WtP* 方式と称す．

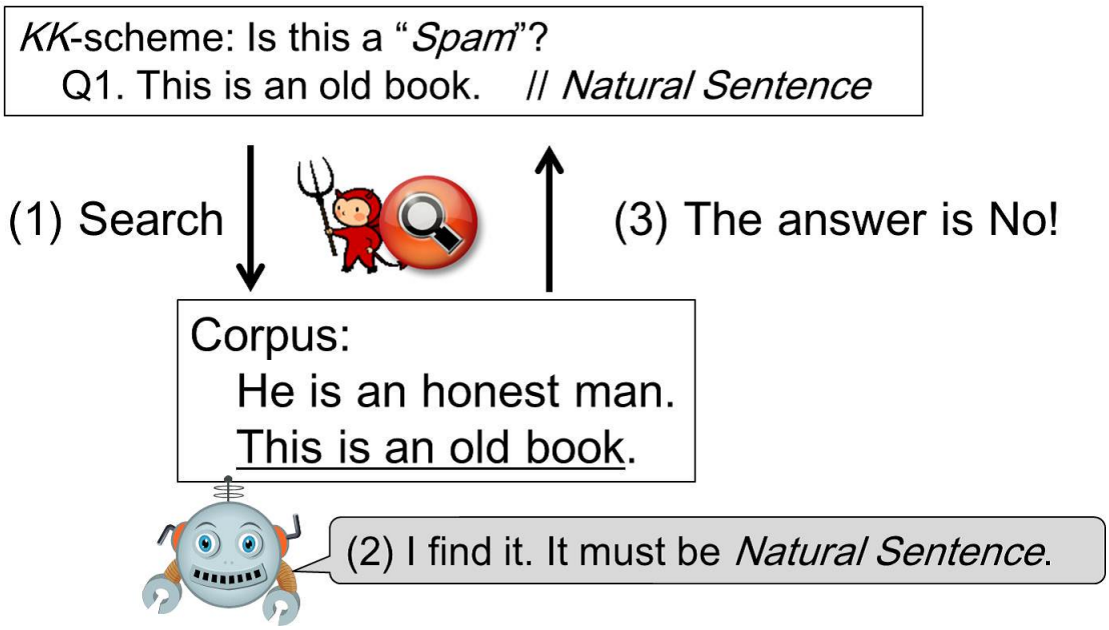
WtP 方式では，“Ant, yellow and red: the 2nd colour is?”，“Enter the biggest number of eight, twenty one or 46”，“Which of hair, finger, heart, knee or toe is part of the head?” などと，一般常識に関するクイズがテキスト形式で出題される．この方式は，数字や単語を変えることで，過去に出題されたことのない問題を容易に生成できるため，自然文を収集する必要がない．

しかしながら，本論文では，クイズ CAPTCHA の問題点として，出題文の「型」における多様性のなさを指摘する．

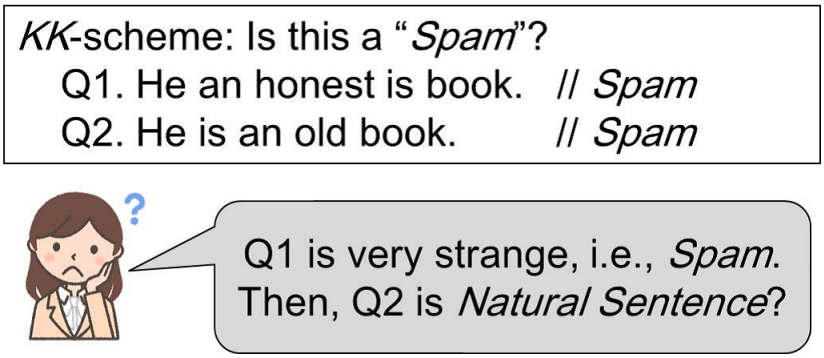
問題点 (L4): 問題文の型における多様性の不足

WtP 方式は，出題文の形式をある一定の型に固定し，その一部を置換することで作問をする．このように明確に定義された (“well-defined”) 型をもつ文章は，一般的な文章に比べると文意文脈解釈は容易であるため，ロボットでも出題文の内容を解読できると推測さ

[2] これらの詳細は，2.2.5 節を参照されたい．



(a) Attack with Search Engines



(b) Priming Effects

図 1.4: *KK* 方式の問題点
Fig. 1.4: Weak Points of *KK*-scheme.

れる．問題文の型における多様性が小さければ，単語辞書と簡単な構文解析を用いたソフトウエアに対してさえも脆弱になる．

1.3.2 聴覚型 CAPTCHA における困難性

まず，大部分の聴覚型 CAPTCHA が持つ問題点について示す．

問題点 (A1): 統計的雑音を用いた方式の脆弱性

聴覚型 CAPTCHA への攻撃に関する研究によれば，白色雑音などの統計的雑音は，非雑音との音の特徴量や “auditory power” の違いを利用した雑音除去技術を用いた ASR に対して脆弱である [8, 10]．挿入する雑音や音声の変形を強くして ASR に対抗する方法もあるが，結果として人間にも解答できないほどに難聴化されてしまう [24, 25, 26]．

問題点 (A2): 記憶作業負荷に起因するーザビリティの悪さ

人間が聴覚型 CAPTCHA に解答するには，認識した単語や数字を記憶する必要がある．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．一度に解答すべき単語や数字が増えるほど人間の記憶作業に対する負荷は高くなり，誤認による正答率の低下の原因となる．

問題文を聞きながら解答すればよいとの意見もあるが，視覚障害者の場合はそれも難しい．なぜなら，視覚障害者らが使用するスクリーンリーダはキーボードの入力結果を読み上げる機能を持つため，音声の聞き取り中に解答を入力をすると，その操作音が聞き取り作業を妨害してしまうためである．

問題点 (A1) に対する先行研究としは，統計的雑音の代わりに，意味論的雑音を用いる方式がある．意味論的雑音とは，背景雑音としての会話や歌などで，解答対象となる発話と同じ音韻 [3]から構成された音である．意味論的雑音の音としての特徴量は解答対象の発話と同等であるため，その識別には意味論的な解釈が必要になる．したがって，ASR が意味論的雑音を雑音と認識して除去することは難しい．

意味論的雑音を用いた聴覚型 CAPTCHA には，Meutzner らによる方式 [37] がある．本論文では，以後これを *MGK* 方式と称す．

MGK 方式では，意味論的雑音としてランダムな音韻列 (以後 RPS: Random Phoneme Sequence と称す) を用いて，単語の発話と RPS を重畳なしに結合した音声をテストとして提示する．利用者は，RPS を無視して単語の発話のみを解答できれば，認証を通過できる．

しかしながら，本論文では，*MGK* 方式の問題として，単語の発話部分に対する難聴化がなされていない点を指摘する．

[3]意味の弁別をなす最小の音声単位 (phoneme).

MGK 方式の問題点 (A3): 難聴化の施されていない音声区間の存在

MGK 方式では単語の音声区間に難聴化が適用されない．単語部分だけでも ASR が高い正答率を示すならば，それは攻撃者とって役立つ知識となるため，*MGK* 方式の弱点になりうる．残念ながら Meutzner らによる安全性の分析では，ASR による評価を単語と RPS で分けずに行っているため，この懸念を払しょくできていない．

1.4 本研究の着想

1.4.1 言語型 CAPTCHA に関する着想

本研究では，1.3.1 節で述べた言語型 CAPTCHA に関する困難性を解決するため，次の手法を検討する．

自然文が問題文として提示されない方式

1.3.1 節で示したように，*KK* 方式の問題点は自然文を利用する点に起因する．そこで本研究では，自然文がテストに含まれない方式の構築を狙う．

最初の着想は，*KK* 方式の自然文に相当する文にも，マルコフ連鎖により合成したワードサラダを用いることである．ワードサラダの利用は，自然文の利用に比べて，次の点で優れている．

- 同一コーパスに対して，多様な文を生成できる．
- 検索エンジンによる文字列の検索やコーパスの特定が困難である．

マルコフ連鎖は，そのパラメータである階数を変えることで，異なる性質をもつワードサラダを生成する．提案方式では，階数の異なるワードサラダの間に存在する「文としての違和感や自然さ」の差を人間に識別させる．階数によって多くの文を合成できるので，問題点 (L1) の自然文の問題点は存在しない．問題点 (L2) の検索エンジンを用いた攻撃に対しては，ワードサラダのみを使用するので，検索エンジンのデータベースと完全一致する確率は低いため，その攻撃に対して頑強になる．

次の着想は，子音交替による文章の改変である．子音交替とは，方言などにみられる単語の子音が変化する現象である．本研究では，問題文として提示する文に子音交替を施すことで，聞き間違い／書き間違い／多様な方言などを模擬する．

子音交替を適用後の文字列は，形態素解析 [38] を妨害するため，適用前に戻すことは難しい．したがって，自然文として問題文に提示される文も，元のコーパスにはない新しい文字列となる．

相対比較問題

一方で提案方式は，*KK* 方式の自然文とワードサラダの識別に比べ，人間には解きにくいと推測される．そのため，単一の文ごとに自然文やワードサラダのどちらかを評価させるのではなく，階数の異なる 2 つのワードサラダを 1 組にし，そのどちらが相対的に自然 (もしくは不自然) かを問う解答方式を採用する．この相対比較問題方式は，常に異なる性質を持つワードサラダを比較するため，問題点 (L3) のプライミング効果の抑制も期待できる．

1.4.2 聴覚型 CAPTCHA に関する着想

図 1.5: 多様な発話のイメージ
Fig. 1.5: Concept of Various Speakers.

1.3.2 節で示した問題点 (A1) は，統計的雑音の挿入による難聴化の限界を示している．そこで本研究では，既存の ASR が不特定話者に対する音声認識を苦手とする点に着目する．

音声認識は，話者の特徴をモデル化した音響モデルを使用するが，その作成は特定話者より不特定話者の方が困難である [39, 40]．本研究では，不特定話者の多様な発話を速度・ピッチの変動や非母国語話者を用いた合成音で模擬し (図 1.5)，発話から得られる音韻ごとの特徴量の揺らぎを大きくすることで，音響モデルの精度を劣化させる．一方で人間は，日常的に多様な発話を認識しなければならないため，この点ではロボットより優位であると期待できる．

さらに本研究では，Meutzner ら [37] の提案した RPS を用いて，言語的な意味を持つ単語と RPS の識別問題を用いる．この方式は，次の点で優れている．

- RPS は意味論的雑音であり，統計的雑音として除去することが困難である [37].
- 解答方式を単純化することで，問題点 (A2) で示した記憶作業の負荷を低減できる．
- 識別問題では，単語認識のような正確な解答は不要である．人間の認知能力は，多様な発話に対しても識別程度ならば頑強であり，ロボットとの能力に差があると期待できる．

1.5 本研究の貢献

本論文では，既存方式の分析と提案方式の実装や実験を通して，次のことを明らかにする．

まず，言語型 CAPTCHA に関して，次の点を明らかにする．

問題文の型におけるの多様性のなさを狙う攻撃に対するクイズ CAPTCHA の脆弱性 クイズ CAPTCHA は，ホワイトハウスが提供している請願サイト *We the People* [35] で使用されていた．本研究では，実際にこのクイズ CAPTCHA の攻撃を行い，その脆弱性を指摘する．

自然文を狙う攻撃に対する *KK* 方式の脆弱性 以下に示す攻撃が， 鴨志田らの検討した攻撃方式に比べて強力であることを示す．これらの攻撃により，*KK* 方式の安全性が危殆化することを示す．

- 自然文とワードサラダにおける生成文の多様性の差を用いた攻撃: 過去に出題された問題文と同一のものが提示された場合，それを自然文として解答する攻撃．
- 検索エンジンを用いた攻撃: 自然文とワードサラダを検索し，コーパスを検出した方を自然文として解答する攻撃．

相対比較方式で自然文として扱う合成文の生成に用いる最適な階数 提案方式の自然文に相当するワードサラダを，様々な階数のマルコフ連鎖から合成し，それぞれについて， 生成文の多様性や検索エンジンを用いた攻撃に対する安全性を検討する．また，人間による異なる階数で合成したワードサラダの識別能力を，実験により調査する．これらの結果をもとに，安全性とユーザビリティを併せ持つ，最適な階数を決定する．

***KK* 方式に対する提案方式の優位性** 提案方式と *KK* 方式の比較により，提案方式の優位性を明らかにする．

子音交替を用いた方式の評価 子音交替を施した文字列を用いる方式に関して，実験によりその安全性とユーザビリティを評価する．

次に，聴覚型 CAPTCHA に関して，次の点を明らかにする．

***MGK* 方式の脆弱性** *MGK* 方式に対して，ASR を用いた攻撃を行い，その脆弱生を示す．

ASR を用いた攻撃に頑強な話者の多様性の特徴 多様な話者を模擬するため，発話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，安全性やユーザビリティを調査し，より適した方式を示す．

***MGK* 方式に対する提案方式の優位性** 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．

1.6 本論文の構成

本論文は，7 章で構成される．

第 1 章では，まず本研究の背景と目的を述べ，次に本研究の概要を示し，その着想と貢献を述べた．

第 2 章では，本論文を通して使用する基本定義を示し，関連研究の概要を述べる．

第 3 章では，言語型 CAPTCHA の 1 種であるクイズ CAPTCHA について，既存方式の脆弱性を明らかにする．既存方式の代表例として， *WtP* 方式の分析を行い，その脆弱性を狙った攻撃ソフトウエアを実装し，安全性を評価する．

第 4 章では，言語型 CAPTCHA の新方式を提案する．まず，既存方式の代表例として，*KK* 方式の分析を行い，その脆弱性について述べる．次に，提案する言語型 CAPTCHA の着想とアルゴリズムを示す．さらに実験によって，既存方式と提案方式の安全性とユーザビリティを評価し，提案方式の優位性を示す．

第 5 章では， 言語型 CAPTCHA の作問にオンライン上の文章を利用するための採取加工技法について検討する．まず，提案方式の着想とアルゴリズムを述べ，複数の言語型 CAPTCHA に対する適用例を示す．次に，実験により提案方式の安全性とユーザビリティを評価し，提案方式が有用な言語型 CAPTCHA の特徴を明らかにする．

第 6 章では，聴覚型 CAPTCHA の新方式を提案する．まず，既存方式の分析からその問題点を明らかにする．次に，提案する聴覚型 CAPTCHA の着想とアルゴリズムを示す．さらに実験によって，既存方式と提案方式の安全性とユーザビリティを評価し，提案方式の優位性を示す．

第 7 章では，本研究について結論づける．

第2章　基本定義と従来研究

2.1　基本定義

2.1.1　記法

本論文で用いる記法について示す．$A(\cdot)$ を確率的アルゴリズムとする．A に乱数 r を用いることで決定的アルゴリズムとして動作する場合を A_r と表す．集合 $\mathcal{S}$ から，確率分布 $\mathcal{D}$ に従って要素 s を取り出すことを $s \leftarrow^{\mathcal{D}} \mathcal{S}$ と表す．特に $D = \$$ の場合は，一様ランダムな分布から要素 s を選択することを表す．

a, b を変数，リテラル，数字のいずれかとし，c を変数とする．$c \leftarrow a \circ b$ は，a, b に対する演算 $\circ$ の結果を c に代入することを表す．$c \leftarrow a$ は，a を c に代入することを表す．$|A|$ は，A が集合や配列であればそのサイズを，A が文字列ならばその文字数を表す．$[a, b]$ は，$a \leq x$ かつ $x \leq b$ となる整数 x の範囲を表す．

2.1.2　マルコフ連鎖モデル

N をマルコフ連鎖の階数とする．N 階マルコフ連鎖は， 直前 N 個の状態に依存して次の状態が決定される確率過程である．状態を表す確率変数 $X_0, \ldots, X_n, X_{n+1}$ について，X_{n+1} の生起確率は，その直前の N 個のみの条件付き確率で，

$$P(X_{n+1} = x | X_n = x_n, \ldots, X_0 = x_0) = P(X_{n+1} = x | X_n = x_n, \ldots, X_{n-N+1} = x_{n-N+1})$$

と与えられる．

マルコフ連鎖に基づく文章の合成

N 階マルコフ連鎖による文章合成は，コーパスに形態素解析 [38] を適用して抽出した形態素[1] N-gram と，そこから連鎖する $N + 1$ 番目の形態素の頻度情報からなるマルコフ連鎖モデルを構築して行う．本論文では，階数に幅を持たせたマルコフ連鎖モデルも対象とし，$N_L \leq N \leq N_H$ となる整数 N を階数とする場合は $[N_L, N_H]$ と表す．

マルコフ連鎖モデルにより合成された文を，ワードサラダと呼ぶ．特に文章合成に使用した階数 N を強調する際は，N 階ワードサラダと称する．

[1] ある言語で意味を持つ最小の単位．

マルコフ連鎖に基づく音韻列の合成

本論文では，RPS の生成にもマルコフ連鎖モデルを使用する．

まず，作問に使用する単語辞書を決める．次に，単語の表記に対応する発話辞書[2]から，単語に使用されている音韻 N-gram と，そこから連鎖する $N+1$ 番目の音韻の頻度情報からなるマルコフ連鎖モデルを構築する．本論文では，階数 1 マルコフ連鎖モデルを使用して，RPS の合成を行う．

2.1.3 *Ham* と *Spam*

本論文では，*Ham* は意味を持つ事物 (meaningful object) を，*Spam* は意味を持たない事物 (meaningless object) を表す．

Ham や *Spam* は，取り扱う事物により詳細な定義は異なるため，注意を要する．事物の種類に応じた詳細な定義については，それぞれ 4 章と 6 章に示す．

2.1.4 安全性定義と評価方法

本論文では， 鴨志田ら [30] の安全性定義を利用する．

評価指標 (FRR, FAR, F-値)

X を出題文を表す確率変数，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* となるテストを出題して，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．h を *Ham* の出題数，s を *Spam* の出題数とし $z=h+s$ とすれば，*Ham* と *Spam* を出題する確率はそれぞれ，

$$\begin{aligned} P(X=H) &= \frac{h}{z} \\ P(X=S) &= \frac{s}{z} = 1-\frac{h}{z} \end{aligned}$$

となる．CAPTCHA の成功率は，これらの同時確率で，

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

[2]例として [41] などがある．

と与える．人間による CAPTCHA 1 問あたりの失敗率を，

$$P_q = P(Y = S, Y = H) + P(Y = H, X = S)$$

とする．

ここで，CAPTCHA z 問により構成された認証方式を考える．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．また，ロボットによる CAPTCHA 1 問あたりの成功率を P_m とし，その正答数が $k \geq \theta$ となる確率を，機械受入率 FAR (False machine Acceptance Rate) と定める．すなわち，FRR および FAR は，1 回あたりの確率 P_q および P_m となる事象が z 回中 k 回発生し，かつ閾値 θ と k が前述の関係を満たす確率である．よって FRR および FAR を，二項分布で

$$\begin{aligned} FRR &= \sum_{k=\theta}^{z} \binom{z}{k} P_q^k (1 - P_q)^{z-k} \\ FAR &= \sum_{k=\theta}^{z} \binom{z}{k} P_m^k (1 - P_m)^{z-k} \end{aligned} \tag{2.1}$$

と与える．

本論文では，KK 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

$$F = \frac{2 \cdot (1 - FAR) \cdot (1 - FRR)}{(1 - FAR) + (1 - FRR)} \tag{2.2}$$

を用いる．

ツールを用いた攻撃に対する安全性

ツールによる特定の処理を表す確率変数を W とする．処理の具体例としては，MS-WORD による文章校正 [30] や検索エンジンによるコーパスの検出がある．$Spam$ が検出される事象を $W = t$ とすれば，その確率 $P(W = t)$ は，

$$\begin{aligned} P(W = t) &= P(W = t, X = S) + P(W = t, X = H) \\ &= P(W = t|X = S)P(X = S) \\ &\quad + P(W = t|X = H)P(X = H) \end{aligned} \tag{2.3}$$

となる．このとき，入力が $Spam$ である確率は，ベイズの定理から，

$$P(X = S|W = t) = \frac{P(W = t|X = S)P(X = S)}{P(W = t)} \tag{2.4}$$

となる．同様に，処理が行われない事象を $W = f$ とすれば，その確率 $P(W = f)$ は，

$$\begin{aligned} P(W = f) &= P(W = f, X = S) + P(W = f, X = H) \\ &= P(W = f|X = S)P(X = S) \\ &\quad + P(W = f|X = H)P(X = H) \end{aligned} \tag{2.5}$$

となる．このとき，入力が *Ham* である確率は，

$$P(X = H|W = f) = \frac{P(W = f|X = H)P(X = H)}{P(W = f)} \tag{2.6}$$

となる．したがって，ロボットによる解答 Y_w を，$W = t$ のとき，

$$Y_w = \begin{cases} S \text{ w./p. } P(X = S|W = t) \\ H \text{ w./p. } P(X = H|W = t) \end{cases} \tag{2.7}$$

$W = f$ のとき，

$$Y_w = \begin{cases} S \text{ w./p. } P(X = S|W = f) \\ H \text{ w./p. } P(X = H|W = f) \end{cases} \tag{2.8}$$

と定めることで，*FAR* を最大化できる．したがって， CAHTCHA 1 問あたりの正答率は，

$$P_{mw} = P(Y_w = S, X = S) + P(Y_w = H, X = H) \tag{2.9}$$

となる．ただし，

$$\begin{aligned} P(Y_w &= S, X = S) \\ &= P(Y_w = S|W = t)P(W = t|X = S) \\ &\quad + P(Y_w = S|W = f)P(W = f|X = S) \end{aligned} \tag{2.10}$$

$$\begin{aligned} P(Y_w &= H, X = H) \\ &= P(Y_w = H|W = t)P(W = t|X = H) \\ &\quad + P(Y_w = H|W = f)P(W = f|X = H) \end{aligned} \tag{2.11}$$

である．

2.2 従来研究

2.2.1 CAPTCHA の定義

Ahn ら [3] は，次のように CAPTCHA を定義した．

(P, V) を確率手的で対話的なプログラムのペアとし，それぞれ乱数 r, r' を用いて対話が行われた結果，出力 $\langle P_r, V_{r'} \rangle \in \{accept, reject\}$ を得るとする．ここで，P をエンティティ (人間またはロボット) とし，V をテストプログラムとする．テストプログラム V があるエンティティ P を認証する確率 $\mathbf{Succ}_P^V$ は，式 (2.12) のようになる．

$$\mathbf{Succ}_P^V = \Pr_{r,r'}[\langle P_r, V_{r'} \rangle = accept] \tag{2.12}$$

テストプログラム V は， AI 問題 $\mathcal{P}$ を利用する．$\mathcal{P}$ は，$\mathcal{P} = (\mathcal{S}, \mathcal{D}, f)$ と定義される．ここで $\mathcal{S}$ は問題の実体の集合，$\mathcal{D}$ は $\mathcal{S}$ の確率分布，f は $f : \mathcal{S} \to \{0,1\}^*$ となる解答の実態を生成する写像である．あるプログラム A が，高々 τ の時間で式 2.13 に示されるように少なくとも δ の確率で解答できる AI 問題 $\mathcal{P}$ を，(δ, τ)-AI 問題と称する．

$$\Pr_{s \leftarrow \mathcal{D}_{\mathcal{S}}, r}[A_r(s) = f(s)] \geq \delta \tag{2.13}$$

CAPTCHA とは， 式 (2.12) の V に相当する．ある人間の集団 α が β より大きい確率で解答できるテスト V を，(α, β)-CAPTCHA と称する．異なる人間の集団 α_0, α_1 では， 対応する β_0, β_1 も異なる点に注意を要する．さらに，式 2.13 に示されるプログラム A を用いたいかなるロボットも η より高い正答率を出力できない場合，(α, β, η)-CAPTCHA と称する．CAPTCHA が機能するためには，$\beta > \eta$ でなれればならない．

2.2.2 CAPTCHA の分類

本論文では，CAPTCHA を次のように大別する．以降の節で，それぞれについて詳細を述べる．

- 視覚型 CAPTCHA
- 聴覚型 CAPTCHA
- 言語型 CAPTCHA

2.2.3 視覚型 CAPTCHA の構成に関する研究

視覚型 CAPTCHA の分類

視覚型 CAPTCHA には，大別すると次の種類がある．

- 文字列 CAPTCHA (String-based CAPTCHAs)
- イメージ CAPTCHA (Image-based CAPTCHAs)

- モーション CAPTCHA (Motion-based CAPTCHAs)

文字列 CAPTCHA については，テキスト CAPTCHA (Text-based CAPTCHAs) と称する研究者も多いが，この表記は言語型 CAPTCHA と混同しやすい点に注意を要する．本論文ではこのような混乱を避けるため，一般にテキスト CAPTCHA と呼ばれる方式については，文字列 CAPTCHA と称する．

文字列 CAPTCHA

現在最も普及している文字列 CAPTCHA は，画像に難読化の施された文字列を埋め込み，利用者にその読解をさせる方式である．

文字列 CAPTCHA では，難読化した画像に対する人間とロボットの認識能力の差を利用する．CAPTCHA が発明された 2000 年当時は OCR 関連技術の発達が未成熟であったので，雑音の挿入や画像の変形で難読化された文字列の認識は，ロボットに解くことは難しいと期待されていた．一方で人間の認識能力は，事物をあるまとまった全体像として認識するゲシュタルト理論 [42] により，ある程度の難読化には頑強である．また，ジオン (Geon) 理論 [43] によれば，視覚システムは物体を基本形状の集合として扱い，ある程度の特徴があればその認識が可能である．

基本的な文字列 CAPTCHA には，次の方式がある．

- *PessimalPrint* [44]
- *Gimpy* [45]
- *BaffleText* [46]

PessimalPrint [44] は，文字列境界線の不明瞭化や点雑音の追加により，紙媒体での印刷やコピーの劣化を模擬する方式である．図 2.1 に例を示す.

Gimpy [45] は，2 つの単語が重複した画像を複数提示し，それらの解読をさせる方式である．図 2.2 に例を示す．Gimpy の簡易版となる *EZ-Gimpy* では，1 つの文字列を画像に埋め込み，歪みや雑音を追加して難読化を施す．難読化の具体的な方法は，画像への歪みの付加や，白／黒の直線やグリッド線／等高線の追加がある．*EZ-Gimpy* で用いる文字列は，初めは単語のみであったが，安全性の強化を目的とし，アルファベットと数字で構成されたランダムなものを使用する場合も作成されている．

BaffleText [46] の例を図 2.3 に示す．この方式は，多様なフォントを用いて文字列を画像に埋め込み，その後に雑音画像とのマスクを取ることで難読化を行うテストとして提示される画像には，単語や発話可能な文字列が埋め込まれる．

文字列 CAPTCHA の派生形には，次の方式がある．

- *Sequenced Tagged CAPTCHA* [47]
- 手書き文字を模擬する方式 [48]

- 解答方式にドラッグアンドドロップを取り入れた方式 [49]
- *iCAPTCHA* [50]
- 文字列 *reCAPTCHA* [51]

Sequenced Tagged CAPTCHA [47] では，各文字に数字がタグ付けされた画像を利用する．文字と数字を重ねて配置することで，機械による各要素の分離・認識を困難にする．利用者は，タグ付けされた数字に従って認識した文字を並び替えて解答する．

手書き文字を模擬する方式 [48] は，多様な文字の形状を取り入れた方式である．図 2.4 に例を示す．この方式は，選択されたフォントを自動的に変形して，手書き文字を模擬した多様な文字画像を作成する。

Desai ら [49] は，利用者に文字画像の認識と，マウス操作による解答を課す方式を提案した．

iCAPTCHA [50] は，リレーアタックに対する安全性を強化した方式である．リレーアタックでは，攻撃者が一旦 CAPTCHA からの問題を貯めこみ，それを第三者に解かせた結果を解答として送信するため，正規利用者に比べて応答時間が長い傾向がある．*iCAPTCHA* は，サーバとクライアント間のインタラクションを頻繁に行うことで生じる応答時間の差を利用して，正規利用者とリレーアタックを識別する．

文字列 *reCAPTCHA* [51] は，図 2.5 のように，役割の異なる 2 つの文字列を問題として提示する．それぞれの文字列には，片方に CAPTCHA としての役割を，もう片方には紙媒体などの電子化作業促進の役割を与える．この方式では，CAPTCHA として使用された文字列の読解ができれば，認証が成功する．文字列 *reCAPTCHA* では，CAPTCHA として使用する文字列については，文字同士を隣接／重畳させ，画像に回転／歪みを与えることで難読化する．電子化作業促進の役割は，OCR のスキャンが失敗した紙媒体をテスト画像として提示し，人間に解読させることで行う．

図 2.1: *PessimalPrint* の例 [44]
Fig. 2.1: Example of *PessimalPrint* [44].

イメージ CAPTCHA

近年の OCR 関連技術向上により，文字列 CAPTCHA の脆弱性が指摘されている [7, 52]. このため，ロボットによる認識がより困難であると期待される写真やイラストを用いたイメージ CAPTCHA が数多く提案されている．

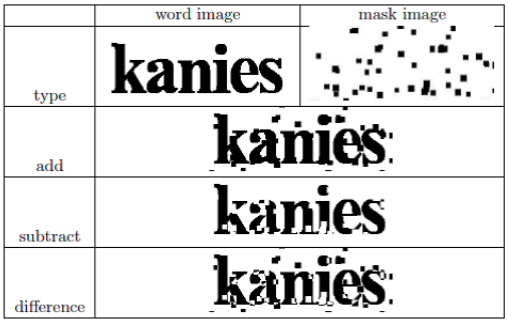
文字列 CAPTCHA では，難読化した画像に対する人間とロボットの認識能力の差を利用していたが，イメージ CAPTCHA ではそれに加えて，画像の意味論的な解釈の難しさ (タグ付けや共通項目の抽出) や視点や向きの違いに頑強な認識の難しさなどを利用する．人間は，前述のゲシュタルト理論 [42] やジオン理論 [43] により，事物をまとまったものとして解釈する能力に長けている．また， 2 次元または 3 次元の事物を心的に回転させる能力であるメンタルローテーション [53, 54] は，人間の高度な認識能力として知られている．

イメージ CAPTCHA には，次の方式がある．

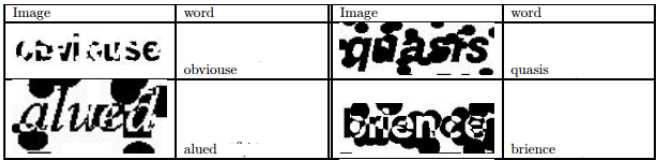
- *Bongo* [45]
- *PIX* [55]
- 3D モデルから 2D 画像を生成する方式 [56]
- *Collage CAPTCHA* [57]
- *Asirra* [20]
- *What's Up CAPTCHA* [58]
- *IMAGINATION* [21]
- 3D CAPTCHA [59]
- *IC-CAPTCHA* [60]
- *Locimetric* 型メンタルローテーション *CAPTCHA* [61]

Example of *Gimpy*　　Example of *EZ-Gimpy*

図 2.2: *Gimpy* の例 [45]
Fig. 2.2: Examples of *Gimpy* [45].



(a) Examples of using a mask degradation



(b) Example of *BaffleText*

図 2.3: *BaffleText* の例 [46]
Fig. 2.3: Examples of *BaffleText* [46].

- 非現実画像 *CAPTCHA* [62]
- *No-CAPTCHA* ／イメージ *reCAPTCHA* [63]

Bongo [45] は，ある特徴を持って分類された 2 種類の画像の集合と，それらとは別の 1 つの画像利用者に提示し，その画像がいずれの集合に属するかを解答させる．図 2.6 の例であれば，7 角形の画像がチャレンジとして提示されているので，右の集合が正解となる．この方式は 2 択問題であるので，利用者に複数回の解答させることで，ランダム推測攻撃に対する安全性を強化する必要がある．

PIX [55] は，共通の特徴を持つ複数画像を利用者に提示し，その共通する特徴を解答させる方式である．例えば,「自転車，バイク，船，飛行機」の画像が提示された場合,「乗り物」が解答になる．

Hoque ら [56] は，3D モデルから光源処理や向きを変えて生成した 2D 画像を CAPTCHA に利用した．この方式は，ほぼ無制限に多様な 2D 画像を生成できるため，過去に出題された画像のデータベースを利用したマッチング攻撃に対して頑強である．

Collage CAPTCHA [57] は，利用者に複数画像と 1 つのタグを提示し，タグに相当する画像を選択させる方式である．

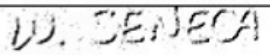
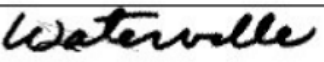
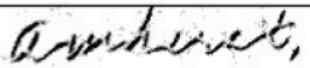
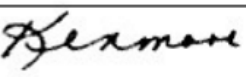
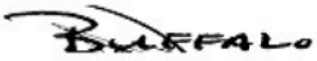
Truth Word	Test Image	Transformations
WSeneca		Add noise and convolve with the mask
Waterville		Spread image, add lines and arcs
West Seneca		Add less noise, lines, rectangle, and convolve with the mask
Amherst		Add less noise, spread image and wave it
Lockport		Add noise, line, rectangle, convolve with mask
Kenmore		Spread and blur image
Buffalo		Add arcs and lines
Rochester		Spread image and add circles/arcs

図 2.4: 手書き文字を模擬した *CAPTCHA* の例 [48]
Fig. 2.4: Examples of *Handwritten CAPTCHA* [48].

Asirra [20] は，複数の犬と猫の画像を利用者に提示し，そこからすべての猫の画像を選択させる方式である．

What's Up CAPTCHA [58] は，提示された画像を回転させ，正しいい向きに直すように利用者に指示する方式である．例えば，ひっくり返ったコップを戻すような作問がなされる．

IMAGINATION [21] は，難読化した画像を用いて，2 段階の処理で認証を行う．この方式は，最初に 8 つのアスペクト比が異なる長方形画像をタイル状に並べたものを提示し，利用者にいずれかの画像の中心をクリックするように指示する．利用者が正しく中心位置をクリックできた場合，次に *IMAGINATION* は，その画像を拡大し選択肢の中から画像に適したタグを選択させる．

3D CAPTCHA [59] は，3D の文字画像を利用者に提示し，利用者に解答させる方式である．この方式は，3D の文字画像は 2D のものより外観に多様性があるため，ロボットにより認識が困難である点を利用している．

IC-CAPTCHA [60] は，ある事物の画像に対し，楕円や長方形などの妨害画像を自動挿入したものを利用者に提示し，その事物の名称を解答させる方式である．

Locimetric 型メンタルローテーション CAPTCHA [61] は，図 2.7 のように，ある 3D の事物を 2 つの異なる角度の視点から提示する．その際，出題画像にはマーカーが追加され

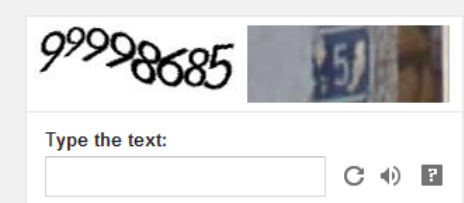


図 2.5: 文字列型 *reCAPTCHA* の例
Fig. 2.5: Examples of *String-based reCAPTCHA*.

ており，回答画像にはマーカーはない．この方式では，回答画像上でマーカーに対応する位置を示した利用者を，正規利用者として扱う．

非現実画像 CAPTCHA [62] では，3D モデルを使用して 2 つの事物がのめりこんだ非現実事物を作成し，通常の 3D モデルと並べて配置し，利用者に非現実事物を解答させる方式である．図 2.8 の例では，画面中央付近の犬と車がのめりこんだ事物が解答となる．

No-CAPTCHA ／イメージ *reCAPTCHA* [63] は，図 2.9 (a) のような No CAPTCHA と，(b) のイメージ型のハイブリッド方式である．No CAPTCHA では，利用者はチェックボックスをクリックするだけで CAPTCHA の認証を実施できる．イメージ方式の場合は，利用者は提示された 9 つの画像から，指示された特定の事物を選択する．Sivakorn ら [11] によるブラックボックス手法を用いた解析によれば，reCAPTCHA システムは利用者の信頼度を cookie やブラウザ環境を用いて評価し，その信頼性に合わせて No CAPTCHA またはイメージ CAPTCHA を提示する．

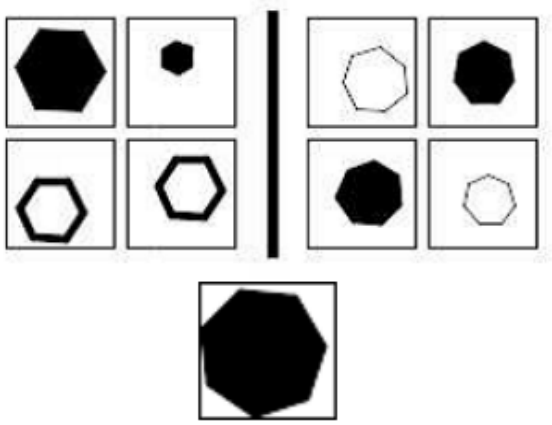


図 2.6: *Bongo* の例 [45]
Fig. 2.6: Example of *Bongo* [45].

イメージ CAPTCHA には，エンターテイメント性を兼ね備えた方式も提案されている．

- *Jigsaw Puzzle CAPTCHA* [22]
- 4 コマ漫画 CAPTCHA [64]

Jigsaw Puzzle CAPTCHA [22] は，その名の通りジグソーパズルを利用者に解かせることで認証を行う．この方式では，各パネルのエッジを，ランダムな鋸型に加工することで，パネル同士のエッジ画像の一致度を利用する攻撃 [65] に対して頑強である．

4 コマ漫画 CAPTCHA [64] は，利用者に対して，ランダムな順番に提示された画像を並び替えて 4 コマ漫画を完成させるように指示する方式である．利用者は，連続した画像から話の流れやユーモアを解釈する必要がある．

モーション CAPTCHA

モーション CAPTCHA は，事物のアニメーションや動画を利用した方式である．事物が動くことにより，静的な場合に比べて，提示する画像がより多様になる．

提示したアニメーションや動画の意味論的な解釈をさせる方式には，次のものがある．

- *Animation CAPTCHA* [66]
- 手話を用いた方式 [67]
- *NuCaptcha* [23]

Animation CAPTCHA [66] は，いくつかの事物がランダムにアニメーションしている動画を利用者に提示する．利用者は，それらの 1 つをクリックし，さらにそれが何かを解答できれば，認証を通過できる．

Sajad Shirali-Shahreza ら [67] は，聴覚障害者向けに手話を用いた方式を提案した．この方式は，利用者に手話動画を提示し，その意味に相当する事物を選択肢から解答させる．

NuCaptcha [23] は，文字が右から左に移動しているアニメーションを提示する．提示される文字は， simple と standard の難易度により異なる難読化がされる．利用者は，解読した文字をテキストボックスに入力して解答する．

提示したアニメーションや動画の意味論的な解釈に加えて，インタラクティブに操作を要求するものには，次の方式がある．

- *DCG CAPTCHA* [68, 69]
- *CAPTCHaStar* [70]

DCG (Dynamic Cognitive Games) CAPTCHA [68, 69] は，アニメーションする複数の事物を提示し，利用者に参照事物と同じ形や意味論的な属性をもつものを，解答領域にドラッグアンドドロップするように指示する．この方式は，応答時間の違いを利用して， 正規利用者とリレーアタックを識別する．

CAPTCHaStar [70] については，図 2.10 を用いて説明する．テスト開始時点では，図 2.10 (b) の状態で，小さな多数の星がランダムに配置されている．これらの星々は，利用者のマウス操作により全体が逐次移動する．利用者は，マウス操作により図 2.10 (d) のような何らかの事物が形成された状態を登録することで，*CAPTCHaStar* から認証を受けることができる．

2.2.4 聴覚型 CAPTCHA の構成に関する研究

聴覚型 CAPTCHA は，視覚障害者向けを中心に提案されている．

多くの聴覚型 CAPTCHA では，人間が持つ音声の選択的聴取能力であるカクテルパーティ効果 [71, 72, 73] などにより，ある程度の雑音には頑強な聞き取りができることが期待されている．また，一部のイメージ CAPTCHA と同様の趣旨で，音から意味論的解釈を必要とする方式もある．

聴覚型 CAPTCHA には，次の方式がある．

- 雑音挿入を利用した初期の方式 [74]
- 音の意味論的な解釈を利用する方式 [75]
- 音韻修復効果を用いた方式 [76]
- 人間とロボットの発話の違いを用いた方式 [77]
- VoIP 環境に適した方式 [78]
- *HearSay CAPTCHA* [79]
- *jCAPTCHA* [80]
- *SoundsRight CAPTCHA* [81, 82]
- ランダムな音韻例を意味論的な雑音として用いた方式 [37] (*MGK* 方式)
- 音声 *reCAPTCHA* [51]

Kochanski ら [74] は，白色雑音の挿入，エコーフィルタの適用，歌の一部を切り取り発話に重畳するなどを用いて発話される単語の難聴化を行った．

Holman ら [75] 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，利用者にその名称を解答させる方式を提案した．利用者は，聴覚と視覚のいずれかを用いてテストを認知できる．

福岡ら [76] は，発話の一部を削除しつつ白色雑音を挿入することで，発話を難聴化する方式を提案した．この方式では，数字の発話を採用している．音韻修復効果 [83] により，単純な音声の削除に比べて，人間に認識しやすい．

Gao ら [77] は，テストとして提示したテキストを利用者に発話させることで，人間とロボットを識別する方法を提案した．この研究では，人間の発話と機械合成音の特徴を分析し，その差を CAPTCHA として利用した．

Soupionis ら [78] は，VoIP で知らない利用者からの呼び出しを受け取った場合に CAPTCHA を課す方式を提案した．この方式では，発話の難聴化に “loud noise” を使用することにより，バックグラウンドノイズの音量が通常発話より小さいことを利用した ASR の雑音除去を妨害している．

HearSay CAPTCHA [79] は，利用者に提示した画像や音声を認識させ，発話で回答させる方式を提案した．

jCAPTCHA [80] は，複数単語を文法を無視した並びで利用者に提示し，その聞き取りをさせる方式である．この方式は，言語の文法を学習した ASR の認識精度を低下させるが，孤立単語認識型の ASR には脆弱である．

SoundsRight CAPTCHA [81, 82] は，動物の鳴き声や楽器の音を複数提示し，利用者に音の名称を解答させる方式である．この方式は，利用者の解答と応答時間を監視し，認証を実施する．

MGK 方式 [37] は，ランダムな音韻例を意味論的な雑音として用いる方式である．この方式では，単語の発話とランダムな音韻列を複数提示し，利用者に単語の発話部分のみを解答するように指示する．

音声 *reCAPTCHA* [51] は，雑音が挿入された数字発話の聞き取りをさせる方式である．雑音の種類には，ヒス雑音や音の高周波成分の変形などがある．

2.2.5 言語型 CAPTCHA の構成に関する研究

言語型 CAPTCHA には，人間の記述した豊富な分量の自然文を作問に必要とする方式と，必要としない方式がある．前者の方式は，問題文として収集した自然文を提示する方式で，数に制限のない作問のためには，十分な自然文を収集する必要がある．

豊富な分量の自然文を必要とする言語型 CAPTCHA には，以下の方式がある．

- 単語の意味の曖昧性を用いた方式 [32]
- 文章が示す話題の識別をさせる方式 [33]
- 機械合成文を用いる方式 [30] (*KK* 方式)
- 機械翻訳文の違和感を用いる方式 [34]
- *Voigt-Kampff Test* [84]

Bergmair ら [32] は，単語の意味の曖昧性解消がロボットには困難である点を CAPTCHA に利用することを検討した．この方式では，ある自然文のある単語を類義語・同義語またはそれ以外の語で置換し，利用者に対して文意が通じる内容かどうかを解釈させる．例え

ば move を置換対象とする場合，run, go などへの置換は意味が通じるが imporess, strike では意味が通じない可能性が高い．

Egele ら [33] は，ロボットによる文章の話題の特定が困難な点を作問に利用した．この方式では，利用者に複数の文を提示し，その中から意味論的に仲間外れの文を選択させる．例えば，料理について記述された複数文の中に異なる話題の文があれば，それが正答となる．

KK 方式 [30] は，人間の記述した自然文とスパムメールなどに使用されているワードサラダ文の識別をテストに用いる．この方式では，自然文とワードサラダ文では，人間の感ずる「文の自然さ」は異なるが，機械でその「文の自然さ」を評価することが困難である点を CAPTCHA に利用している．この方式については，4 章にて詳細な分析を行う．

山本ら [34] は，人間の記述した自然文と，人間にとって違和感のある機械翻訳文の識別をテストに用いた．機械翻訳文には，母国語と非母国語の間で機械翻訳を繰り返すことで，人間が大きな違和感を持つ文を利用する．この方式では,「文の違和感」の評価がロボットには困難である点を CAPTCHA に利用している．

研究ではないが，1968 年に発表されたディックの SF 小説 [84] においては,「このカバンは官給品なんだ．赤ん坊の皮でできている.」などと，異様な内容の文章を聞かせ，異様部分に対する身体的・感情的反応の時間遅れを計測する *Voigt-Kampff Test* のアイデアが示されている．

次に示す方式では，単語のみを用いたり，ある特定のパタンに沿った文章を問題文として提示するため，自然文の収集を必要としない．

- クイズを用いた CAPTCHA [35, 36] (*WtP* 方式)
- *Knock Knock Jokes* を用いた方式 [85]
- *SemCAPTCHA* [86]

ホワイトハウスのオンラインイン請願システム [35] では，ユーザアカウント取得の際に，一般常識で解けるクイズ形式の CAPTCHA を課していた．この *WtP* 方式については，4 章にて詳細な分析を行う．

Ximenes [85] らは，*Knock Knock Jokes* の解釈能力を用いた方式を提案した．この方式は，駄洒落という音的な特徴を CAPTCHA に利用している．

SemCAPTCHA [86] は，利用者に 3 つの単語を提示し，その中から仲間外れを選択させる方式である．例えば,「リンゴ，バナナ，ラッパ」の場合は，ラッパが答えになる．

2.2.6 CAPTCHA への攻撃に関する研究

視覚型 CAPTCHA への攻撃

文字列 CAPTCHA への攻撃

Mori ら [87] は，機械学習による context shape matching により，Gimpy で 33%， EZ-Gimpy で 92% の攻撃に成功した．この方式の文字の特徴量である context shape vector は，文字を点と辺の集合とみなし，その相対位置を距離と角度の 2 次元対数極座標ヒストグラムで表したものである．この方式では，始めに難読化が施されていない文字画像を用いて context shape vector を学習し，テストデータのもつ context shape vector との比較で認識を行う．

Chellapilla ら [88] は，CAPTCHA への攻撃手法として標準的な構成となった，(1) ノイズ削減の前処理，(2) 1 文字ごとの分割，(3) 深層学習や SVM (Support Vector Machine) [89] による各文字の認識，の 3 段階処理を提案した．これ以降に，多くの文字列 CAPTCHA への攻撃 [90, 91, 92, 7, 93, 52] が報告されている．特に Goodfellow ら [93] は，深層学習を用いたストリートビュー上の数字を認識する技術を開発し，それを CAPTCHA の攻撃に転用することで，2013 年の文字列 reCAPTCHA に対する 99.8% の正答率を示した．

また，単語を用いた文字列 CAPTCHA への攻撃には，単語辞書を利用した攻撃方式 [94, 95] も提案されている．

イメージ／モーション CAPTCHA への攻撃

IMAGINATION [21] などのイメージ CAPTCHA は，複数画像を 1 枚にまとめることで，再分割を困難にして個別画像の認識を難しくしている．Zhu ら [65] は，エッジ検出アルゴリズムにより，画像の再分割を可能にした攻撃を提案した．

Golle [96] は，SVM [89] を用いて Asirra に対する攻撃を成功させた．

Bursztein [97] は，2011 年に NuCaptcha [23] に対しての攻撃を報告している．この攻撃方法は，静止画 CAPTCHA に対する 3 段階の処理に加えて，アニメーションの静止画フレームへの分割と，各フレーム間を解析して CAPTHCA 部分の抽出を行う．

Sivakorn ら [11] は，2016 年の No-CAPTCHA ／イメージ reCAPTCHA に対する攻撃を報告した．この報告によれば，reCAPTCHA は信頼性の高い利用者にはチェックボックスのクリックだけで認証を行う．Sivakorn らは，ウェブの自動巡回を 9 日間行い生成した cookie を用いて reCAPTCHA を「信頼」させ，意図的にチェックボックス reCAPTCHA を出題させることに成功した．一方で reCAPTCHA は，信頼性の低い利用者にはイメージ CAPTCHA を提示する．Sivakorn らは，深層学習によるイメージへのタグ付けや，出題されたイメージの履歴を利用し，70.78% の確率でイメージ reCAPTCHA への攻撃に成功した．

聴覚型 CAPTCHA への攻撃

初期の 聴覚型 CAPTCHA を提案した Kochanski ら [74] は，自分たちの提案方式について ASR を用いた攻撃を実施し，その安全性を評価した．彼らによれば，SNR (Signal to Noise Rate) を 5db 程度とした白色雑音の挿入による単語発話の難聴化を行うと，人間が 90% 以上の正答率であるのに対し，2002 年当時の ASR の正答率は約 10% である．

Tam ら [98, 99] は，AdaBoost [100]，SVM [89]，k 近傍法 [101] といった複数の機械学習による分類機と PLP (Perceptual Linear Predictive) [102] や MFCC [103] などの音の特徴量を組み合わせ，それらが聴覚型 CAPTCHA への攻撃成功率に与える影響を比較した．彼らの結果によれば， 2008 年の Google，Digg，reCAPTCHA といった商用サイトの聴覚型 CAPTCHA に対して，それぞれ 67%，75%，45% の確率で攻撃に成功する．

Bursztein ら [104, 105, 8] は，白色雑音や周期的な雑音である統計的雑音による難聴化が ASR に対して効果が小さいことを明らかにした．彼らの方式では，“auditory power” である RMS (Root Mean Square) に着目したサンプリングにより雑音除去を行う．彼らの結果によれば， 2008 年の Microfsoft や Google の聴覚型 CAPTCHA に対して，それぞれ 75%，33%の確率で攻撃に成功する．また，2011 年の eBay，Yahoo!，Microsoft の聴覚型 CAPTCHA に対して，それぞれ 82%，45.5%，49% の確率で攻撃に成功する．

佐野ら [10] は，HMM (Hidden Markov Model) ベースの機械学習アルゴリズムを用いて，2012–2013 年の聴覚型 reCAPTCHA を 58.75 % の確率で解答できることを示した．彼らの方式では，まず [8] と似た方法で雑音を除去してから，HMM による音声認識を行う．

言語型 CAPTCHA への攻撃

1.3 節に示したように，[32, 33, 34] で提案された言語型 CAPTCHA は，作問に使用する自然文に起因する脆弱性の克服が課題となる．

山本ら [34] は，言語型 CAPTCHA の新方式を提案する際に，作問に使用する自然文の取得方法について検討している．この方式では，認証フェーズとは別に自然文の収集フェーズがある．自然文の収集フェーズでは，利用者に 1 枚の画像を提示し，その内容を説明する文の作成を促す．彼らの主張によれば，作成された文をインターネット上で検索し一致するものを棄却することで，新規でかつ非公開の自然文を収集できるとしている．しかしながら，この方法は，結託した複数の攻撃者が自然文を登録する攻撃に対して脆弱になる．

Bergmair ら [32] は，前述のとおり単語の曖昧性を利用した CAPTCHA を提案したが，同時にその安全性を評価している．Bergmair らは，提案した比較的単純な単語の曖昧性の問題に対しては，2004 年当時の NLP (Natural Language Processing) 技術でも 65% の確率で正答されたと報告している．

クイズ方式の CAPTCHA に関しては，IBM の Watson [106] のように，自然言語で質問を受け付け正しい解答をする人工知能の登場により，無力化しつつあるとの意見 [107] がある．

言語型 CAPTCHA を直接攻撃した研究ではないが，ワードサラダやフィッシングメールの検出に関する研究は数多く報告されている．

基本的な自然文とワードサラダの識別方法は，文中の文字／形態素／単語 N-gram の出現確率を用いる方法である．森本ら [108] の報告によれば，この方式は N の値により大きく精度が変わるという問題がある．ワードサラダの生成に使用されたマルコフ連鎖モデルの階数が大きい場合，自然文との識別は困難になる．森本ら [108] は，この問題を克服するため，離散型共起表現の出現数に着目した方式を提案した．離散的共起表現とは,「もし…ならば」などと，離散的だが組み合わせて使用することが多い表現である．

Ntoulas [109] は，ワードサラダで構成されたウェブページを，その内容から判別する方式を提案した．この方式では，N-gram のほかに，ページの特徴をアンカーテキストやページの可視領域などから取得し，それらを用いてウェブページの識別を行った．

Park ら [110] は，金融機関などを装い利用者のパスワードを盗むフィッシングメールを見分ける能力について，人間と SVM [89] を用いたロボット間での比較をした．この報告では，フィッシングメールの識別という限定された作業でさえ，人間にのみ識別可能な文章の存在があると指摘している．

リレーアタック

リレーアタックは，攻撃者が中継者のように振る舞い，第三者を利用して CAPTCHA を解く攻撃方式である．リレーアタックでは第三者（人間）を利用するため，チューリングテストとしての CAPTCHA を破っているわけではない点に注意を要する [1].

Chellapilla [90] は，低賃金で雇用できる労働者に CAPTCHA を解かせることを検討した．Podec Trojan [111] は，同様の趣旨で CAPTCHA を解くウイルスである．

TrojanCaptcharA [112] は，ポルノサイトを囮にして感染者に CAPTCHA を解かせるウイルスである．

Egele ら [113] の方式では，利用者からのリクエストを遮断し，攻撃者が解きたい正規サイトの CAPTCHA を割り込ませる．攻撃者は，利用者が CAPTCHA を解いた後にそのリクエストを再開させることで，利用者に不正を気付かれることなく CAPTCHA を解ける．

2.2.7 CAPTCHA のユーザビリティに関する研究

Sauer ら [114] は，聴覚型 CAPTCHA の 視覚障害者に対するユーザビリティを報告した．視覚障害者は 2008 年の音声 reCAPTCHA [51] に対して，1 問あたり 60–65 秒を要し，その正答率は 46% であった．

Bigham ら [24] は，晴眼者を含めた聴覚型 CAPTCHA のユーザビリティを報告した．彼らは，視覚障害者の約 8%，晴眼者の約 44% が聴覚型 CAPTCHA を解くのをやめてしまうと報告している．正答率に関しても，聴覚型 CAPTCHA は 40% 程度にとどまり，視覚型 CAPTCHA の約 80% に比べると低い．Bigham らは，視覚障害者に対する CAPTCHA

のユーザインタフェースの問題も指摘し，より細かい粒度で音声の再生を可能にするように提案している．

Bursztein ら [25] は，複数の商用サイトで使用されていた視覚型と聴覚型 の CAPTCHA について，大規模なユーザビリティ調査をしている．彼らの報告は，聴覚型 CAPTCHA は視覚型よりも，認知に個人差があることを示した．3 人の解答が完全一致する場合と完全に異なる場合の確率を調べたところ，視覚型は 70%， 5% であるのに対して，聴覚型は 31.2%，33.6%であった．また，Bursztein らが調査した聴覚型 CAPTCHA は，視覚型 CAPTCHA に比べて 2 倍の応答時間である 20 秒台を要するが，その正答率は半分程度にとどまることを示した．

Belk ら [115] は，利用者の好みを取り入れることで，CAPTCHA に対する応答時時間を短縮できないかを検討した．この報告は，利用者の好む認知方式 (言語型とイメージ型) によって，彼らの文字列 CAPTCHA や イメージ CAPTCHA を解答する応答時間に違いがある可能性を示唆している．

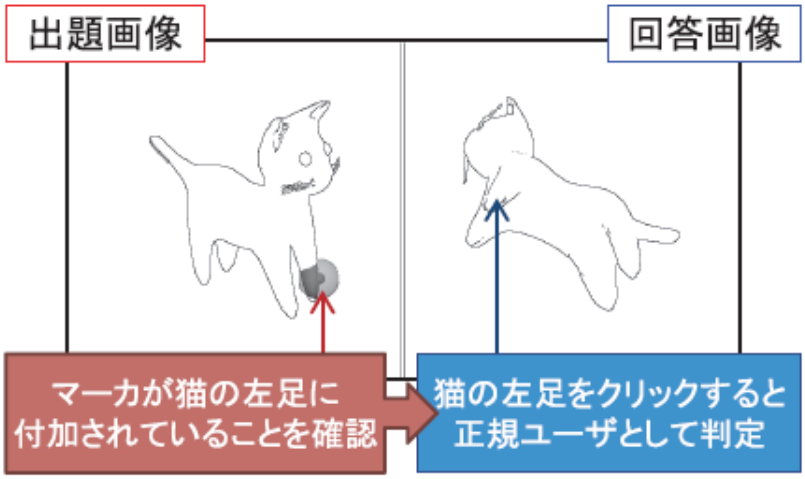
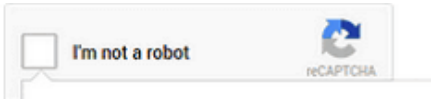


図 2.7: *Locimetric* 型メンタルローテーション *CAPTCHA* の例 [61]
Fig. 2.7: Example of *Locimetric and-based Mental Rotation CAPTCHA* [61].

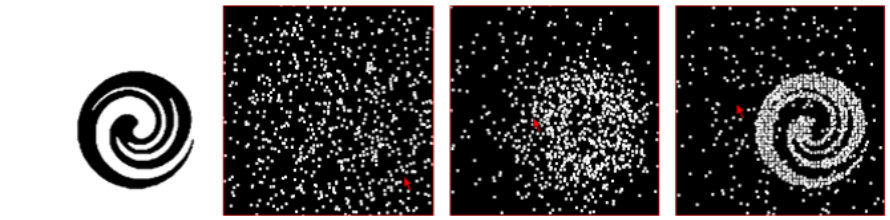
図 2.8: 非現実画像 *CAPTCHA* の例 [62]
Fig. 2.8: Example of Unreal-image *CAPTCHA* [62].



(a) *No CAPTCHA reCAPTCHA*

(b) *Image reCAPTCHA*

図 2.9: *No-CAPTCHA and Image-based reCAPTCHA* の例
Fig. 2.9: Examples of *No-CAPTCHA and Image-based reCAPTCHA*.



(a) A random starting picture. (b) A sample unsolved challenge. (c) An almost solved challenge. (d) A correctly solved challenge.

図 2.10: *CAPTCHaStar* の例 [70]
Fig. 2.10: Example of *CAPTCHaStar* [70].

第3章　クイズ CAPTCHA の脆弱性

3.1　導入

言語型 CAPTCHA は，視覚や聴覚に制限を受けないアクセシビリティの高い方式である．その代表例には，*KK* 方式 [30] や *WtP* 方式 [35] などがあり，後者は視覚型や聴覚型 CAPTCHA の代替として，アメリカ合衆国の政府系サイトであるオンライン請願システム *We the People* で利用されている．

本章では，既存のクイズ CAPTCHA である *WtP* 方式を分析し，その脆弱性を明らかにする[1]．

WtP 方式のアルゴリズムは公開されていないため，その作問手法は不明である．そこで本論文では，*WtP* 方式の安全性を実験的に評価する．まず， 実際に *We the People* のサイトにおいてアカウント作成の手続きを行い，提示されたクイズ CAPTCHA を収集し作問内容を分析する．分析では，*WtP* 方式が脆弱となる理由の仮説を立て，それを実験的に検証する．次に，分析結果から効果的な攻撃方法を検討し，実際に攻撃用ソフトウエアの実装を通して，*WtP* 方式の安全性を評価する．

3.2　*WtP* 方式

WtP 方式は，利用者がサイトの新規アカウントの作成手続きに入ると，図 3.1 のようなクイズをテキストで提示する．利用者は，視覚または聴覚で問題文を認識し，その解答をフォームに入力する．*WtP* 方式は，クイズの正答を入力した利用者に対しては，以降のアカウント作成作業を許可する．

3.3　*WtP* 方式に対する攻撃方法の提案

3.3.1　*WtP* 方式の脆弱性に関する仮説

本研究では，*WtP* 方式に対する脆弱性分析の着眼点として，2 つの仮説を検討する．

仮説 1. *WtP* 方式では，クイズの問題文を一定の「型」で固定し，その一部を変更して作問する．問題文の「型」の種類は，限定的で多様性が小さい．

[1] *KK* 方式については，4 章にて，提案方式と比較した安全性の評価を行う．

たとえば，図 3.1 のクイズは，リスト “35, eleven, twenty nine, eleven and 4:” と条件（属性）指定文 “the 4th number is?” で構成されている．このリストの一部を “blue” や “red” などの英単語に置換し，条件文を “the 2nd color is?” などに変更すれば，別のクイズが作成できる．この作問手法例では，提示したリストから特定条件を満足する事物を選択させるクイズを作問できる．

仮説 1 が成立するクイズは，問題文の型の種類が限定的であるため，一般的な文章に比べて文意や文脈の解釈は容易になる．さらに問題文の型を正規表現や Backus-Naur 記法で定義できれば，非常に簡単なパタンマッチングのプログラムで作問内容を認識できる．

仮説 2. *WtP* 方式は，基本的な単語で記述された文をクイズとして提示する．利用者は，そのクイズを一般的な知識で解くことができる．

人間の知る情報の量には個人差があるため，一般常識で解けないクイズは CAPTCHA の利用者を制限してしまう．また，人間は情報を蓄えるという点ではロボットに及ばないため，解答に必要となる情報の量を増やすことでクイズを難しくしても，CAPTCHA としては意味がない．したがって，*WtP* 方式が CAPTCHA として適切に実装されていれば，仮定 2 は成立するはずである．図 3.1 の例では，利用者は数字の認識とその大小関係が分かれば，クイズに解答できる．

仮説 2 が成立するならば，クイズの解答に特別なデータベースは不要である．特にクイズの内容が *SemCAPTCHA* [86] のような単語のグループ分け問題である場合は，一般的で規模の小さい語彙データベースでも有用となる．

3.3.2 *WtP* 方式の作問結果の分析

本研究では，これらの仮説に基づく攻撃方法を検討する．方針としては，*WtP* 方式の問題文は型や語彙の種類に制限があるという仮説を利用して，簡単な構文解析プログラムと語彙データベースによってクイズに解答をする．本節はその前段階として，実際に *We the People* から収集した *WtP* 方式のクイズ 1000 問について，仮説 1 と仮説 2 が成立するかどうかを分析する．

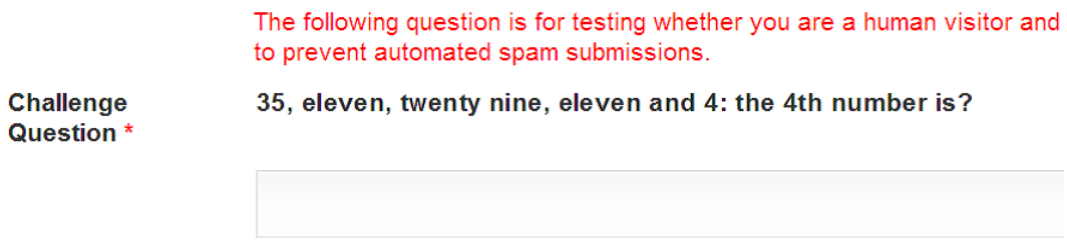


図 3.1: ホワイトハウスで使用されているクイズ CAPTCHA の例 (2014 – 2016 年 12 月現在)
Fig. 3.1: Example of CAPTCHA used on White House Website in 2014 – December, 2016.

表 3.1: 2014 年 2–3 月に使用された *WtP* 方式の問題の分類

Table 3.1: Category of Questions in *WtP*-scheme in February–March, 2014.

Type of Questions	Samples	Appearance Rate [%]
Type A	*Ant, yellow and red: the 2nd colour is?* *In the number 9501081, what is the 3rd digit?*	29.6
Type B	*The list bank, snake, mosquito, restaurant, finger and cake contains how many body parts?* *From days Monday, Friday, Sunday or Wednesday, which is part of the weekend?*	26.5
Type C	*Of the numbers thirty one, 91, 30, seventy nine, forty five or sixty one, which is the largest?* *Enter the biggest number of eight, twenty one or 46:*	21.5
Type D	*Enter the number eighty seven thousand nine hundred and sixty nine in digits:* *What is thirty eight thousand seven hundred and seventy as digits?*	17.0
Type E	*What's 16 - three?* *3 plus ten is what?*	3.8
Type F	*If the cake is brown, what colour is it?* *What day is today, if tomorrow is Thursday?*	1.6

表 3.2: *WtP* 方式を攻撃する C# プログラムの正規表現の例

Table 3.2: Examples of Regular Expressions of our C#-Program against *WtP*-scheme.

Type A	`^([^:]+)\s*:\s+(?:the\s+)?(.+)(KEYWORD)\s+is\s*[\?\.:]?\s*`
	`^(?:What\s+is|What's)\s+(?:the\s+)?(.+)\s+(KEYWORD)\s+in\s+(.+)\s*[\?\.:]?\s*$`
	Note that "KEYWORD" is "digit|weekend|number|color|colour|body\s+part".

Type B	`^The\s+list\s+(.+)\s+contains\s+how\s+many\s+(KEYWORD)\s*[\?\.:]?\s*$`
	`^How\s+many\s+(.+)\s+in\s+the\s+list\s+(KEYWORD)\s*[\?\.:]?\s*$`
	Note that "KEYWORD" is "digit|weekend|number|color|colour|body\s+part".

Type C	`^Which\s+of\s(.+)is\s+the\s+(highest|largest|biggest|lowest|smallest)\s*[\?\.:]?s*` `^([^:]+)\s*:\s+the\s+(highest|largest|biggest|lowest|smallest)\s+is\s*[\?\.:]?\s*`
Type D	`^Enter\s+the\s+number\s+(.+)\s+in\s+digits\s*[\?\.:]?\s*$` `^(?:What\s+is|What's)\s+(.+)\s(?:as\s+a\s+number|as\s+digits)\s*[\?\.:]?\s*$`

Type E	`^(?:What\s+is|What's)\s+(.+)\s+(OP)\s+(.+)\s*[\?\.:]?\s*$`
	`^(.+)\s+(OP)\s+(?:is\s+what|=|equals)\s*[\?\.:]?\s*$`
	Note that "OP" represents the string of "plus|add|minus|subtract|multiply|divide|\ + |\ − |\ ∗ |/".

Type F	(Customized Patterns)
	`(?:^The\s+colou?r\s+of|is\s+what\s+colou?r|what\s+colou?r\s+is\s+it)`
	`^(.+)[,:]\s+which\s+.+\s+weekend\s*[\?\.:]?\s*$`

仮説 1 が成立するならば，*WtP* 方式の問題文は，特定かつ少数のルールで定義できるはずである．また，仮説 2 が成立するならば，作問に使用される語彙やその内容に制限があるはずである．

そこで本研究では，収集した *WtP* 方式の問題文に対して正規表現への置き換えを行い，それらを少数の型に分類する実験を試行した．

結果を表 3.1 に示す．表 3.1 は，表 3.2 の正規表現を利用して，*WtP* 方式のクイズを 6 種類の型に分類した際の作問例と，その型に属するクイズの出現率を示している．表 3.1 と表 3.2 における各型の特徴は，次の通りである．

WtP 方式における問題文の型の特徴

Type A 単語のリストから，特定の条件に合致する単語のリスト中の位置 (順番) を，利用者に解答させる作問．

Type B 単語のリストから，特定の条件に合致するすべての単語またはその数を，利用者に解答させる作問.

Type C 数字のリストから，最も大きいまたは小さい数を，利用者に回答させる作問.

Type D 英単語として表記された数字を，利用者にアラビア数字に置換させる作問.

Type E 利用者に簡単な計算問題を解かせる作問.

Type F その他の作問．ただし，作問の内容は，Type A,B の派生型である.

作問に使用された単語は，表 3.2 に示されるように,「数字，色，曜日，体の部位」といった一般常識の語彙が利用されていた.

3.3.3 問題文の型に応じた攻撃方法

表 3.1 と表 3.2 の結果は，少なくとも今回収集した作問例に関しては，仮説 1 と仮説 2 が成立することを示している．そこで，これらの仮説に基づく攻撃方法を構築する．攻撃アルゴリズムは，提示されたクイズの型を表 3.2 の正規表現で分類し，その結果に応じた攻撃方法を用いる.

Type A, B の作問に対しては，質問内容に沿ってリストに含まれる単語を意味論的に分類する必要がある．この分類については，語彙のデータベースである FrameNet II [116, 117] [2] を利用する．Type A, B の作問に対する攻撃方法を次に示す.

WtP における **Type A, B** の作問に対する攻撃方法

1. 問題文を解析し，どのような属性の単語についての質問かを調べる．具体的には，表 3.2 にある “KEYWORD” や “digit” に相当する部分を検出する．ここで属性を表す単語をキーワードと称する.
2. 語彙データベースからキーワードが属する意味論的なグループを調べる.
3. 問題文を解析し，単語リスト部分を抽出する．各単語に対して，語彙データベースからその単語の属する意味論的なグループを調べて，キーワードと同じ意味論的なグループに属する単語をすべてチェックする.
4. Type A では，チェックした単語のリスト中の位置 (順番) を出力する．Type B では，チェックした単語の一覧または数を，質問に合わせた形式で出力する．チェックされた単語の数が 0 の場合は，エラーを出力する．また，Type A において，複数単語にチェックが入った場合もエラーを出力する.

[2] 辞書と類似語・同義語辞典を併せ持つデータベースの一種

Type C, D, E に分類された作問は，計算問題に過ぎない．これらに対しては単純な計算プログラムで解答できる．

Type F の作問に対する攻撃は，Type A, B への方法と最後の出力処理以外は同様になる．出力処理については，表 3.2 の Type F で示した 2 種類の型に合わせた個別対応を行う．

3.4 評価

3.4.1 実験方法

本研究では，3.3 節で示した方法に基づき攻撃用のソフトウエア実装した．このソフトウエアに対して，*We the People* から新たに収集した 300 問の作問結果を入力し，その正答率を調べる．

3.4.2 評価結果

本ソフトウエアは，99.7% の確率で正答を出力した．

本ソフトウエアの誤答は，表 3.2 の正規表現による分類の失敗に起因する．その理由は，3.3.2 節の分析では検討していない問題文の型が，評価実験で用いた 300 個の作問中に存在したためである．具体的な作問内容は “Which of hair, finger, heart, knee or toe is part of the head?” であり，“part of the head” をキーワードとして正しく認識できずエラーを出力した．

3.5 考察

本研究では，公開された語彙データベースと軽量なプログラムのみで *WtP* 方式の攻撃に成功した．この理由は，*WtP* 方式の問題文が明確に定義された (“well-defined”) 型をもつためである．今回試作したソフトウエアが誤答した例においても，その型を知ってしまえば対応は容易である．

この脆弱性の問題は *WtP* 方式のようなクイズ CAPTCHA に限らない．型に制限のある文章は，一般的な文章に比べて文意文脈解釈は容易である．したがって，型に制限のある文章を CAPTCHA に利用することは難しいと考えられる．

3.6 まとめ

本章では，クイズ CAPTCHA の代表例である *WtP* 方式の安全性を評価した．

WtP 方式のアルゴリズムは公開されていないため，本研究では，その作問例から *WtP* 方式の脆弱性となる特徴を推測した．次に，*We the People* のサイトから *WtP* 方式の作問を

収集し，推測した脆弱性についての分析をした．*WtP* 方式の作問はその問題文の型と語彙の種類に制限があるため，一般的な文章と異なり，正規表現によるパタンマッチングで作問内容の分類と解釈ができる．最後に，提案した攻撃方法に従い実装したソフトウエアを用いて，*WtP* 方式の安全性を評価した．本ソフトウエアは *WtP* 方式の作問を 99.7%の確率で正答した．したがって，*WtP* 方式は本ソフトウエアに対して脆弱である．

本章で実装したソフトウエアは，IBM の Watson [106] と異なり，巨大なデータベースやサーバは不要である．本ソフトウエアは，誰にでも利用できる軽量な語彙データベースと小規模のプログラムで構成されているため，非常に低いコストで *WtP* 方式を攻撃できる．

第4章　言語型 CAPTCHA の提案: 機械合成文の不自然度相対識別問題

4.1　導入

3 章では，言語型 CAPTCHA の 1 種であるクイズ CAPTCHA の安全性を評価した．クイズ CAPTCHA は，その問題文が明確に定義された (“well-defined”) 型を持つため，それを狙った攻撃に対して脆弱であった．

本章では，問題文が明確に定義できない (“ill-defined”) 型を持つ方式について考える．その代表例には *KK* 方式 [30] がある．

KK 方式は，ワードサラダや自然文を問題文として提示する．そのため *KK* 方式は，問題文の型の種類に関する制限はないが，作問の際にコーパスを必要とする．公開文章をコーパスに利用すると，*KK* 方式は検索エンジンを用いた攻撃に対して脆弱となる．非公開文章をコーパスに利用した場合は，その分量が少ないため，問題文として使いまわした自然文を狙った攻撃に対して，*KK* 方式は脆弱となる．以上のことを，実験により明らかにする．

さらに本章では，新しい言語型 CAPTCHA の特徴とアルゴリズムを示す．提案方式では，自然文に相当する文もマルコフ連鎖に基づき生成することで，*KK* 方式の持つ脆弱性を解決する．また実験により，提案方式の安全性と人間によるユーザビリティを評価する．

4.2　準備

4.2.1　*Ham* と *Spam*

本章では，*Ham* を「意味論的自然文」，*Spam* を「意味論的不自然文」と定義する．

鴨志田らは *KK* 方式 [30] において，*Ham* には自然文を，*Spam* にはワードサラダを用いた．

提案方式では，*Ham* と *Spam* の双方にワードサラダを用いる．*Ham* を合成するマルコフ連鎖モデルの階数 N_{Ham} と *Spam* の階数 N_{Spam} は，$N_{Ham} > N_{Spam}$ の関係を満たす．提案方式における *Ham* と *Spam* の「自然さ」の尺度は，1 つの問題を構成する *Ham* と *Spam* の組における相対的なものであることに注意を要する．

4.2.2 生成文の多様性とコーパスの多様性

与えられたコーパスから N 階マルコフ連鎖モデルを構築し，それによって生成された W_A 個のワードサラダのうち，重複を除いて互いに異なるものが W_U 個あるとする．本章では，$100 \times W_U/W_A$[%] を生成文の多様性と定義し，問題新規性の評価指標として扱う．例えば，生成した 100 個のワードサラダ中に 10 個の重複があった場合，その多様性は 90% になる．

コーパスに含まれる形態素 M-gram の集合を $\mathcal{D}_M$ とし，形態素 1-gram からなる集合を $\mathcal{D}_1$ とする．ある形態素 M-gram A $(= a_n a_{n-1} \ldots a_{n-M+1}) \in \mathcal{D}_M$ から連鎖する $M+1$ 番目の形態素の候補集合 $C_{(M,A)}$ は，

$$C_{(M,A)} = \{ \quad c \in \mathcal{D}_1 \mid P(X_{n+1} = c | X_n = a_n, \ldots, X_{n-M+1} = a_{n-M+1}) > 0 \}$$

となる．すなわち，生起確率が 0 より大きな候補の総数を表す．本章では $C_M = \sum_{A \in \mathcal{D}_M} |C_{(M,A)}| / |\mathcal{D}_M|$ を，形態素 M-gram でのコーパスの多様性と定義する．

4.3 *KK* 方式とその脆弱性

KK 方式の作問例を図 4.1 に示す．また，*KK* 方式のアルゴリズムを次に示す．

問題: 不自然な文を選択してください.

No. 1 エジソンなど，鉄鋼業や先住民の影響をし，ブラジル，人種を受けている．

No. 2 この原則はアメリカ合衆国憲法修正第 14 条に端的に現れている．

No. 3 1927 年の程度が海兵隊を求められた．

No. 4 この絵は 2 ドル紙幣の裏面図版に使用されている

解答: 1 と 3 が *Spam*.

図 4.1: *KK* 方式による作問例 [30]
Fig. 4.1: Sentences Synthesized by *KK*-scheme [30].

KK 方式のアルゴリズム [30]

1. コーパスから N 階マルコフ連鎖モデルを作る．

2. 自然文を h 個，ワードサラダを s 個，計 z 個の文をランダムな順で利用者に与える．

3. 利用者は z 個の文を，それぞれ *Ham* か *Spam* に分類して解答する．

4. 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．

KK 方式の特徴は，問題文に自然文を使用している点である．このため *KK* 方式では，検索攻撃を避けるため，秘匿文章をコーパスにすることを推奨している．しかしながら，非公開文章はその量が不足が不足するため，作問の際に同じ問題が使いまわされる可能性が高い．これは，Sivakorn ら [11] によるイメージ CAPTCHA への攻撃と同様の趣旨で，出題履歴を利用した攻撃に脆弱である．

KK 方式の脆弱性の詳細については，4.5 節と 4.6 節において，提案方式と安全性を比較して示す．

4.4 提案方式

4.4.1 提案方式の概要

KK 方式と提案方式の違いは，(1) *Ham* の生成方法と (2) 解答方式である．

(1) *Ham* の生成方法: 提案方式では，*Ham*，*Spam* ともにマルコフ連鎖で合成する．ただし，それぞれに使用する階数は $N_{Ham} > N_{Spam}$ を満たす．

Ham にワードサラダを用いることは，以下の点で自然文の利用に比べて優れている．よって，ロボットに対する *FAR* の改善が期待できる．

- 同一コーパスに対して，より高い生成文の多様性をもつ．
- 検索エンジンによるコーパスの特定が困難である．

提案方式は，上記の利点により，公開文章をコーパスとして使用できる．事前にある程度大きなコーパスを準備しておけば，問題を生成しながらコーパスを公開文章から収集し，マルコフ連鎖モデルを更新できる．定期的なモデルの更新により，生成文の多様性を維持することができる．

(2) 解答方式: ワードサラダは，自然文に比べて違和感の強い文章が生成されやすいため，人間に対する *FRR* の悪化が懸念される．その対策として，提案方式では 1 問ごとに *Ham* と *Spam* を両方提示し，そのどちらが相対的に自然 (もしくは不自然) かを問う方式を取りいれる．

問題: より不自然な文を選択してください.

No. 1 ($N_{Ham} = 2, N_{Spam} = 1$)

A 離れとは言い出せなかったが，その彼とほぼ同時に停留所に着くだろ

B 菜ならんで水を終ると，彼の自転車の間，つらくなって，一声であっ

No. 2 ($N_{Ham} = 3, N_{Spam} = 1$)

A 三歩下がり，邦子は気にいっていた．片方の壁には食器棚があり，

B 東亜戦争が，そして転んでいる海岸にあおむけになり，御隠居と戦っ

No. 3 ($N_{Ham} = 4, N_{Spam} = 1$)

A 亜紀子はきめた．水の表面がきらきらと輝きながら，小さく揺れて

B 加えた．入口までも四歳にあたえた．車を覚えにいるとすごいの気持

解答: 全て B が *Spam*.

図 4.2: 提案方式による作問例
Fig. 4.2: Sentences Synthesized by Our Proposal.

図 4.2 に提案方式の作問例を示す．作問例 1 は，*Ham* 単体では A，B とも自然さを感じることはできない．作問例 2 でも，人によっては違和感を感ずるかもしれない．しかし，*Spam* との相対比較にすることで，解答が容易になっている．

4.4.2 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. $N_{Ham} > N_{Spam}$ となる 2 つの階数を選択し，コーパスからそれぞれのマルコフ連鎖モデルを作る．
2. *Ham* と *Spam* を z 組作成する．
3. 各組に対して，*Ham* と *Spam* をランダムに選択肢 A と B に割り当て，利用者に提示する．利用者に，各組についてより自然 (*Ham*)，もしくは不自然 (*Spam*) な選択肢を解答させる．

表 4.1: 実験に用いたコーパスの特徴 (文字数，行数) = (80783, 5248)
Table 4.1: Features of our Corpus; (Number of Characters, Lines) = (80783, 5248).

N-gram	1	2	3	4	5	6	7
Number of Unique Words	7,893	34,469	60,790	73,632	77,532	77,526	76,395
Diversity of the Corpus (C_N)	4.403	1.785	1.231	1.075	1.023	1.008	1.002

4. 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．

コーパスは，問題新規性を満たすため，一定の作問数ごとに更新が必要になる．コーパスのサイズに依存して，十分な問題新規性が維持される作問数は変化するため，事前に多様性を実験的に確認することが望ましい．

4.5 評価

4.5.1 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．本実験は，2014 年 11 月から 2015 年 2 月の間に行った．

実験 1 生成文の多様性の評価

実験 2 検索エンジンを用いた攻撃による *Ham* と *Spam* の識別性の評価

実験 3 人間による評価

4.5.2 実験方法

共通設定と表記方法

次に，実験に共通する設定を示す．

- 青空文庫 [118] に登録されている 5 種類の現代語仮名遣いの文章を，まとめて 1 つのコーパスとして扱う．表 4.1 にその特徴を示す．
- 提案方式で *Spam* に用いるワードサラダの階数を $N_{Spam} = 1$ とする．*Ham* に用いるワードサラダの階数 N_{Ham} は，固定階数である $1, 2, 3, 4, 5$ と，幅を持つ階数 $[N_L, N_H] =$ $[1, 2]$,$[1, 3]$,$[2, 3]$,$[2, 4]$,$[3, 4]$,$[3, 5]$,$[4, 5]$ を用いて評価する．

- *KK* 方式を，$N_{Ham} = 7$ と $N_{Spam} = 1$ のワードサラダ識別問題として扱う．表 4.1 から， 7-gram におけるコーパスの多様性は $C_{N=7} \approx 1$ なので，本章では 7 階ワードサラダを自然文として扱う．
- 生成するワードサラダの文字数は，30 から 40 文字の範囲とする．

グラフの表記について，その注意点を示す．階数をパラメータとして扱う場合，N と $N_{diff} = N_H - N_L$ を使用する．固定階数の場合，$N_{diff} = 0$ とする．幅を持つ階数の場合，$N = N_H, N_{diff} = N_H - N_L$ とする．

実験 1 (多様性)

マルコフ連鎖モデルからワードサラダを，階数ごとに 50,000 個生成し，生成文の多様性を確認する．

実験 2 (検索エンジンによる識別性)

次の手順を，$N_{Ham} = 2, 3, 4, 5$, $N_{diff} = 0, 1, 2$ について，それぞれ 10 回行う．

1. N_{Ham} 階マルコフ連鎖モデルから，ワードサラダ 10 個を *Ham* として合成する．1 階マルコフ連鎖モデルから，ワードサラダ 10 個を *Spam* として合成する．
2. *Ham* と *Spam* を順番に 1 つずつ取り出して組を構成する．全ての組に対して，次の処理を行う．
   - Yahoo! 検索エンジンにそれぞれを問い合わせる．
   - 検索結果の上位 10 件にコーパスが含まれている場合，検索エンジンがコーパスを検出したと判断する．コーパスの検出が成功した場合，さらに次の判定を行う．
   - 次のいずれかの条件を満たした場合，検索エンジンにより *Ham* と *Spam* を正しく識別できたとする．
     - *Ham* のみコーパスが検出された．
     - *Ham* の方が *Spam* より上位検索結果で，コーパスの検出がされた．

     検索では各文について通常方式と完全一致方式の両方を実行し，上位となる方を利用する．

利用する検索エンジンは， Yahoo!，Google，Bing を候補に挙げ，事前調査し Yahoo! を選択した．Bing 検索は，ワードサラダからコーパスを検出する確率が低いため除外した．Google 検索は，プログラムによる連続した検索ができなかったため，効率の観点から除外した．

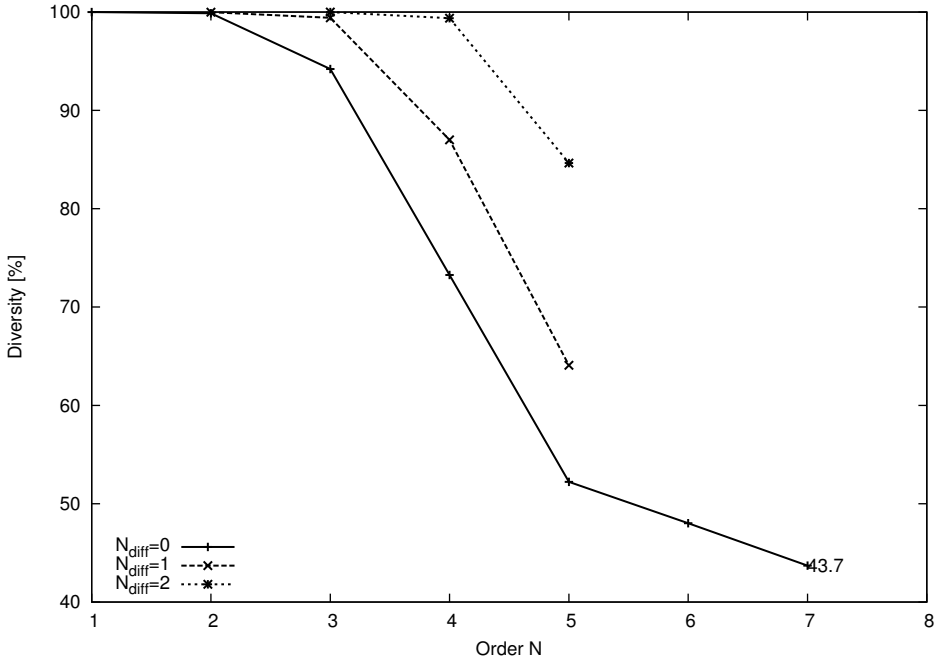


図 4.3: *N* 階ワードサラダの多様性
Fig. 4.3: Diversity of Sentences Generated by Markov Chain.

実験 3 (人間による評価)

次の手順で生成したエクセルファイルを用いて，テストを実施した．被験者は，男性 13 名，女性 3 名の日本人 16 名である．年齢構成は 18 歳から 65 歳であり，視力の状態は全盲 (光覚・手動弁・指数弁) が 2 件，弱視が 3 件，晴眼が 11 件であった．

1. 実験 2 で生成した *Ham* と *Spam* の組を，各 *Ham* の階数ごとに，無作為に 10 個選択する．

2. 選択肢として *A* と *B* の 2 つを用意し，各組ごとに *Ham* と *Spam* を無作為に割り当てる．

4.5.3 実験結果

実験 1 (多様性)

図 4.3 に，*N* 階ワードサラダの多様性を示す．

固定階数 ($N_{diff} = 0$) のグラフから， 自然文 ($N = 7$) に比べて，ワードサラダの多様性が十分高いことがわかる．特に $N < 4$ で顕著である．$N = 4$ を境に傾きが急峻になるが，以降

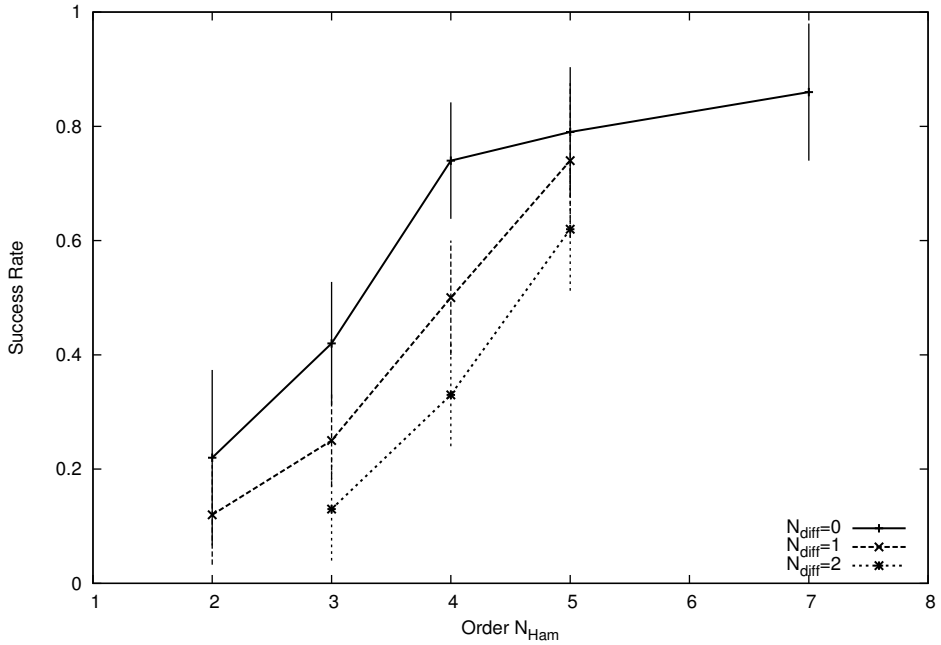


図 4.4: Yahoo! 検索エンジンによる *Ham* と *Spam* の識別能力
Fig. 4.4: Distinguishability Rate by Yahoo! Search Engine.

は N の増加に対して緩やかに遷移している．この理由は，表 4.1 のコーパスでは $C_{N\geq 4} \approx 1$ のためと推測される．

幅を持つ階数 ($N_{diff} = 1, 2$) で生成したワードサラダの多様性は，固定階数 N_H と N_L で生成したものの値に対して，N_L よりの中間位置にあることから，N_L の影響が強く出たと推測される．

実験 2 (検索エンジンによる識別性)

表 4.2: 検索エンジンによるコーパス検出確率
Table 4.2: Conditional Probabilities of Sentence to be Detected.

Order N	1	2	3	4	5	7	[1,2]	[1,3]	[2,3]	[2,4]	[3,4]	[3,5]	[4,5]
$P(W = t\|X = x)^{\dagger}$	12	19	44	78	85	89	6	9	25	33	56	58	75

†: If $N = 1$, then $x = S$. Otherwise, $x = H$.

表 4.2 に，検索エンジンによるワードサラダ単文ごとのコーパス検出率を示す．また，図 4.4 に，検索エンジンによる *Ham* と *Spam* の識別結果を示す．図中の縦線は，$\pm 1\sigma$ の幅を示す．

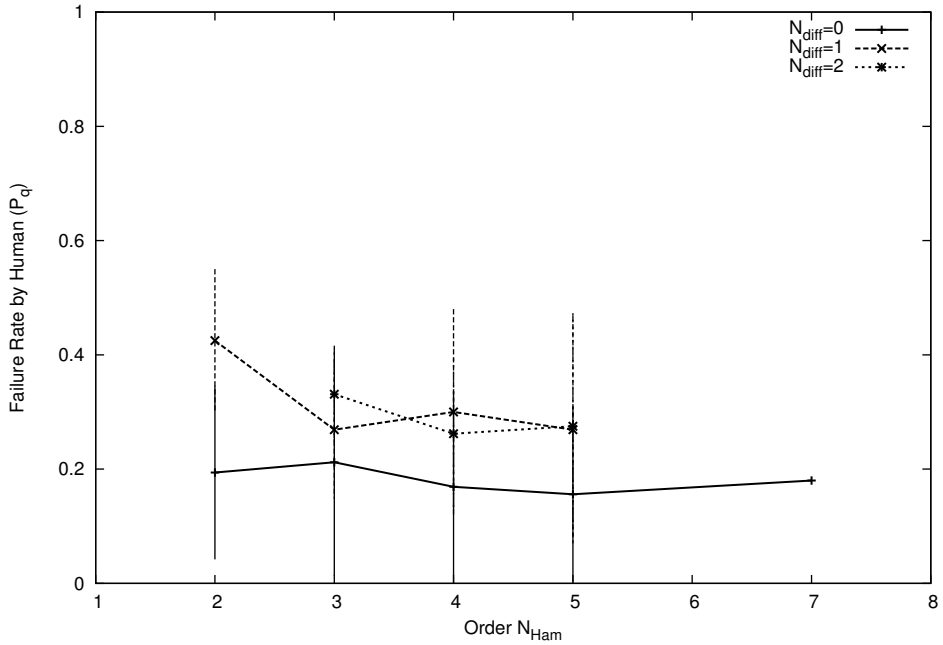


図 4.5: 人間による *Ham* と *Spam* の識別結果 (失敗率 P_q)
Fig. 4.5: Distinguishability Rate by Human.

固定階数 ($N_{diff} = 0$) のグラフから，ワードサラダ同士の識別は，自然文 ($N = 7$) とワードサラダの識別に比べて，検索エンジンには困難である．検索エンジンを用いた攻撃者は，*KK* 方式に相当する $N_{Ham} = 7$, $N_{Spam} = 1$ の 2 つを 86%で識別可能だが，$N_{Ham} = 2$, $N_{Spam} = 1$ になると 22% しか識別できない.

N_{Ham} に幅を持つ階数 ($N_{diff} = 1, 2$) と固定階数 (N_H, N_L) を用いた場合を比較する．検索エンジンによる $N_{Ham} = [N_L, N_H]$ と $N_{Spam} = 1$ のワードサラダとの識別結果は，N_L よりの中間位置にある．実験 1 の結果と同様に，N_L の影響が強く出たと推測される.

実験 3 (人間による評価)

図 4.5[1]に，主観実験により取得した，人間の CAPTCHA 1 問あたりの失敗率 P_q を示す. 図中の縦線は，$\pm 1\sigma$ の幅を示す.

固定階数 ($N_{diff} = 0$) のグラフから，*KK* 方式 ($N_{Ham} = 7$) の結果 18% に対して，提案方式の結果は 15.6–21.2% であり，階数に依存せずほぼ一定の P_q となることがわかった．ま

[1] $N_{Ham} = 7$ のデータは，鴨志田らの結果から導出したものであるが，平均値のみをプロットしている．また解答方式は，単文ごとに *Ham* と *Spam* を識別するものである.

た，幅を持つ階数 ($N_{diff} = 1, 2$) では固定階数に比べて若干高い P_q が出ているが，$N_L = 1$ の場合を除き，階数の変化に対してはあまり影響が見られなかった．

著者らは，人間による識別結果も実験 1 や 2 と同様に，階数変動の影響を強く受けると予想していた．実験 3 の結果は，人間による文の自然さの認識力が，想像以上であることを示している．

4.6 考察

4.6.1 *KK* 方式の脆弱性

KK 方式について，鴨志田らの検討したランダム推測攻撃とワード攻撃 [30] に加えて，本章で示した，生成文の多様性や検索エンジンを用いた攻撃についての安全性を検討する．

ランダム推測攻撃

h, s の値を知る攻撃者によるランダム推測の攻撃成功率 P_{mr} は，次のように示される．

$$\begin{aligned} P_{mr} &= P(Y = S, X = S) + P(Y = H, X = H) \\ &= P(Y = S)P(X = S) + P(Y = H)P(X = H) \\ &= \left(\frac{s}{z}\right)^2 + \left(\frac{h}{z}\right)^2 \end{aligned}$$

他の攻撃については， $s = 5, h = 15$ を例に挙げて計算方法を示す．

ワード攻撃

ワード攻撃成功率 P_{mw} の計算方法を示す．問題文 X が MS-WORD 2007 による文章校正を受ける事象を $W = t_w$，校正を受けない事象を $W = f_w$ とすれば，[30] より $P(W = t_w|X = S) = 0.24$, $P(W = t_w|X = H) = 0$ となる．$P(X = S) = 0.25$, $P(X = H) = 0.75$ なので，式 (2.3) より，$P(W = t_w) = 0.06$, $P(W = f_w) = 0.94$ となる．この分類器では，文章校正がされた場合には必ず *Spam* と解答する．そうでなければ式 (2.6) より，分類器は $P(X = H|W = f_w) = 0.798$ の確率で *Ham*，$P(X = S|W = f_w) = 0.202$ の確率で *Spam* と解答する．式 (2.7), (2.8) の議論より，式 (2.10), (2.11) において $P(Y_w = H, X = H) = 0.798$, $P(Y_w = S, X = S) = 0.394$ となる．よって，式 (2.9) から $P_{mw} = 0.697$ となる．

生成文の多様性の差を用いた攻撃

この攻撃は，過去に出題された問題文を収集し，新たに提示された問題が過去に出題されたものと一致するかどうかを調べ，その結果を攻撃に利用する．*Ham* と *Spam* の多様性が異なる場合に，有効な攻撃手段である．

生成文の多様性の差を用いた攻撃の成功率 P_{md} の計算方法を示す．問題文 X が過去に出題されたものと一致する事象を $W = t_d$，一致しない事象を $W = f_d$ とすれば，実験 1 の $N = 1,7$ の結果より $P(W = t_d|X = S) = 0$, $P(W = t_d|X = H) = 0.563$ となる．この分類器では，$W = t_d$ であれば必ず *Ham* と解答する．そうでなければ式 (2.6) より，$P(X = H|W = f_d) = 0.567$ の確率で *Ham* を，$P(X = S|W = f_d) = 0.433$ の確率で *Spam* と解答する．以降はワード攻撃と同じ議論により，$P_{md} = 0.716$ が計算される．

検索エンジンを用いた攻撃

この攻撃は，*Ham* と *Spam* の間に存在する検索結果の違いを利用する．

検索エンジンを用いた攻撃の成功率 P_{ms} の計算方法を示す．問題文 X が検索エンジンによりコーパスを特定された事象を $W = t_s$，特定されない事象を $W = f_s$ とする．表 4.2 にから $N = 1$ のデータを *Spam*，$N = 7$ のデータを *Ham* とすれば，$P(W = t_s|X = S) = 0.12$, $P(W = t_s|X = H) = 0.89$ となる．この分類器では，$W = t_s$ であれば式 (2.4) より $P(X = H|W = t_s) = 0.957$ の確率で *Ham* を，$P(X = S|W = t_s) = 0.043$ の確率で *Spam* と解答する．そうでなければ，式 (2.6) より $P(X = H|W = f_s) = 0.273$ の確率で *Ham* を，$P(X = S|W = f_s) = 0.727$ の確率で *Spam* と解答する．以降はワード攻撃と同じ議論により，$P_{ms} = 0.823$ が計算される．

攻撃方式の比較

図 4.6 に，$P(X = S) = s/z$ について，*KK* 方式に対する各攻撃方式の成功率を示す．本章で検討した 2 つの攻撃手法は，鴨志田らの検討したワード攻撃より，高い性能を示している．

攻撃手法により，*Ham* と *Spam* のどちらの検出を得意とするかは異なる．そのため s, h の比率により攻撃成功確率は異なるが，それぞれの最低値は $(P_{mr}, P_{mw}, P_{md}, P_{ms}) = (0.500, 0.567, 0.683, 0.796)$ となることから，*KK* 方式は検索エンジンを用いた攻撃に最も脆弱である．

P_{md} については，使用するコーパスとそこから生成する文の数により，強い影響を受けることに注意を要する．例えば，同一コーパスからの作問数が増えたり，小さいコーパスを利用した場合は，P_{md} の値は上昇してしまう．

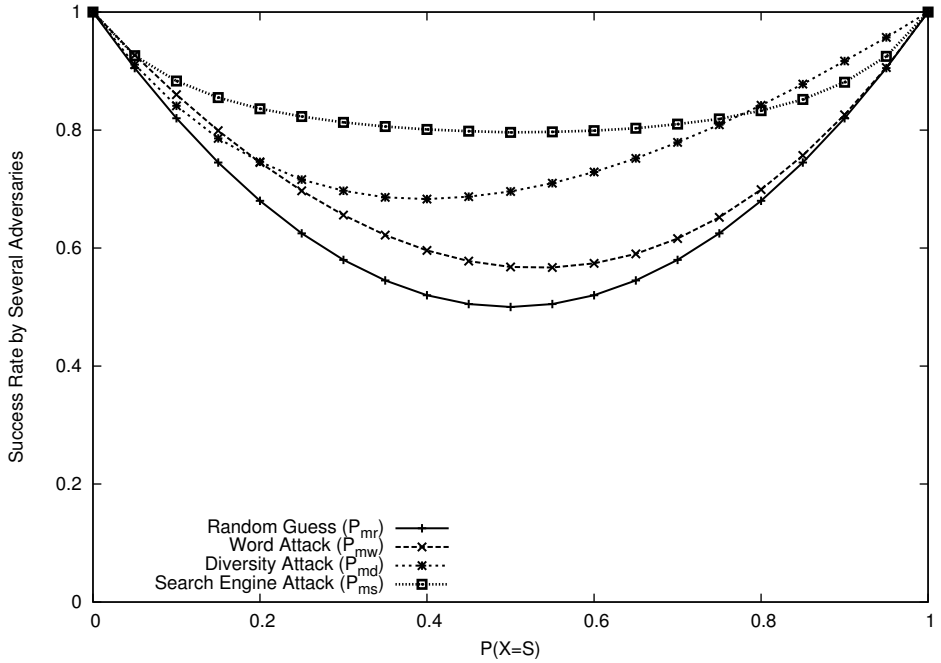


図 4.6: *KK* 方式に対する $P(X = S)$ ごとの攻撃成功率

Fig. 4.6: Attack Success Rate Given Known Probability of *Spam* $P(X = S)$ in *KK* scheme.

4.6.2 最適な N_{Ham} の決定と既存方式との比較

実験 1，2 の結果と 節 4.6.1 の検討より，検索エンジンを用いた攻撃が最も高い成功率を示すことから，この結果を提案方式と *KK* 方式の *FRR* として用いる．図 4.7 に，表 4.2 と節 4.6.1 で示した計算法から，CAPTCHA 1 問あたりの検索エンジンを用いた攻撃の成功率を示す．このグラフでは，図 4.5 の結果より，最適な N_{Ham} の候補となる固定階数のデータのみを示している．なお，同様の理由により，以後の検討は固定階数の場合のみで行う．

図 4.5 と図 4.7 から得た *FRR* と *FAR* より，総合的な指標として，式 (2.2) に示される *F*-値を用いて，提案方式と *KK* 方式を比較する．他の既存方式については，論文ごとに提示された *FRR* と *FAR* から *F*-値を計算する．

表 4.3 に，既存方式と提案方式の比較結果[2]を示す．表 4.3 のアクセシブルな方式とは，特定知覚に依存せず，問題の認識と解答ができるものと定める．テキストを提示する形式の CAPTCHA は，聴覚障害者は視認により，視覚障害者は彼らの使用する一般的な補助ソフトであるスクリーンリーダにより，問題への対処ができる．商用方式については，ア

[2]商用方式の *FRR* は，*FAR* と比べて調査時期が古いことに注意を要する．特に音声型 CAPTCHA は，その難化が指摘されている [27, 26, 119] ため，これらの *FRR* が悪化している可能性がある．

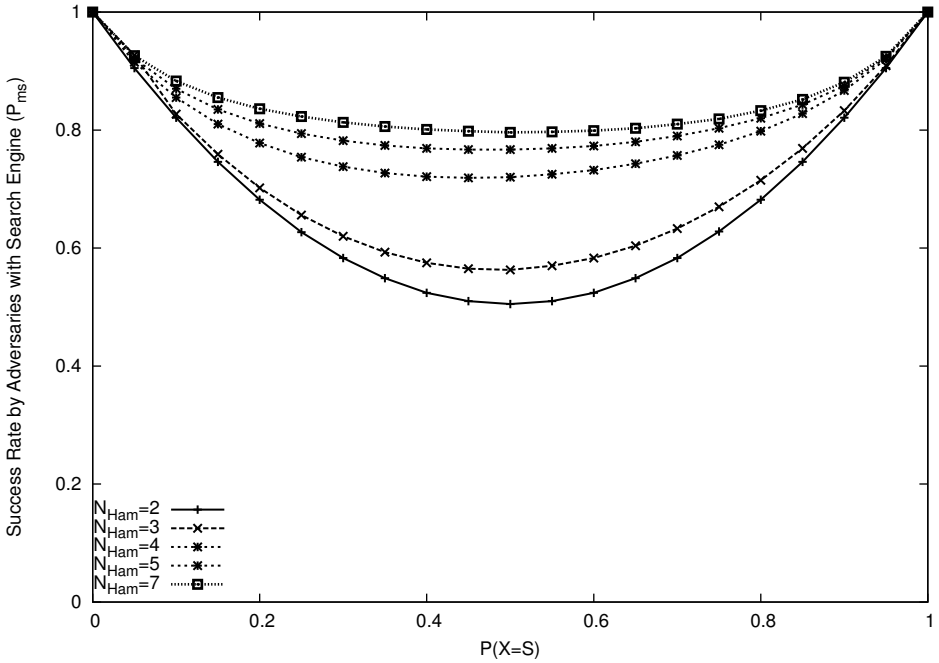


図 4.7: 提案方式に対する $P(X = S)$ ごとの検索エンジンを用いた攻撃成功率
Fig. 4.7: Attack Success Rate Given Known Probability of *Spam* $P(X = S)$ in our Proposal.

クセシビリティ性を満たさないため，参考データとしての扱いに留める．

表 4.3 の結果から，提案方式は $N_{Ham} = 2$ の場合に　F-値 0.61 で最適となる．アクセシブルな方式としては，提案方式が最もよい F-値を示した．

4.6.3 Gap Amplification

CAPTCHA 1 問あたりの *FRR* と *FAR* において，$FAR > FRR$ が成立する場合は，Gap Amplification [3] により安全性を強化できる．

提案方式において，節 4.6.2 で導出した最適条件 $H_{Ham} = 2, H_{Spam} = 1$ の問題を，20 問出題する場合を考える．式 (2.1) における (z, P_q, P_m) はそれぞれ (20, 0.194, 0.505) となることから，正答数の閾値 θ をパラメータとした CAPTCHA 20 問あたりの *FRR* と *FAR* は，二項分布より図 4.8 のように計算できる．$FRR = FAR$ となる *ERR* (Equal Error Rate) の条件において，提案方式は $\theta = 6, 7$ の付近で， *FRR* と *FAR* をともに約 10%に改善できる．

表 4.3: CAPTCHA 1 問あたりにおける既存方式と提案方式の比較
Table 4.3: Comparison between Conventional Schemes and our Proposal.

Scheme	N_{Ham}	FRR	FAR	F-ratio	Accessible?
Visual-eBay[†1]	—	0.07	0.514	0.64	No[†5]
Visual-Yahoo[†1]	—	0.12	0.053	0.91	No[†5]
Visual-reCapcha[†1]	—	0.25	0.223	0.76	No[†5]
Audio-eBay[†2]	—	0.37	0.829	0.27	No[†6]
Audio-Yahoo[†2]	—	0.32	0.455	0.61	No[†6]
Audio-reCapcha[†3]	—	0.53	0.586	0.44	No[†6]
[34][†4]	7	0.00	0.796	0.35	Yes
KK [30]	7	0.180	0.796	0.33	Yes
Our Proposal	2	0.194	0.505	0.61	Yes
	3	0.212	0.563	0.56	
	4	0.169	0.720	0.42	
	5	0.156	0.767	0.37	

†1: The values of FRR and FAR are referred from [25] and [52], respectively.
†2: The values of FRR and FAR are referred from [25] and [8], respectively.
†3: The values of *FRR* and *FAR* are referred from [25] and [10], respectively. Note that our study in 2013 showed *FRR* of Audio-reCAPTCHA was 100% regarding 5 Japanese with visual impairment.
†4: The value of FRR is referred from the results of informal experiments described in [34].
†5: For the hearing impaired.
†6: For the visually impaired.

4.6.4 コーパスの違いによる生成文の多様性

図 4.9 に，異なるコーパスを用いた場合の生成文の多様性を示す．各コーパスの特徴は，異なり語数については図 4.11 に，コーパスの多様性 (C_N) については図 4.10 に示す．実験では，文章は各々が異なる特徴を持つことを想定し，意図的に特徴の異なるコーパスを用いた．図 4.9 は，コーパス 0–4 ごとに，N 階ワードサラダを 50,000 個づつ生成した時の生成文の多様性の分布である．なお，節 4.5 で用いたコーパスは，コーパス 0–4 を 1 つにまとめたものである．

図 4.9 から，階数 N に対する生成文の多様性は，コーパスの異なり語数と C_N の影響を受けることがわかった．特に C_N の影響は，N が増加するほど顕著に表れた．したがって，生成文の多様性を確保するためには，分量の大きいコーパスの採用や，小さい N 階ワードサラダの利用が効果的である．また，異なり語数と C_N を調整したコーパスの選択をすることで，ある程度は生成文の多様性を制御できると推測する．

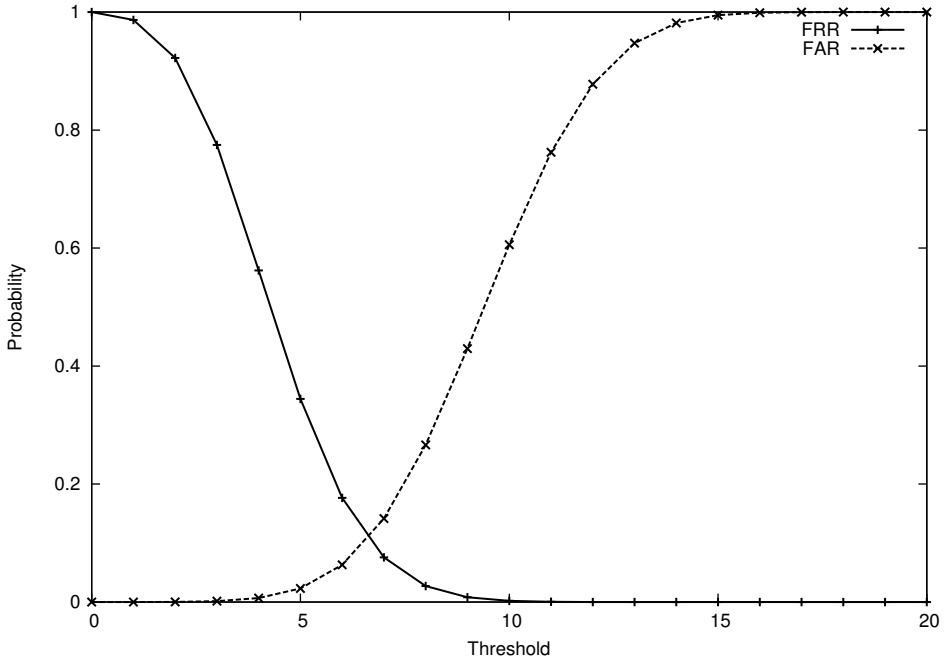


図 4.8: 提案方式 ($N_{Ham} = 2, N_{S\,pam} = 1$) における FAR と FRR の分布
Fig. 4.8: Probability Distributions of FAR and FRR of Our Proposal ($N_{Ham} = 2, N_{S\,pam} = 1$).

4.6.5 失敗率 P_q, $1 - P_m$ の分布

図 4.12 に，提案方式における CAPTCHA 1 問あたりの人間と機械の失敗率 P_q, $1 - P_{ms}$ の度数分布を示す.

図 4.5 の結果から，P_q と N_{Ham} への依存性が小さいので，人間の失敗率 P_q については $N_{diff} = 0$ となる 40 問を 1 つの群として扱う．ロボットの失敗率には，最も強力な検索攻撃を代表例とし，$N_{Ham} = 2$ における $1 - P_{ms}$ を用いる．ロボットの失敗率は，コーパスより生成した 640 問と節 4.6.1 に示した算出法から，16 問ごとの検索結果をもとに計算した 40 個のデータを用いる.

P_q は，(平均値, 標準偏差) = $(0.183, 0.106)$ となる分布であった．$1 - P_{ms}$ は，(平均値, 標準偏差) = $(0.485, 0.018)$ となる分布であった．提案方式の P_q と $1 - P_{ms}$ の分布は，重なりが少なく，かつその中央値同士が離れているため，CAPTCHA としての機能が期待できる.

P_q と $1 - P_{ms}$ は，平均値と中央値がほぼ一致した山なりの分布である．したがって，P_q や P_{ms} の平均値を代表値とした複数回試行への二項分布による近似 (節 4.6.3) は，一定の信頼性があると考える.

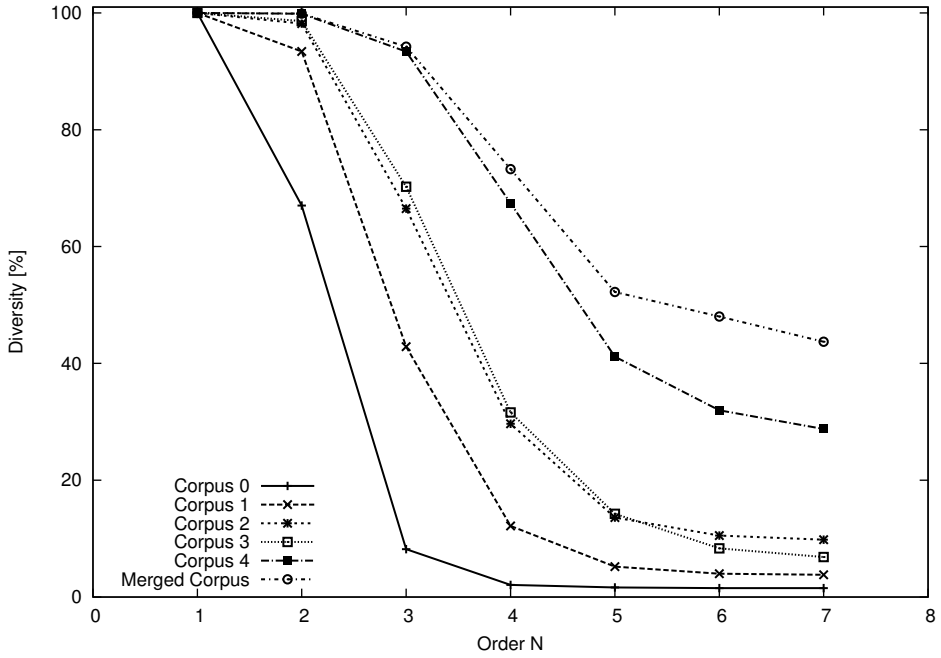


図 4.9: 異なるコーパスから生成された N 階ワードサラダの多様性 ($N_{diff} = 0$)
Fig. 4.9: Diversity of Sentences Generated by Different Corpora ($N_{diff} = 0$).

4.6.6 課題: 利用者の応答時間

図 4.13[3]に，実験 3 の正答率が 16 人中 8 番目であった被験者の応答時間を示す．図中の縦線は，$\pm 1\sigma$ の幅を示す．被験者数を増やした信頼性の確保は，今後の課題とする．

ワードサラダ識別型 CAPTCHA は，既存の方式に比べ応答時間が長いことが知られている．提案方式の問題 20 問を用いて CAPTCHA を構成した場合，認証には 200 秒程度の応答時間が見込まれる．一方で，商用 CAPTCHA の平均応答時間は，画像型で 11 秒，音声型で 43 秒程度である [25]．注意を要する点として，*FRR* が高い方式は，認証を受けるまでに複数の試行が必要になることがある．特に音声型 CAPTCHA は，この理由により，認証までの応答時間が増大しやすい．

KK 方式 [30] と比べて提案方式は，1 問あたり 2 つのワードサラダを読む必要があるので応答時間が長い．特に $N_{Ham} = 2$ の場合に顕著であるが，これは実験 3 の順番として $N_{Ham} = 2$ を最初に行った影響も考えられる．$N_{Ham} = 2$ の後半 5 問に限れば，応答時間の平均は 10.8 秒であり，$N_{Ham} = 3, 4, 5$ の場合と同程度である．

以上の検討から，提案方式の応答時間の改善は，重要な課題であると考えられる．対策としては，ワードサラダの文字数を削減し，利用者の読む文章量を削減する方法がある．

[3] $N_{Ham} = 7$ のデータは [30] から取得し，平均値のみをプロットしている．

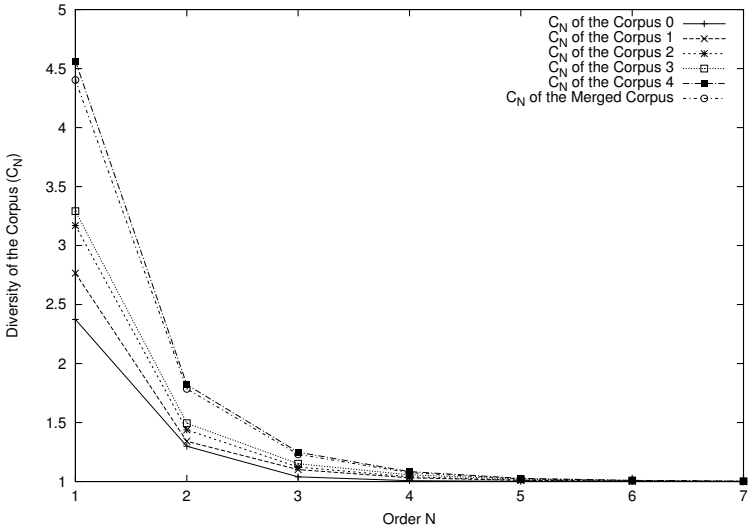


図 4.10: コーパスの多様性
Fig. 4.10: Diversity of our Corpora.

4.7 まとめ

本章では，まず既存の言語型 CAPTCHA として *KK* 方式 [30] を例に挙げ，その脆弱性を示した．自然文を用いたワードサラダ識別型 CAPTCHA の安全性の問題を指摘し，*KK* 方式が自然文の収集困難性と検索エンジンを用いた攻撃に対し脆弱であることを実験的に示した．

提案方式は，階数の異なるマルコフ連鎖モデルから生成された 2 種類のワードサラダ間に存在する「文の自然さ」の差を CAPTCHA に利用する．提案方式では，ワードサラダのみを用いることで，自然文に起因する脆弱性の問題を解決した．また，2 種類のワードサラダの比較結果を解答する方式により，人間による正答率の低下を抑制した．本章では，提案方式と *KK* 方式を *F*-値を指標として比較し，その優位性を示した．

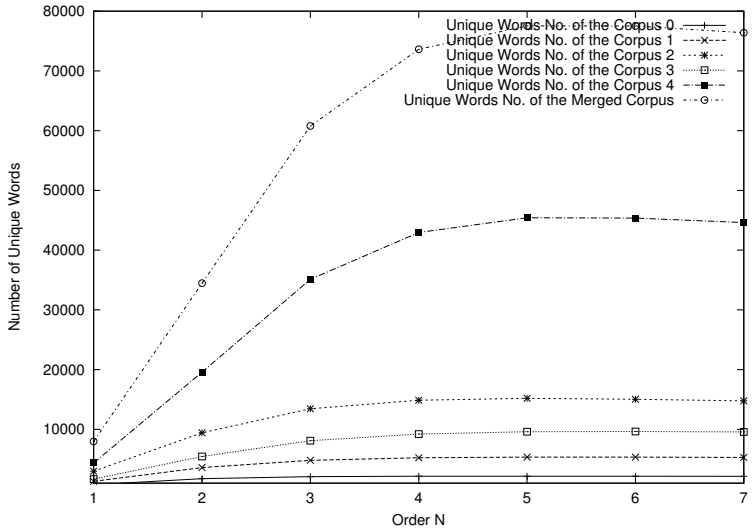


図 4.11: コーパスの異なり語数の多様性

Fig. 4.11: Number of Unique Words of our Corpora.

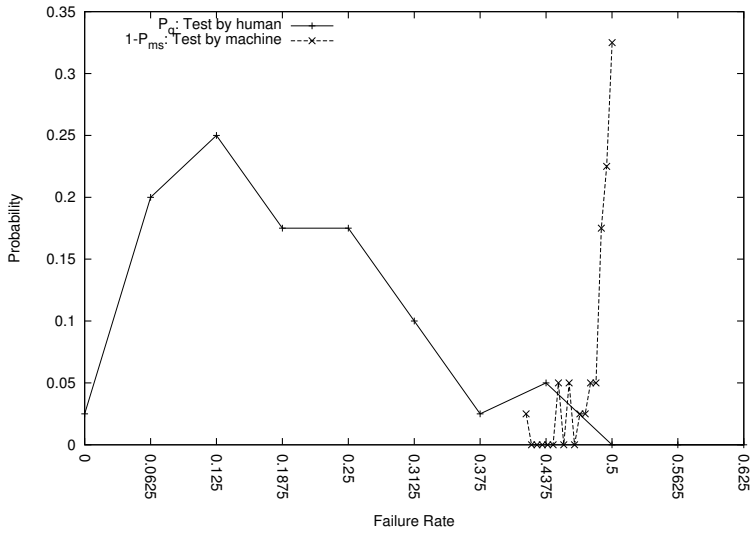


図 4.12: 失敗率 (P_q, $1 - P_{ms}$) の度数分布

Fig. 4.12: Frequency Distribution of Failure Rate (P_q, $1 - P_{ms}$).

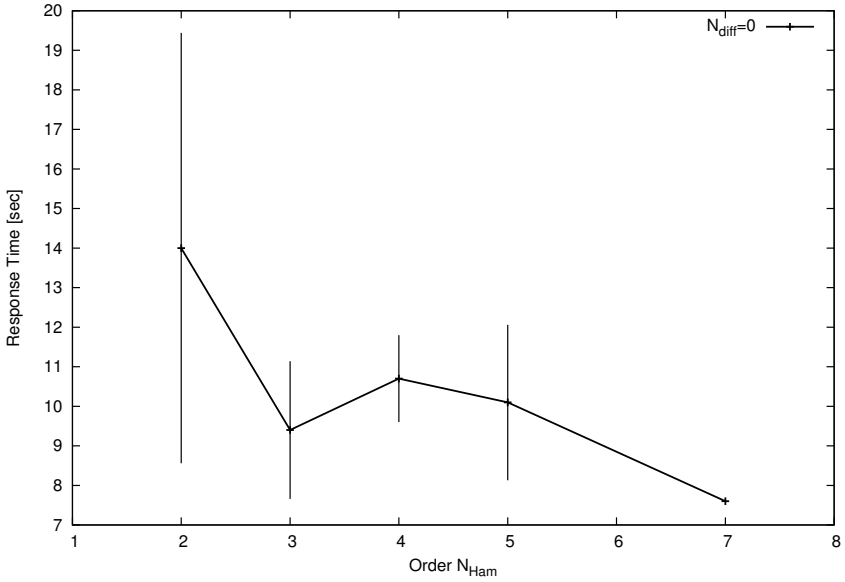


図 4.13: 1 問あたりの利用者の応答時間
Fig. 4.13: Response Time per Question.

第5章　子音交替を用いたインターネット上の文章の採取加工技法

5.1　導入

これまで述べたように，自然文を問題文として提示する言語型 CAPTCHA では，自然文の持つ脆弱性が問題になる．4 章で提案した相対比較問題は，人間の「文の自然さ」の識別能力を利用する [30, 34] などの方式には有用だが，[33] のような文の話題を問う方式には使用できない．

この問題の解決方法として，本章では，インターネット上の文章の採取加工技法について検討する．具体的には，取得した自然文に対して，子音交替を用いた文字置換を施すことで，検索攻撃に対抗する．子音交替とは，語の置かれた環境によって子音が変化するという方言などでもみられる現象である．人間には比較的馴染みのある規則なので，子音交替後の文でも人間の文意文脈解釈能力が働くと期待できる．

本章では，提案する作問技法として，子音交替による文字列置換やコーパスからの自然文の抽出方法についてアルゴリズムを示す．また，提案方式を具体的に 3 種類の異なる文意文脈解釈問題に適用する．さらに実験によって，提案方式のユーザビリティや安全性を評価し，提案方式に適した文意文脈解釈問題の特徴を示す．

5.2　提案する作問技法

5.2.1　言語型 CAPTCHA の基本構成

提案するテストの作成と実行は，大別すると下記の手順をたどる．

1. **作問要求の受付:** 利用者が認証などの手続きで，CAPTCHA が課せられるべき状況に到達すると，作問プログラムが起動する．
2. **素材文章の収集:** 作問プログラムはインターネットに接続し，コーパスから問題文の素材となる文章を収集する．
3. **作問:** 取得した文章を元に，複数の文を作成する．このうち特定の文だけに，「内容の自然さ」や「話題分野の一致」といった文意解釈上の特徴が含まれているようにする．

4. **出題**: 利用者に作成した問題文を提示し，特定の特徴を持った文を選ぶようにと指示する．問題は文字情報として提示されれば十分なので，視覚ディスプレイでも，点字ディスプレイ，音声読み上げのいずれの手段にも対応できる．

5. **解答の確認**: 解答結果がシステムが期待する答えと一致すれば，認証に合格したとする．

5.2.2 コーパスの選定

文意文脈解釈問題の作成には，素材となる文章を必要とする．コーパスの選択と，そこから素材文章を獲得する方法としては，次のものが考えられる．

1. **インターネット から文章を収集する方法**: Twitter やブログなどで公開された文章集合をコーパスとし，そこから素材文章を収集する方法である．多様な文章を大量に得られる利点がある．反面，それは公開情報であるから，攻撃者も同じ文章を検索によって収集できるという欠点がある．それゆえコーパスと同じ文章を，改変せずにそのまま回答選択肢に使う方式 [30, 34] には特に不向きである．

2. **秘匿されたデータベース内の文章を利用する方法**: 出題者だけが知り得る文章集合をコーパスとし，そこから文をランダムに選び素材文章とする方法である．一般に，問題文の秘匿のみを安全性の根拠にするのは好ましくない．また，秘匿されたコーパスであっても，一度問題として使えば，そこで公開されてしまうので，秘匿が成り立たない．攻撃者がテストシステムへの出題要求を大量に繰り返すことで，いつかは既知で解答を分析済みの問題文に遭遇できてしまう欠点がある．

3. **利用者に文章を作成させ入力させる方法**: テストの途中に，利用者に文章を作成させ入力させる段取りを持つ方法である．これにより，新規な文書がテストシステムに補給され，別の機会に素材として使用できる．しかし，攻撃者も文章生成に参加することになり，攻撃者に都合の良い素材文章を登録できてしまう欠点がある．

さらに方式 2 と 3 では，素材文章を方式 1 ほどに大量に低コストで蓄積できるわけではないので，数に制限なく作問することにも適さない．よって本研究では，数に制限のない新規作問を実現するために方式 1 を採用する．

5.2.3 文章からの文字列抽出

本研究では，インターネット上の文章の豊富な分量により，制限のない新規作問を狙う．しかしながら，素材文章を常に元の文単位で利用すると，作問に使用出来る分量が文の数に制限される．さらに，常に文頭から利用する単純な方式を取ると，実際に利用可能な分量がさらに制限される．

本研究では，この問題を回避するために文字列抽出関数を用いる．この関数は，複数文からなる文章，切り出す文字数の範囲 $[\ell_{min}, \ell_{max}]$，切り出す文字列の数 j を入力とし，指定された範囲の長さの j 個の文字列 str を出力する．

次に処理内容を示す．文字列の先頭形態素を自立語にする理由は，なるべく自然な形で抽出し，ワードサラダと識別可能にするためである．

1. 入力された文章を形態素解析し，その形態素の配列 ary を取得する．配列の最後に終端記号を追加する．
2. $i = 0, \mathcal{SS} \leftarrow \emptyset$ とし，次の処理を行う．
   (a) $ary[i]$ が自立語でなければ，$i \leftarrow i+1$ とし，処理 2-a に進む．$ary[i]$ が自立語であれば，処理 2-b に進む．
   (b) 総文字列長が ℓ_{min} 以上となる配列要素までの形態素を結合し，文字列 ss を得る．終端記号を読みだした場合は，処理 3 に進む．
   (c) $|ss| \leq \ell_{max}$ であれば，$\mathcal{SS} \leftarrow \mathcal{SS} \cup \{ss\}$ とする．$i \leftarrow i+1$ とし，処理 2-a に進む．
3. $(str_0, \ldots, str_{j-1}) \xleftarrow{\$} \mathcal{SS}$ を出力する．

5.2.4 子音交替による文字置換

言語型 CAPTCHA が抱えるコーパスの特定という問題を解決するため，問題文として表示する文字列の子音を改変する．

これは，方言などに見られる単語の子音の違いを指す．例えば，ザ行からダ行への子音交替では,「ざぶとん」を「だぶとん」と改変する．漢字に対しては，MECAB [38] や CABOCHA [120] の「読み」情報を利用し仮名に変換[1]してから，子音交替処理をする．

次の例は，図 5.2 のワードサラダ選択問題 1 に対して，子音交替処理を施したものである．

1. ナナジスジニユウヒョクヲタベタ.
2. モヒハレタナラユウエロチニコマッタ！
3. オガネガタリズホロガガエナイ.
4. ヤクタイノナイコトヲガロガエテイルト,

本手法はロボットへの対抗策として期待できる．改変率が一定以上ならば，Web 検索によって改変した文章から元のコーパスを特定するのは困難であると仮定する．漢字を仮名に書き換え，なおかつ子音を改変された文を検索クエリとして，改変前の文字列と類似す

[1]変換精度は，これらのツールの性能に依存する

るものを，インターネット上の全文章という非常に大きい文章集合から見つけ出すことは，「あいまい検索」などの技術をもってしても，いまだ難しいと期待できるからである．

子音交替によって生成された文は，人間にとっては聞き間違いや書き間違いの 1 種として感じられる．また，方言などにみられる現象であるため，子音交替による変化は，比較的慣れ親しんだものである．したがって，文字改変の割合がある程度小さければ，人間は元の文章と同様の意味を理解できると期待できる．これらの確認は，5.4 節と 5.5 節で行う．

子音交替の手順の詳細を説明する．言語 L に含まれるすべての行の種類を $\mathcal{U}_L$ とし，文字列 *str* に含まれるすべての行の種類を $\mathcal{U}_s$ とする．本章では L を日本語とするので，行とは，ア行，カ行，$\cdots$ を指す．$u(\in \mathcal{U}_s)$ 行から v 行への子音交替を i 箇所適用することを $f(u, v, i)$ と表す．*str* に対して $f(u, v, i)$ を j 回適用する子音交替関数は，$\mathcal{F} := \{f(u_g, v_g, i_g)\}_{g\in[1,j]}$ を用いて，$\mathbf{F}^{\mathcal{F}}(str, j)$ と表す．ただし，$\exists g, g'$ に対して $g \neq g'$ ならば $u_g \neq u_{g'}$ である．子音交替関数は，次のように動作する．(i, j) の値は，ロボットへの頑強性と人間による理解のしやすさを考慮して決める必要がある．

1. *str* より $\mathcal{U}_s$ を計算する．
2. u, v を次のように選択する．

$$u \overset{\$}{\leftarrow} \mathcal{U}_s, \quad v \overset{\$}{\leftarrow} \mathcal{U}_L \backslash u, \quad \mathcal{U}_s \leftarrow \mathcal{U}_s \backslash u$$

3. *str* 中に含まれる u 行の文字数を超えない範囲で， i を決定する．
4. $f(u, v, i)$ を *str* に適用する．
5. 2–4 の処理を， j 回繰り返す．

5.3 提案する作問技法を適用した文意文脈解釈問題の作成

5.2 節に示した作問技法に基づき，文意文脈解釈問題を構成する．本章では 1 種類の話題を識別するテストと，2 種類の文の自然さを識別するテストを示す．

子音交替は，適用前の文に比べると適用後の文は不自然になるため，文の自然さに関する影響が大きいと推測できる．このため，問題の正答率が文の自然さに影響されにくい方式として，話題の識別問題への適用を検討する．また，文の自然さとは明確な定義が難しいため，テスト方式によって子音交替の影響が異なることも考えられる．よって，文の自然さの識別問題に対しては，2 種類の方式を検討する．

共通話題識別テスト

共通話題識別テストとは，共通する話題の文脈に現れる複数の文を利用者に提示し，共通話題が何であるかを答えさせるテストである．

- 国務をきちんとこなして欲しい。
- 参院態勢でみんな混乱　日本はどうなる？
- 資本主義社会では、経済理想をもって、
- 健康診断に費用がかかる。

選択肢: (1) 政治、(2) 健康、(3) 経済、(4) 天気 **正解: (1)**

- 読み物としてもおもしろい農学書
- 正史にその創作は含まれていない。
- 都市の変遷が伺える
- 経験がその人の年輪として刻まれている

選択肢: (1) 文学、(2) 社会、(3) 歴史、(4) 農業 **正解: (3)**

- 財務会計論が難しい …
- 売り手が投機筋しか見つからない
- 趣味にローンするのは反対
- 相手を探し、試合を開始しなさい。

選択肢: (1) 運動、(2) 探検、(3) 教育、(4) 金融 **正解: (4)**

図 5.1: 共通話題識別テスト作成例（子音交替適用前）

Fig. 5.1: Examples of Semantic Cognition Test regarding Common Topic without Consonant Gradation.

ワードサラダ文選択問題 1

1. 七時過ぎに夕食を食べた。
2. もし晴れたなら遊園地に困った!
3. お金が足りず本が買えない。
4. 益体のないことを考えていると、

正解: (2)

ワードサラダ文選択問題 2

1. また、子牛は旅行とは速くならなければ
2. バイクを下取りに出すときの気持ち
3. 休日に一人で楽しんでもいいじゃないか
4. 短距離が得意でもマラソンどうだろうか?

正解: (1)

ワードサラダ文選択問題 3

1. 酒やジュースなど飲み物に入れるため
2. 駅舎とホームは構内踏切で連絡している。
3. 定休日と食べるの子どもは旅行と言えるの
4. 治療の指標になる。この pH の測定は

正解: (3)

図 5.2: ワードサラダ文識別テスト作問例（子音交替適用前）

Fig. 5.2: Examples of Semantic Cognition Test by Differentiating Word Salad Sentences from Natural Sentences without Consonant Gradation.

機械翻訳文選択問題 1

1. 経済が 2 パーセント成長すると予測します。
2. ビールのボトルケースを買ったとき、
3. はじめまして！よろしくお願いします！
4. 少年と明日よく再び掛かっていること。

正解: (4)

機械翻訳文選択問題 2

1. 飲めば飲むほどにやめられない味だ。
2. 彼女は最も有名な家族の 1 つから来る。
3. 私は毎日マラソンをする習慣がある。
4. なんでもお気軽にご相談ください。

正解: (2)

機械翻訳文選択問題 3

1. 交互の調子が悪い宇宙から何かを出すんだ。
2. それから、正確に狙いをつけます。
3. 私は彼らを追いかけた。すると、
4. お腹がすいた、お昼になにを食べようか？

正解: (1)

図 5.3: 機械翻訳文識別テスト作成例（子音交替適用前）

Fig. 5.3: Examples of Semantic Cognition Test by Differentiating Machine-Translated Sentences from Natural Sentences without Consonant Gradation.

図 5.1 に作問例を，以下に共通話題識別テストのアルゴリズムを示す.

共通話題識別テストのアルゴリズムの概要

1. 利用者に提示する 1 問あたり文の数として，共通話題に属する数 c と，共通話題と異なる偽の話題に属する数 d を入力する．制約条件として，$c \geq d$ を要する．本章では $d = 1$ とする．

2. 共通話題を表すキーワードの候補となる集合 $\mathcal{K}$ を入力する．$\mathcal{K}$ は一般的な知識に属するものが望ましい．制約条件として，$|\mathcal{K}| \geq 2$ を要する．

3. 1 つの作問ごとに次の処理を行う．

   (a) $\mathcal{K}$ から正答となるキーワード *cKey* と偽のキーワード *dKey* を選択する．

   (b) *cKey* と *dKey* のシソーラスとして $c\hat{K}ey$ と $d\hat{K}ey$ を取得し，$c\hat{K}ey$ と $d\hat{K}ey$ は含むが *cKey* と *dKey* は含まない文章をコーパスから取得する．

   (c) 取得した文章に文字列抽出関数を用いて，c 個の共通話題に属する文と d 個の偽の話題に属する文を取得し，それぞれに対して子音交替による文字置換を行う．なお，文字列抽出関数で切り出した文字列については，その中に $c\hat{K}ey$ や $d\hat{K}ey$ が含まれていなくてもよい．

   (d) 以上のように取得した文を問題文とし，$\mathcal{K}$ を選択肢とする．

4. 作問結果をランダムな順番で利用者に提示する．利用者は，作問ごとに提示された文からその大勢を占める話題を選択する．

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．

ワードサラダ文識別テスト

ワードサラダ文識別テストでは，*KK* 方式と同様に自然文とワードサラダを利用者に提示し，それらを識別させるテスト方式である．

図 5.2 に作問例を，以下にワードサラダ文識別テストのアルゴリズムを示す.

ワードサラダ文識別テストのアルゴリズムの概要

1. 利用者に提示する 1 問あたり文の数として，自然文の個数 h とワードサラダの個数 s を入力する．制約条件として，$h \neq s$ AND ($h = 1$ OR $s = 1$) を要する．本章では $s = 1$ とする．

2. コーパスから N 階マルコフ連鎖モデルを作る．

3. 1 つの作問ごとに次の処理を行う．

(a) コーパスから取得した文章に文字列抽出関数を適用して，自然文として h 個の文字列を取得する．

(b) マルコフ連鎖モデルからワードサラダとして s 個の文字列を取得する．

(c) 取得した文字列に対して子音交替による文字置換を行う．

4. 作問結果をランダムな順番で利用者に提示する．利用者は，作問ごとに提示された文からワードサラダを選択する．

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．

機械翻訳文識別テスト

機械翻訳文識別テストとは，[34] の方式と同様に，自然文と機械翻訳文を利用者に提示し，それらを識別させるテスト方式である．ただし，機械翻訳文の生成方法は [34] と異なり，ある言語の文章を 1 回だけ機械翻訳したものを用いる．

図 5.3 に作問例を，以下に機械翻訳文識別テストのアルゴリズムを示す．

機械翻訳文識別テストのアルゴリズムの概要

1. 利用者に提示する 1 問あたり文の数として，自然文の個数 h と機械翻訳文の個数 s を入力する．制約条件として，$h \neq s$ AND ($h = 1$ OR $s = 1$) を要する．本章では $s = 1$ とする．

2. 2 つの言語 L_{NS} と L_{MT} を選択する．本章では L_{NS} を日本語，L_{MT} を英語とする．

3. 1 つの作問ごとに次の処理を行う．

   (a) L_{NS} コーパスから取得した文章に文字列抽出関数を適用して，自然文として h 個の文字列を取得する．

   (b) L_{MT} コーパスから取得した文章に機械翻訳を適用する．さらに文字列抽出関数を適用して，機械翻訳文として s 個の文字列を取得する．

   (c) 取得した文字列に対して子音交替による文字置換を行う．

4. 作問結果をランダムな順番で利用者に提示する．利用者は，作問ごとに提示された文から機械翻訳文を選択する．

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．

5.4 評価

5.4.1 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．

実験 1 生成文の多様性の評価

実験 2 検索エンジンを用いた子音交替に対する攻撃の評価

実験 3 人間による評価

5.4.2 実験方法

作問プログラムとその設定

各作問プログラムは，テスト方式ごとに分離して実装した．サーバへの組み込みはしておらず，単体動作する Windows Form プログラムである．通常の対話的な作問と解答の他に，実験用に作問結果をテキストファイルに直接出力することができる．

共通話題識別テストの作問プログラムへの入力データを次に示す．

共通話題識別テストの作問設定

- 1 問あたりの文の数: 5 ($c = 4, d = 1$)
- 1 文あたりの文字数: 30–80
- キーワードの集合: $\mathcal{K}$ = {“スポーツ”, “天気”, “経済”, “食事” }
- コーパス： 検索 API[2] で取得可能なインターネット上の文章

1 問あたりの文の数を 5 とした理由は，予備実験から得た経験に基づく．この値が大きいと，共通話題の情報を十分に含む作問が行えるが，文章量が増え利用者には使いづらい．値が小さすぎると，十分な情報を含めることができないため，利用者の正答率が落ちてしまう．

ワードサラダ文識別テストの作問プログラムへの入力データを次に示す．

ワードサラダ文識別テストの作問設定

- マルコフ連鎖モデルの階数: $N = 1$
- 1 文あたりの文字数： 40–80

[2]GAPI.Net 0.5.0.1(http://gapidotnet.codeplex.com/)

- コーパス： 青空文庫 [118]

機械翻訳文識別テストの作問プログラムへの入力データを次に示す.

機械翻訳文識別テストの作問設定

- 1 文あたりの文字数： 40–80
- 機械翻訳ソフトウエア: オンラインのソフトウエア (付録 A.2)
- 日本語コーパス： 日本語版 Wikipedia [121]
- 英語コーパス： 英語版 Wikipedia [122]

コーパスの特徴

作問プログラムがコーパスから収集した文章について，その特徴を示す．実験では，これらの文章を用いて作問を行う.

共通話題識別テスト 109,042 文字， 1,209 行，形態素 9,726 種

ワードサラダ文識別テスト 87,260 文字，1,717 行，形態素 4,361 種

機械翻訳文識別テスト 日本語 48,813 文字，1,267 行，形態素 4,019 種

英語 単語 1,990 種

実験 1 (多様性)

各テスト方式について 10,000 回の作問を繰り返し，共通話題識別テストでは 50,000 個の生成文を，その他のテスト方式ではそれぞれ 40,000 個の生成文を取得し，生成文の多様性を確認する.

実験 2 (検索エンジンによる攻撃)

子音交替を施した問題文のコーパスを特定する攻撃の評価として，まず次のような実験方法を検討した.

1. コーパスの漢字を仮名に書き換えた文章をウェブに公開し，既存の検索エンジンにそのページをインデックス化してもらう.
2. 子音交替を適用した文章を検索語として，既存の検索エンジンに問い合わせる.

3. 「あいまい検索」などの検索語の補正機能により，取得元のコーパスが特定されたかを確認する．評価指標は，2 で指定した文章のコーパスが，どのように選出されたかで判断することにする．本章では，(A) と (B) の場合を，検索によりコーパスが特定されたと定義する．

   (A) 第 1 候補に選出された．

   (B) 第 2–10 候補中に選出された．

   (C) 選出されない，または第 11 候補以降で選出された．

しかしながら，このように公開されたページに高いランクが付くとは考えにくい．ページランクが低ければ，コーパスが検索結果の上位に現れにくく，評価が甘くなるかもしれない．

問題となるのは，子音交替の際に漢字を仮名に書き換える点である．一方で，漢字表記を維持したまま子音交替を施すと，置換対象の文字が限定されてしまう．そこで，子音交替とは異なる文字列置換アルゴリズムによる次の実験方法を用いる．このアルゴリズムでは，文中の置換箇所については音交替の場合と同程度に制御できるため，代替案として有用であると判断した．

1. コーパスには青空文庫の文章を用いる．青空文庫を採用した理由は，有名なサイトであるため，検索結果の上位に現れやすいと推測したためである．

2. コーパスから文字列置換アルゴリズム用のモデルを作成する．モデルは，3 つの連続する文字に対して，中心位置の文字を値，その前後の文字の組を鍵とするハッシュテーブルとする．このモデルに鍵となる文字の組を入力すると，コーパスの文字出現頻度に従い，1 つの文字が出力される．

3. 5.2.3 節で示した方式に従い，コーパスから自然文を 100 個生成する．子音交替の代わりに，次の文字置換をすべての生成文に対して行う．この置換作業は，子音交替と同様に 1 文あたり 2–5 回行う．

   - 生成文中からランダムに i 番目の文字 c を選択する．
   - 手順 2 で生成したモデルに $(i-1, i+1)$ 番目の文字の組を入力し，出力として文字 $\hat{c}$ を得る．
   - 形態素 c を $\hat{c}$ で置換する．

4. Google の Web 検索を用いて，手順 2 で生成した文を人手で検索し，前述の A–C の指標に従い評価する．

実験 3 (人間による評価)

被験者として，日本語を母国語とする 12 人の全盲／弱視の視覚障害者らに参加して頂いた．本章では，その主要な情報取得経路により被験者を分類する．すなわち，スクリーンリーダなどの音声による情報取得者 9 人と，拡大鏡などを利用した視覚による情報取得者 3 人のグループに大別 [3] する．

被験者らへのテストの提示は，次のように行う．

- 素材文章の収集と作問は事前に行う．作問結果をテキストファイルに出力し，人為的なフィルタリングはせず，被験者にそのまま提示する．作問内容は，被験者全員に対して共通の 10 問とする．
- 子音交替の影響を検討するため，子音交替適用なし／ありに分けて，それぞれ 10 問ずつ作問する．子音交替は，問題文 1 行あたり 2–5 文字を改変対象とする．
- 解答方式を択一選択方式に統一し，選択肢数を 4 とする．

5.4.3 実験結果

表 5.1: 実験結果

Table 5.1: Results of our Experiments.

Test Type	Is Consonant Gradation Applied?	Accuracy [%]		
		$U_A + U_V$ †1	U_A	U_V
Common	No	69	72	60
Topic	Yes	72	77	57
Word Salad	No	90	92	84
Sentence	Yes	64	65	65
Machine-Translated	No	55	53	60
Sentence	Yes	28	28	30

†1: U_A and U_V represents users with audio and visual inferface, respectively.

実験 1 (多様性)

生成文の多様性は，共通話題識別テストが 99.65%，ワードサラダ文識別問題が 99.94%，機械翻訳文識別問題が 99.59% であった．

[3] 点字ディスプレイの利用者も数名いたが，全員が音声情報の補助的利用にとどまるため，今回は別グループには分類しない．

実験 2 (検索エンジンによる攻撃)

検索によるコーパス特定の実験結果は， (A) 18 個，(B) 4 個，(C) 78 個であった．よって，文字列置換によって検索によるコーパス特定攻撃を抑制する効果は 78% である．小説中の人名などの特徴を持つ文が，特定されやすい傾向にあった．

実験 3 (人間による評価)

実験結果を表 5.1 に示す．表 5.1 の “Consonant Gradation” は，子音交替を意味する．“Accuracy” は，各テスト 10 問に対する被験者らの平均正答確率である．

共通話題識別テスト

作問結果の一部を図 5.4 に示す.

付録 A.1 で示した検定方式では，共通話題識別テストは，子音交替の有無による有意差はない．いずれの場合でも本実験の被験者らは約 70% の確率で正答できていた．

ただしこのテストでは，図 5.4 に示す通り，極端に正答率の低下する作問がされる場合があった．この理由は，正答となる話題の情報が期待通りに自動抽出できず，被験者が撹乱用の偽の話題に引きずられるためだと推測される．

ワードサラダ文識別テスト

作問結果の一部を図 5.5 に示す.

ワードサラダ文識別テストは，付録 A.1 で示した検定方式では，子音交替の有無による有意差が出た．子音交替適用後の正答率は，64% にとどまる．

機械翻訳文識別テスト

作問結果の一部を図 5.6 に示す.

機械翻訳文識別テストの子音交替適用後の正答率は，0% であった．よって，機械翻訳文識別テストは，CAPTCHA として機能していない．

5.5 考察

5.5.1 生成文の多様性

4 章の図 4.3 の実験結果と比べると，5.4 節の結果が示す生成文の多様性は高い値を示している．この理由は，5.2.3 節で述べた文字列抽出関数と，子音交替の効果が考えられる．

子音交替なしの例（正答率 92%）

- 税制課へのお問い合わせ。メールは専用フォームからの送信となり
- 30 日今日は雪、雨、強風で天候が悪く、山荘で天候回復を待つお客様
- 団体等監査（平成 22 年度事業対象）の結果は次のとおりです。なお
- リフティングチャージは安上がり？について。海外に住んでいる家族
- 2 月 15 日もうかりまっかあきまへんな ” なんていう会話がいつ頃

共通話題の選択肢: (1) スポーツ、(2) 天気、(3) 経済、(4) 食事
正解: (3)

子音交替なしの例（正答率 0%）

- 男女同権論を唱えて 150 年、なかなか日本社会の男性優位は崩れ
- 皆様にお届け致します。第２５回例会兼新年会開催日。平成２５年
- 【一等米】日本晴 10Kg 玄米 3、500 円全国の食味基準は滋賀
- 政策として民間金融機関では対応が困難な分野に対して財政融資、
- しております。udck_002071_1。jpg。おはようござい

共通話題の選択肢: (1) スポーツ、(2) 天気、(3) 経済、(4) 食事
正解: (2)

子音交替ありの例（正答率 92%）

-)」や「au けーたいちゃくしんわりびき (ちゅう 3)」のごりようでしゃいんかんのつうしんちようの
- いすが 100 こもしゅっぴんされてはいないとおもうのせ、あくまでもすうじじょうのはなしですが
- _05_30_01。JPG。りえきをあげるにはどうしたらよいでしょうか？じちゅはりえきをあげることは、それほど
- いっちするぺーじ : 7 けんのあいてむがひっとしました。1 ぺーじめをひょうじしています。Thereisenjoi ![みかぐぎゃ
- こうむいんをくしゃのかんり、くにのしゅっしやせいふほゆうかぶしきのばいきゃくとう

共通話題の選択肢: (1) スポーツ、(2) 天気、(3) 経済、(4) 食事
正解: (3)

子音交替ありの例（正答率 42%）

- ねん 7 ぢゅき 24 にちたにやまびろこさんの「こくもつのあめがふる」れす。(´ д ` ) かんそうのえどうしよっかなー…。( ´ д `)
- 2013/7/15(つしぃ) ごぜん 11 : 45・きんきょうぼんとはのっとこい、いやいや、こいびんくいろえとってもきれいです ˆ
- しょうかいしておりまく。ひ、き、どようひぃにしかたべられないようびげんていのげきうまていしょくです。はんつくのめだまやき
- 5 つき 12 にちきのうはひさしきゅみにまとまっやあめでしたね。あめののちはくうきがすんで、けしきがどりくっき
- だんけつしてたたかっていまぬ。ぎぃなさんそときのたいいくたいかいうんどうかいはどんなものだった

共通話題の選択肢: (1) スポーツ、(2) 天気、(3) 経済、(4) 食事

正解: (2)

図 5.4: 実験で使用した共通話題識別テストの作問例
Fig. 5.4: Examples of Semantic Cognition Test regarding Common Topic.

子音交替なしの例（正答率 100%）

1. 雨だった。その雨を部屋の窓から見ながら、邦子は、すべてをほぼあきらめた。そして、
2. 西野だということに、やがて気づいた。美代子はすでに西野から離れて久しく、三枝子も
3. 言うよりも、なんだろう、いちばん若い仲間という感じ。私がいっしょにいて当然なのに
4. 直視したのをうちに砂の持ちなさいとボールには化粧してとしいて着ているのかの脚見えた名前なのか

ワードサラダ文を選択せよ。 (1)–(4)
正解: (4)

子音交替ありの例（正答率 75%）

1. あゆいていくばめんできゃ、はいごにいちだいのびあのによるむちょうのおんがくがものしづか
2. ちょくしぼを、うだれたにみれたまつからあいこだった。どてていっ。ぺんろみえた。そして
3. ぎゃべりかただけでぶけど、なんちぇすてきなひとだろうと、あのときわたしはおもったの
4. あいだで、かれじゃすそしだけまよった。それとれをてみとったかれは、どちらをもとおやま

ワードサラダ文を選択せよ。 (1)–(4)
正解: (2)

子音交替ありの例（正答率 42%）

1. 41 ごうびぇんひゅだいどころ。さよこ、ははおやげてきたがたこ、いいをぼくあがっていた
2. もとにもどり、げんひゃぎ　はかんざきけいこだった。かのこはいつものみずぎをきていた
3. それぷつづけてかた、ようすけはえりこにあいずした。たかいふらいをあよほうこうへ、
4. いちぶぶんとして、ちょうひょりをおよぐちとたちやすいちゅうえあろびぐすのためのぷーる

ワードサラダ文を選択せよ。 (1)–(4)

正解: (1)

図 5.5: 実験で使用したワードサラダ文識別テストの作問例

Fig. 5.5: Examples of Semantic Cognition Test by Differentiating Word Salad Sentences from Natural Sentences.

子音交替なしの例（正答率 33%）

1. 倉橋啓太郎 - 寺島進 （第 1 話、第 9 話・第 11 話回想）
2. 小都羽総合病院皮膚科医師。モジャモジャ頭のアフロヘアーと体型から、
3. 病院初診受付（クラーク）。MR の山崎に買収されやすい。人事異動で、売店配属に。
4. 今日、神聖な名前僧院の 16 人の姉妹が聖ベネディクトのルールによって生きています。

機械翻訳文を選択せよ。 (1)–(4)

正解: (4)

子音交替ありの例（正答率 8%）

1. この「そうぜい」は、ぶんぢぃん（しょうせつか・じじん・かじん・はいじん・ちょさくか
2. ぢょうせつ『さんごくしえんぎ』でもとうじょうするが、ここではむちょうなしょうぐん
3. とやまみさお・もりまつとしおへんちょ『てじぃこぶりくぐんへんせいそうらん』
4. ぎりしゃ・ろーじゃのべんきのじしょおよびしんわ、1075

機械翻訳文を選択せよ。 (1)–(4)

正解: (1)

図 5.6: 実験で使用した機械翻訳文識別テストの作成例

Fig. 5.6: Examples of Semantic Cognition Test by Differentiating Machine-Translated Sentences from Natural Sentences.

文字列抽出関数は，収集した文章から問題文中の 1 行に相当する文字列の切り出しを，文頭以外の複数箇所から実施している．さらにこの関数は，段落ごとの文章を入力とするので，切り出した結果が複数文にまたがることもある．よって，単純な文の数以上に，新規文を生成できる．また，仮に同じ文字列が切り出された場合でも，子音交替による文字の置換がおこなわれることで，同一文字列が問題文として提示される可能性は低くなる．

5.5.2 コーパス特定の困難性

5.4.3 節の実験結果では，文字列置換を施した後の文に対しても，検索エンジンは 22% の確率でコーパスを特定している．この結果は，4 章の表 4.2 と比較すると，2–3 階ワードサラダの値に相当している．これは，検索エンジンによるコーパスの特定を，文字列置換処理が妨害していることを示す.

文字列置換に子音交替を利用した場合には，攻撃者は漢字を仮名に書き換えた文章をデータベース化する必要がある．新規文章が常時生成される Twitter などをコーパスに利用すれば，攻撃者は常にデータベースの更新が必要になり，完全なデータベースの構築は困難になると考えられる．また，このような仮名文は，漢字仮名混じり文に比べ使用される文字種が少ないため，あいまいで特徴の少ない文になり，形態素解析を利用した検索語の修復も困難になると考えられる．したがって，子音交替を適用した場合は，本代替案で

の実験と比較して，良い結果が得られると考えている．

5.5.3　人間の正答率

ワードサラダ文識別テストや機械翻訳文識別テストでは，子音交替による正答率の低下が顕著である．この理由は，自然文とワードサラダや機械翻訳文の間にある「文の自然さ」の差が，子音交替の影響で小さくなったためだと考えられる．

機械翻訳文識別テストの正答率は特に悪い．この改善には，山本ら [34] のように複数回の機械翻訳の適用，精度を意図的に落とした機械翻訳の使用，より機械翻訳が難しいと期待できる口語的文章の利用が考えられる．

共通話題識別テストは，子音交替の影響が少ないため，他の方式に比べて高い正答率を示している．その理由は，共通話題識別テストは文の自然さを問う方式ではないため，文が示す意味が失われない限り正答率への影響は軽微であるためだと考えられる．したがって，子音交替は，共通話題識別問題などの文の自然さの変化に頑強なテスト方式に適している．

5.6　まとめ

本章では，言語型 CAPTCHA において，インターネット上の文章を利用した作問の際の採取加工技法について検討した．提案方式では，子音交替による文字置換により，ロボットによるコーパスの特定を妨害し，自然文の安全な利用を可能にする．

また，提案方式を 3 種類の文意文脈解釈問題に組み込み，被験者の協力を得て実験を行った．実験結果から，話題の識別テストなど，問題文中の文の自然さが正答率への影響が小さい方式に提案方式が適していることを明らかにした．

第6章　聴覚型 CAPTCHA の提案: 多様な話者を模擬して発話された単語とランダムな音列の識別問題

6.1　導入

聴覚型 CAPTCHA は，主に視覚障害者のウェブアクセシビリティの確保を目的として研究されている．聴覚型 CAPTCHA は，言語型 CAPTCHA と比べると，聴覚障害者の利用は困難だが，言語の習熟はさほど必要ないという特徴を持つ[1]．よって，聴覚型 CAPTCHA は，言語型 CAPTCHA と異なるアクセシビリティを持つ方式であるため，その研究の意義は大きい．たとえば，言語の習熟が十分でない幼年層や言語に関する学習障害を持つ視覚障害者らに，聴覚型 CAPTCHA は適している．

本章では，意味論的雑音を用いた既存方式として，Meutzner らの提案した *MGK* 方式 [37] を分析する．*MGK* 方式は，単語の発話に RPS を混合して ASR に対抗する．しかしながら，単語部分に関しては難聴化がなされていないため，その部分に対する ASR の高い正答率を攻撃者が利用する恐れがある．本章では，実験により，以上の攻撃者を想定した場合に *MGK* 方式の安全性が危殆化することを明らかにする．

提案方式については、その特徴とアルゴリズムを示す．提案方式では，多様な話者による発話を模擬することで適切な音響モデルの作成を困難にして，単語部分についても ASR による認識を妨害する．さらに実験により，ASR を用いた攻撃についての提案方式の安全性と人間によるユーザビリティを評価する．

6.2　準備

6.2.1　*Ham* と *Spam*

本章で述べる聴覚型 CAPTCHA では，複数の「音の塊」を無音区間を介して結合した音声ファイルをテストに用いる．この「音の塊」は，発話された単語もしくは RPS である．

[1]鴨志田ら [30] は，言語の習熟度としてネイティブ，学習済み／学習中の非ネイティブ，未学習の非ネイティブらによる *KK* 方式に対する正答率を比較している．言語の習熟度が高いほど，*KK* 方式の作問に対する正答率も高い．

本章で「単語」とは，ある言語において意味を持つ言葉の最小単位を表す．RPS は，音韻の組み合わせをモデル化したマルコフ連鎖により合成された音韻列を表す [2]．ただし，単語と同じ音を持つ RPS は除外する．

本章では，*Ham* を単語とし，*Spam* を RPS とする．

6.2.2 音声認識処理

聴覚型 CAPTCHA に対する攻撃方法としては，ASR を用いることが考えられる．本章では，既存方式や提案方式の安全性を検討する際に，ASR の代表例として，音声認識の分野で広く利用されている HTK (Hidden Markov Model Toolkit [123]) を用いる．HTK では，HMM (Hidden Markov Model) [124, 125] を音響モデルとして使用する．

音声認識処理は，その語彙数や適用される仕事 (タスク) の制限により，孤立単語認識と大語彙連続音声認識に大別される．聴覚型 CAPTCHA で出題されるテストはその多くが単語や数字の認識であるため，本論文では孤立単語認識の概略を示す．

音響モデル

単語辞書を Ψ とし，その要素を $\psi_i \in \Psi$ とする．ここで単語 $\psi_j \in \Psi$ の発話を考える．発話により得られた音声波形をある期間ごとに分割し，それぞれの特徴ベクトル O を式 (6.1) のように得る．ここで o_θ とは，ある期間 θ に対応する特徴ベクトルである．特徴ベクトルには，MFCC (Mel-Frequency Cepstrum Coefficients) [103] などが使用される．

$$O = o_1, o_2, \ldots, o_\theta, \ldots, o_\Theta \tag{6.1}$$

孤立単語認識とは，式 (6.2) となる ψ_i を求めることであり，$\psi_i = \psi_j$ ならば認識は成功となる．事後確率 $P(\psi_i|O)$ を直接計算することは，認識する必要のある特徴ベクトル O とそのラベルをすべて集める必要があるため，現実には難しい．そのため，ベイズの定理より，式 (6.3) を得る．式 (6.3) では，i の変化に無関係な項である $P(O)$ を最後に削除している．

$$\arg\max_i \{P(\psi_i|O)\} \tag{6.2}$$

$$\arg\max_i \left\{ \frac{P(O|\psi_i)P(\psi_i)}{P(O)} \right\} = \arg\max_i \{P(O|\psi_i)P(\psi_i)\} \tag{6.3}$$

[2] Meutzner らの論文 [37] では RPS の生成方法を明示していないため，本章で示した生成方法と異なる可能性がある点に注意を要する．

式 (6.3) において，事前確率 $P(\psi_i)$ は単語辞書サイズやその出現率から計算できる [3]．よって，尤度 $P(O|\psi_i)$ を計算すれば，式 (6.3) を求めることができる．音声認識処理では，尤度を直接的に求めるのは難しいので，この部分をある確率密度関数に置き換えて計算する．HTK では，単語 ψ_i に対応する HMM Ξ_i を用いて，$P(O|\psi_i) = P(O|\Xi_i)$ と仮定して 式 (6.3) を求める．

HMM では，観測できない状態 $X = x(1), x(2), \ldots, x(\Theta)$ を順に遷移しながら，観測できる音韻や無音などに相当する特徴ベクトル O を出力する．HMM は，隠れ状態 X，状態 $x(i)$ から $x(j)$ の遷移確率 $\mu_{x(i)y(j)}$，それぞれの期間 θ で状態 $x(\theta)$ に遷移した際に o_θ を出力する確率密度 $\nu_{x(\theta)}(o_\theta)$ によって，その特徴が定義される．

HMM に基づく確率密度関数 $P(O|\Xi)$ は，式 (6.4) のように表される．ただし，$x(0)$ と $x(\theta+1)$ は，それぞれ開始状態と終了状態を表す．

$$P(O|\Xi) = \sum_X \mu_{x(0)x(1)} \prod_{\theta=1}^{\Theta} \nu_{x(\theta)}(o_\theta)\mu_{x(\theta)x(\theta+1)} \tag{6.4}$$

式 (6.4) を単語 ψ_i ごとに作成したものが $P(O|\Xi_i)$ となる．また，単語辞書に登録されたすべての要素に対する HMM をまとめて，音響モデルと称する．

音響モデルの訓練と認識処理

音響モデルの訓練には，音声波形から抽出された特徴ベクタと対応するラベルを組とした訓練データを用いる．訓練データによって，単語ごとの HMM である Ξ_i を音響モデルとして作成する．音声波形を音韻ごとに分割する処理や HMM の訓練には，k 平均法 [126] や Baum-Welch [124] が使用される．

認識処理では，単語辞書，単語の出現ルール (文法)，音響モデルを用いる．認識対象となる発話の特徴量 $\hat{O}$ に対して，すべての単語ごとにその尤度と事前確率を計算し，単語辞書から音韻列に対する最も確からしい単語を出力する．

HTK などの既存の ASR は，認識した音韻列を言語として意味を持つかどうかの判定をせず，最も確からしい単語を単語辞書から選択して出力する．このような ASR の音声認識アルゴリズムは，不明瞭な発話に対する認識精度を向上させる効果はあるが，発話の中に単語と同様の特徴量を持つ RPS が加わると，それを単語として認識してしまう．

6.3 既存方式とその問題点

2.2.6 節で示したように，統計的雑音を用いた既存の聴覚型 CAPTCHA は脆弱である．

[3] 本論文では，出現率を一様であるとし，$P(\psi_i) = 1/|\Psi|$ とする．

MGK 方式の特徴

佐野ら [10] は，聴覚型 CAPTCHA を頑強にする方法として次のガイドラインを示した．

ガイドライン 1 統計的雑音は ASR の認識を妨害する効果はほとんどないので，意味論的雑音を使用する．

ガイドライン 2 信号と雑音の auditory power を同等にする．

Meutzner の提案した *MGK* 方式 [37] は，このガイドラインに従い構成された方式である．*MGK* 方式のアルゴリズムを次に示す．また，*MGK* 方式の概念図を図 6.1 に示す．

MGK 方式のアルゴリズム

1. 単語辞書と発話辞書から，使用する単語と，それに含まれる音韻を抽出する．
2. 抽出した音韻をランダムに組み合わせて RPS を作成する．
3. h 個の単語と s 個の RPS をランダムな順に並び替え，無音区間で結合した 1 つの音声ファイルを利用者に与える．
4. 利用者は h 個の単語についてのみ，聞き取った内容を文字列で入力する．
5. すべての単語を正しく解答したならば利用者を受理，そうでなければ拒否する．ただし，各単語の解答については，その編集距離 (Levenshtein distance) が 1 以内のものは正答として扱う．

MGK 方式では，単語と同じ音韻で構成されるが言葉として意味を持たない RPS を発話に混ぜることで，6.2.2 節で示したように，既存の ASR が RPS を雑音として無視できない点を利用している．

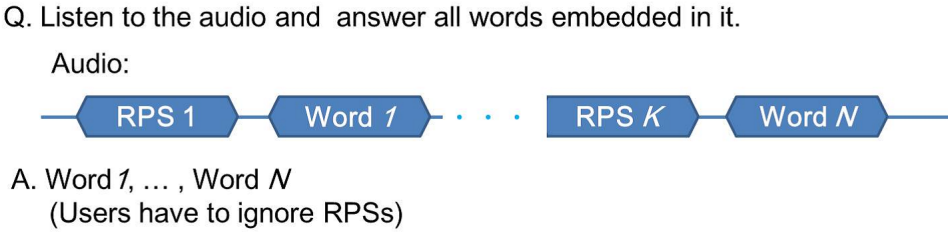
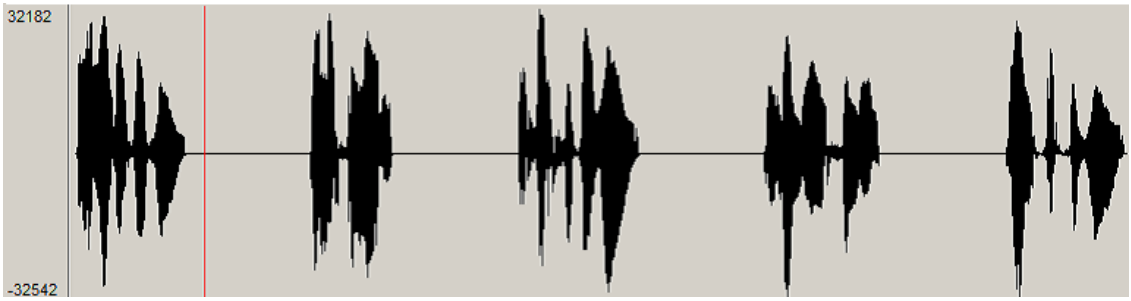


図 6.1: *MGK* 方式
Fig. 6.1: *MGK*-scheme.



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 6.2: 提案方式 1 の音声ファイルの波形例
Fig. 6.2: Example of a Waveform regarding our Proposal 1.

MGK 方式の問題点

Meutzner らは，生成した音声ファイルごとの ASR の認識結果から *MGK* 方式の安全性を評価した．その結果では，ASR が RPS を単語と誤認識することによる *MGK* の頑強さを示したが，単語部分を ASR が正しく認識することによる危殆化の問題については議論がされていない．*MGK* 方式では，単語に相当する音声区間は難聴化がされていないため，この検討は必要であると考える．*MGK* 方式の脆弱性の詳細については，6.5 節において，提案方式と安全性を比較して示す．

既存方式である統計的雑音の挿入や音声の重畳による単語区間の難聴化は，適用が難しいことに注意を要する．統計的雑音はガイドライン 1 により使用できないし，ガイドライン 2 を満たしながら音声の重畳をすると人間の正答率が落ちてしまう．我々の事前調査においても，*MGK* 方式の無音区間を削除し音声同士を重畳した場合，人間の正答率は明らかに落ちていた．

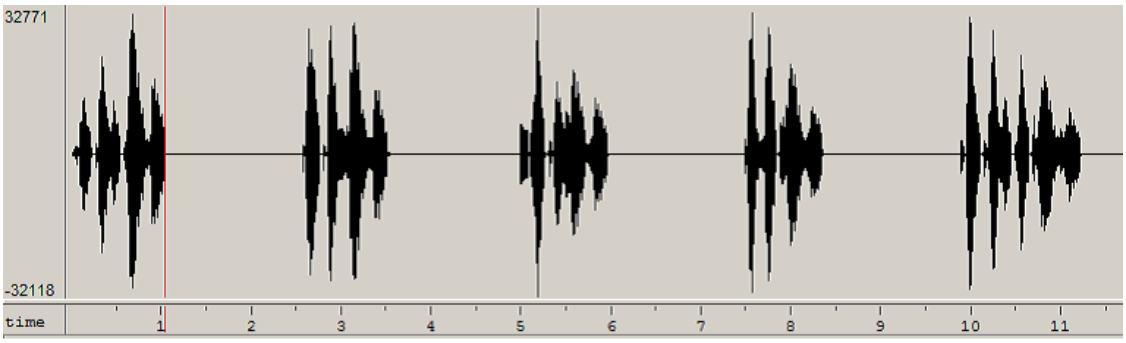
また，人間に複数単語の記憶を促す点も *MGK* 方式の問題点である．事前調査にて *MGK* 方式を体験した被験者からは，この点に関する不満が出ている．

6.4 提案方式

6.4.1 提案方式の概要

MGK 方式と提案方式の違いは，(1) 多様な話者を模擬した合成音の利用，(2) 単語と RPS の識別問題の利用，(3) 解答方式 である．

(1) 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．



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 6.3: 提案方式 2 の音声ファイルの波形例
Fig. 6.3: Example of a Waveform regarding our Proposal 2.

参考文献 [39] 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 [40] 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．一方で人間の場合は，日常的に色々な人と会話を行うため，多様な発話の認識能力は頑強であると期待できる．

(2) 単語と RPS の識別問題の利用: Darlene ら [127] 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，RPS 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，RPS と単語の識別が人間には容易であることを期待する．

(3) 解答方式: 提案方式では，1 問あたりの内容を簡略化する代わりに，複数問の解答を利用者に促す．

識別問題の場合，単語そのものを記憶する必要はなく，その認識も正確に行う必要がない．このため記憶作業負荷が低減し，解答の入力も簡略化されるため，ユーザビリティが向上する．

6.4.2 方式定義

提案方式のアルゴリズムを以下に示す．

表 6.1: 実験に用いた音声の方式ごとの特徴
Table 6.1: Features on Sounds Used in Experiments.

Scheme	Speed and Pitch	Speaker
MGK [37]	Fixed	Native (Japanese)
Our Proposal 1	Modified	Native (Japanese)
Our Proposal 2	Fixed	Non-native (Spanish)

提案方式のアルゴリズム

1. 単語辞書と発話辞書から，　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，本章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，各音声同士を適当な無音区間を挟んで結合する．音声ファイルを利用者に出力する．

4. 利用者は，出力し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 6.2 と図 6.3 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 6.2 は日本語話者の音声で，図 6.3 はスペイン語話者の音声である．

6.5 評価

6.5.1 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．

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

6.5.2 実験方法

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」 [128] 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 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 [37] 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 [4] を，外国語話者を使用する場合は eSpeak [5] を使用した．本実験では，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 6.1 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 として出力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，利用者に出力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．

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．

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

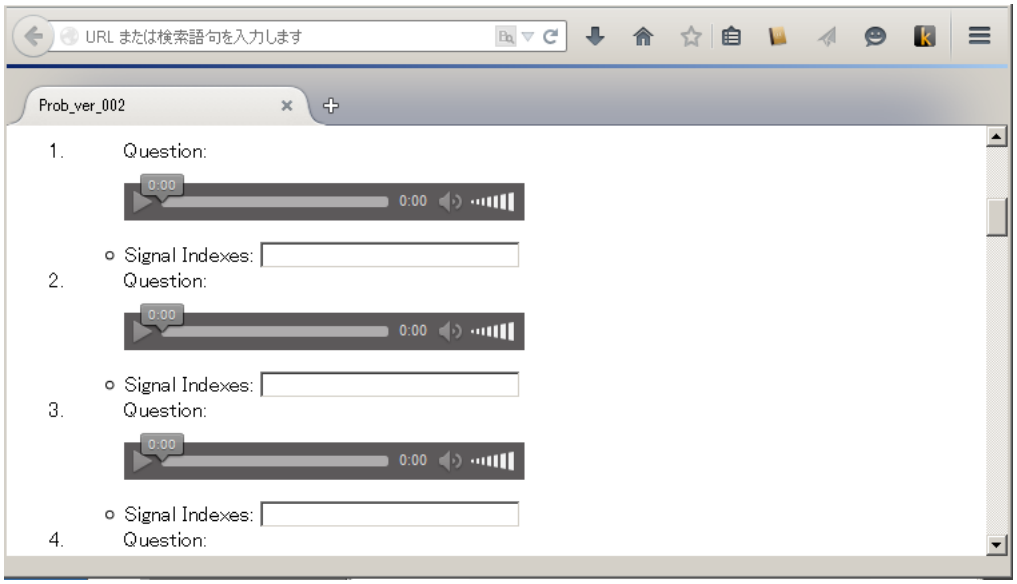


図 6.4: 提案方式の音声型 CAPTCHA 出題画面の例
Fig. 6.4: Appearance of our audio-CAPTCHA.

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．

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 6.4) 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出力する．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．

表 6.2: HTK による単語やランダムな音韻列の条件付き検出率 [%]
Table 6.2: 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 [37]	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

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．

実験 3

本章では，人間の要した解答時間を調査し，人間の正答率と合わせて，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．

6.5.3 実験結果

実験 1

表 6.2 に，HTK による単語や RPS の認識結果を示す.

すべての方式において，HTK は *Spam* を *Ham* と誤検出した．この理由は，RPS が単語と同じ音韻で構成されるためである．参考文献 [37] 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.

表 6.3: 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 6.3: 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

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* を最大で 73%の確率で検出した．

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．

実験 2

表 6.3 に，人間による単語と RPS 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

表 6.4: 被験者らの 1 音声ファイル †1 あたりの解答時間 [秒]
Table 6.4: 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.

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．

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 6.4 に示す．表 6.4 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を除いた．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 RPS の違いや，単語間に挟んだ無音期間の違いのためである．

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．この応答時間は，2.2.7 節で示した既存方式に対して 1/4 – –1/13 程度である．提案方式では利用者に複数問の解答を促すが，既存方式に比べて極端に応答時間が増加しないことを表している．

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．

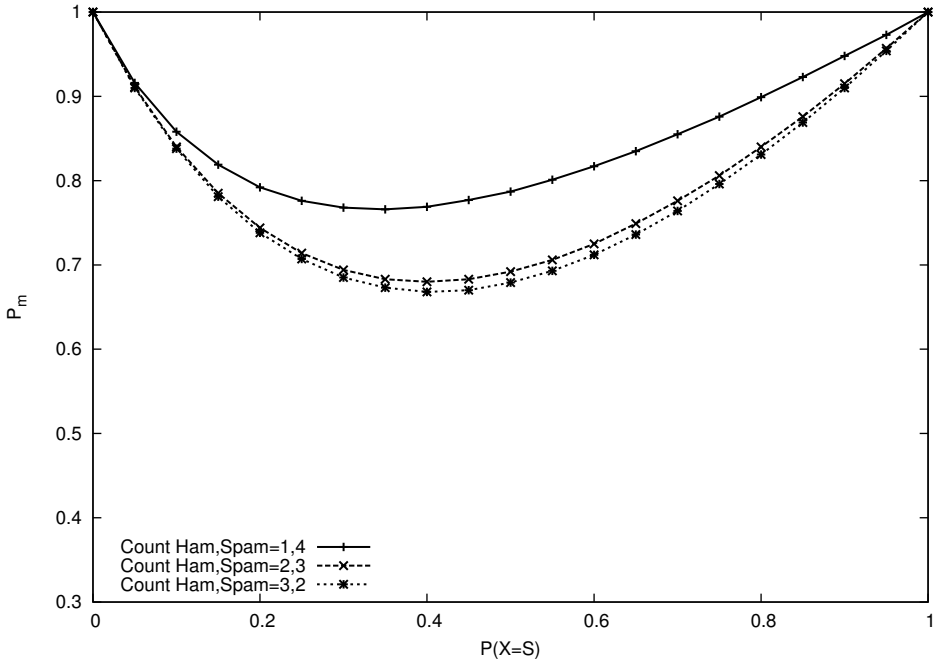


図 6.5: *MGK* 方式で生成された問題 1 問当たりの機械の攻撃成功率 (P_m)
Fig. 6.5: Machines' Success Rate for each Question (P_m) regarding *MGK*-scheme.

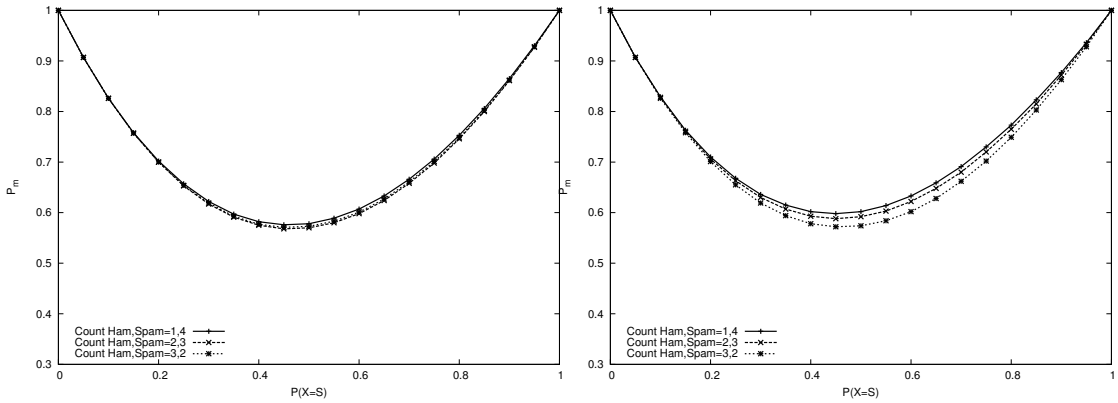
6.6 方式の比較

本節では，2.1.4 節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: 本章では，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，表 6.2 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 (2.9) 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 6.5 と図 6.6 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．



[1] Our Proposal 1. [2] Our Proposal 2.

図 6.6: 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. 6.6: Machines' Success Rate for each Question (P_m) regarding our Proposal 1 and 2.

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 6.3 の結果から，式 (2.8) を用いて P_h を計算する．図 6.7 に，問題中の *Spam* 含有率に対する P_h の変動を示す．

表 6.3 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．

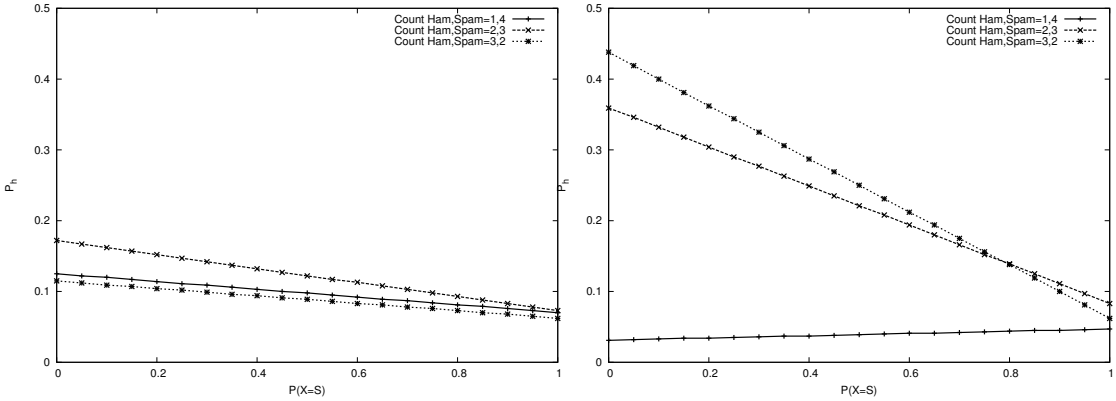
提案方式と既存方式の比較: 方式ごとに，P_m が最小となる (s,t) 条件を選択し，その *FRR* と *FAR* から式 (2.2) に従い *F* 値を計算する．*F* 値を用いた比較結果を，表 6.5 に示す．

表 6.5 [6]より，提案方式 1 が最もよい性能を示すことが分かる．

6.7 まとめ

本章では，ランダムな音韻列である RPS と単語の識別問題を利用した聴覚型 CAPTCHA を提案した．識別問題の利用により，既存の聴覚型 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，ロボットによる音声認識を用いた攻撃を困難にした．本章では，実験により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．

[6] 4 章で比較に使用した聴覚型 CAPTCHA の *FAR* は，reCAPTCHA のもの以外は本章を執筆した 2016 年の段階では弱い攻撃者による結果なので，ここでの比較には使用しない．



[1] Our Proposal 1. [3] Our Proposal 2.

図 6.7: 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. 6.7: Humans' Failure Rate for each Question (P_h) regarding our Proposal 1 and 2.

表 6.5: 既存方式と提案方式の比較

Table 6.5: Comparison between Conventional Schemes and our Proposal.

Scheme	N_{Ham}	FRR	FAR	F-ratio
MGK [37]†1	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 value of FRR is referred from [37]. 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.

第7章　結論

本論文では，視覚障害者向けの新しい言語型 CAPTCHA と聴覚型 CAPTCHA を提案した．提案方式は，聴覚により認知ができるため，視覚障害者にとってアクセシブルな方式である．

言語型 CAPTCHA の提案をするにあたり，既存方式の問題点を分析した．出題文の型が明確に定義された方式の脆弱性を指摘するとともに，具体例としてアメリカ政府系サイトで使用されていたクイズ CAPTCHA に対して，実装したソフトウエアにより 99% の確率で攻撃が成功することを示した．また，自然文を問題文に使用する方式は，自然文の収集困難性を狙ったデータベースマッチングによる攻撃や，検索エンジンを用いた攻撃に脆弱であることを示した．

提案した言語型 CAPTCHA では，複数階数のマルコフ連鎖モデルを用いて合成した「文の自然さ」の異なる機械合成文の相対比較問題によって，自然文の持つ脆弱性と認知バイアスの影響による人間の正答率低下を克服した．

また，既存の自然文を用いた言語型 CAPTCHA の安全性を向上する施策として，インターネット上の文章を利用した作問とその採取加工技術について検討した．提案方式では，子音交替の規則により問題文の文字置換を行うことで，検索エンジンによるコーパスの特定を妨害した．さらにこの方式が，文章の表す話題の識別テストに適していることを明らかにした．

聴覚型 CAPTCHA については，既存方式は白色雑音などの統計的雑音による難読化を行うのに対し，提案方式では，非母国語話者による発話や発話速度・ピッチの変動による多様な話者の模擬を利用した．本研究では，このような不特定話者に対する適切な音響モデルの作成は困難であるため，音声認識を用いた攻撃に頑強であることを実験により示した．さらに，解答方式をランダムな音韻列と単語の識別問題とすることで，既存の聴覚型 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．

以上の結果は，本論文で提案した言語型 CAPTCHA と聴覚型 CAPTCHA が，既存方式に比べて安全性が高く，かつ実際に人間が解けるユーザビリティを備えていることを示している．

本研究に協力頂いた視覚障害者の方達からは，提案方式のアクセシブルな点について期待の言葉を頂くことができた．その一方で，ユーザビリティの改善を希望する声は多い．実験で聞き取った被験者らの意見には，問題が難しく感じるとの意見がしばしば見受けられた．提案方式の解答方式は選択肢式であるため，利用者は認証を通過するのに複数問を

解く必要がある．利用者がストレスなく複数問題を解くには，利用者の感ずる問題の難しさを低減しなければならない．

また，安全性については，危殆化の恐れがある．現在の AI 関連技術の進捗は著しため，ロボットの攻撃手法はますます高度化していくと考えられる．提案方式についても，機械学習などの研究結果を注視し，安全性の分析を継続する必要がある．また，より高度な人間の認知能力の利用や複数方式の組み合わせの検討なども重要である．

参考文献

[1] The Official CAPTCHA Site, http://www.captcha.net/ (retrieved May 10, 2013).

[2] BBC News: Bots now Account for 61% of Web Traffic, http://www.bbc.com/news/technology-25346235 (retrieved August 1, 2016).

[3] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using Hard AI Problems for Security. In *Proceedings of EUROCRYPT*, volume 2656 of *LNCS*, pages 294–311. Springer-Verlag, 2003.

[4] Kumarasubramanian Abishek, Ostrovsky Rafail, Pandey Omkant, and Wadia Akshay. Cryptography Using CAPTCHA Puzzles. In *Public-Key Cryptography - PKC 2013*, volume 7778, page 89, 2013.

[5] *Proposal and Evaluation of Cyber Defense System Using Blacklist Refined Based on Authentication Results*. Conference Publishing Service, 2016.

[6] Captcha Chronicles: Game over for Pokemon GO Bots and Trackers, http://pokemongohub.net/captcha-game-pokemon-go-bots-trackers/ (retrieved September 1, 2016).

[7] Elie Bursztein, Matthieu Martin, and John Mitchell. Text-based CAPTCHA Strengths and Weaknesses. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, pages 125–138. ACM, 2011.

[8] 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*, pages 19–31, 2011.

[9] Project Stiltwalke, http://www.dc949.org/projects/stiltwalker/ (retrieved May 1, 2013).

[10] 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.

[11] Suphannee Sivakorn, Iasonas Polakis, and Angelos D. Keromytis. I Am Robot: (Deep) Learning to Break Semantic Image CAPTCHAs. In *Proceedings of the 1st IEEE European Symposium on Security and Privacy*, EuroSP '16, 2016.

[12] Arthur I Karshmer, Harley R Myler, and Richard D Davis. The Architecture of an Inexpensive and Portable Talking-tactile Terminal to Aid the Visually Handicapped. *Computer Standards & Interfaces*, 6(2), Jan 1987.

[13] Bart Bauwens, Filip Evenepoel, and Jan Engelen. SGML As an Enabling Technology for Access to Digital Information by Print Disabled Readers. *Computer Standards & Interfaces*, 18(1):55–69, Jan 1996.

[14] S. Lewthwaite. Web Accessibility Standards and Disability: Developing Critical Perspectives on Accessibility. *PubMed*, 36(16):1375–1383, Jul 2014.

[15] Tim Berners-Lee. Long Live the Web. *Scientific American*, 303(6):80–85, Nov 2010.

[16] 障害者によるインターネットの利用率, http://barrierfree.nict.go.jp/relate/statistics/hc_internet.html (2013 年 11 月 30 日に参照).

[17] 障がいのある方々のインターネット等の利用に関する調査研究 [結果概要], http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2012/disabilities2012.pdf (2013 年 11 月 30 日に参照).

[18] Web Content Accessibility Guidelines (WCAG) 2., https://www.w3.org/TR/2008/REC-WCAG20-20081211/ (retrieved November 25, 2016).

[19] Inaccessibility of CAPTCHA, https://www.w3.org/TR/turingtest/ (retrieved March 15, 2013).

[20] Jeremy Elson, John R. Douceur, Jon Howell, and Jared Saul. Asirra: a CAPTCHA that Exploits Interest-aligned Manual Image Categorization. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, CCS '07, pages 366–374. ACM, 2007.

[21] Ritendra Datta, Jia Li, and James Ze Wang. Exploiting the human-machine gap in image recognition for designing CAPTCHAs. *IEEE Trans. Information Forensics and Security*, 4(3):504–518, 2009.

[22] Haichang Gao, Dan Yao, Honggang Liu, Xiyang Liu, and Liming Wang. A Novel Image Based CAPTCHA Using Jigsaw Puzzle. In *Proceedings of the 2010 13th IEEE International Conference on Computational Science and Engineering*, CSE '10, pages 351–356. IEEE Computer Society, 2010.

[23] NuCaptcha II Most secure and usable Captcha, http://www.nucaptcha.com/ (retrieved March 5, 2016).

[24] 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*, CHI '09, pages 1829–1838. ACM, 2009.

[25] 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*, SP '10, pages 399–413. IEEE Computer Society, 2010.

[26] Sajad Shirali-Shahreza and M. Hassan Shirali-Shahreza. Accessibility of CAPTCHA Methods. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, pages 109–110. ACM, 2011.

[27] BBC News: Blind Federation Criticises Captcha Security Test, http://www.bbc.com/news/technology-22754006 (retrieved February 5, 2014).

[28] Accessible Rich Internet Applications (WAI-ARIA) 1.1, https://www.w3.org/TR/2016/CR-wai-aria-1.1-20161027/ (retrieved November 25, 2016).

[29] 障害者基本法, http://www8.cao.go.jp/shougai/suishin/kihonhou/s45-84.pdf (2016 年 11 月 27 日に参照).

[30] 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.

[31] David M Schnyer Daniel L Schacter, Ian G Dobbins. Specificity of Priming: a Cognitive Neuroscience Perspective. *Nature Reviews Neuroscience*, 5(11):853–862, Nov 2004.

[32] Richard Bergmair and Stefan Katzenbeisser. Towards Human Interactive Proofs in the Text-Domain (Using the Problem of Sense-Ambiguity for Security). In *In Proceedings of the 7th International Conference, ISC 2004*, LNCS, pages 257–267. Springer, 2004.

[33] Christopher Liam Ivey. System and Method for Delivering a Human Interactive Proof to the Visually Impaired by Means of Semantic Association of Objects, USPTO Applicaton 20120232907, 2012.

[34] 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.

[35] We the People: Your Voice in Our Government, https://petitions.whitehouse.gov/ (retrieved February 5, 2014).

[36] Rob Tuley. TextCaptcha service: Text CAPTCHA Logic Questions, http://textcaptcha.com/ (retrieved January 14, 2014).

[37] 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.

[38] Taku Kudo. MeCab : Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/ (retrieved May 1, 2013).

[39] 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.

[40] 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.

[41] The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict (retrieved September 1, 2016).

[42] Gyorgy Kepes. *Language of Vision*. Dover, 1995.

[43] I. Biederman. Recognition-by-components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147, 1987.

[44] Henry S. Baird, Allison L. Coates, and Richard J. Fateman. PessimalPrint: a Reverse Turing Test. *IJDAR*, 5(2–3):158–163, 2003.

[45] Luis von Ahn, Manuel Blum, and John Langford. Telling Humans and Computers Apart Automatically. *Commun. ACM*, 47(2):56–60, Feb 2004.

[46] Monica Chew and Henry S. Baird. BaffleText: a Human Interactive Proof. In *Proceedings of SPIE, Document Recognition and Retrieval X*, volume 305. SPIE, Jan 2003.

[47] Anand Gupta, Ashish Jain, Aditya Raj, and Abhimanyu Jain. Sequenced Tagged Captcha: Generation and its Analysis. In *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pages 1286–1291, Mar 2009.

[48] A. Rusu, A. Thomas, and V. Govindaraju. Generation and use of handwritten CAPTCHAs. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(1):49–64, 2010.

[49] *Drag and Drop: A Better Approach to CAPTCHA*, 2009.

[50] H. D. Truong, C. F. Turner, and C. C. Zou. iCAPTCHA: The Next Generation of CAPTCHA Designed to Defend against 3rd Party Human Attacks. In *2011 IEEE International Conference on Communications (ICC)*, pages 1–6, Jun 2011.

[51] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895):1465–1468, 2008.

[52] Elie Bursztein, Jonathan Aigrain, Angelika Moscicki, and John C. Mitchell. The End is Nigh: Generic Solving of Text-based CAPTCHAs. In *8th USENIX Workshop on Offensive Technologies (WOOT 14)*. USENIX Association, Aug 2014.

[53] Roger N. Shepard and Jacqueline Metzler. Mental Rotation of Three Dimensional Objects. *Science*, 171(3972), 1971.

[54] Roger N. Shepard and Lynn A. Cooper. *Mental Images and their Transformations*. MIT Press, 1982.

[55] Luis Von Ahn, Manuel Blum, , and John Langford. Telling Human and Computers Apart Automatically. In *Communications of the ACM*, volume 47, pages 57–60, Feb 2004.

[56] M. E. Hoque, D. J. Russomanno, and M. Yeasin. 2D Captchas from 3D Models. In *Proceedings of the IEEE SoutheastCon 2006*, pages 165–170, Mar 2006.

[57] Mohammad Shirali-Shahreza and Sajad Shirali-Shahreza. Collage CAPTCHA. In *9th International Symposium on Signal Processing and Its Applications, ISSPA 2007*, pages 1–4, 2007.

[58] Rich Gossweiler and Maryam Kamvar and Shumeet Baluja. What's Up CAPTCHA? A CAPTCHA Based On Image Orientation. In *WWW 2009*, 2009.

[59] M. Imsamai and S. Phimoltares. 3D CAPTCHA: A Next Generation of the CAPTCHA. In *Proceedings of Information Science and Applications 2010*, ICISA '10, pages 1–8. IEEE, 2010.

[60] 田村 拓己, 久保田 真一郎, 油田 健太郎, 片山 徹郎, 朴 美娘, and 岡崎 直宣. 文字認識攻撃に耐性を持つランダム妨害図形を用いた画像ベース CAPTCHA 方式の提案. **情報処理学会論文誌**, 56(3):808–818, Mar 2015.

[61] 藤田 真浩, 池谷 勇樹, 可児 潤也, and 西垣 正勝. Locimetric 型メンタルローテーション CAPTCHA. **情報処理学会論文誌**, 57(9):1954–1964, Sep 2016.

[62] 藤田 真浩, 池谷 勇樹, 可児 潤也, and 西垣 正勝. 非現実画像 CAPTCHA：常識からの逸脱を利用した 3DCG 画像 CAPTCHA. **情報処理学会論文誌**, 56(12):2324–2336, Dec 2015.

[63] reCAPTCHA: Easy on Humans, Hard on Bots, `https://www.google.com/recaptcha/intro/index.html` (retrieved October 1, 2016).

[64] 可児 潤也, 鈴木 徳一郎, 上原 章敬, 山本 匠, and 西垣 正勝. 4 コマ漫画 captcha. **情報処理学会論文誌**, 54(9):2232–2243, Sep 2013.

[65] Bin B. Zhu, Jeff Yan, Qiujie Li, Chao Yang, Jia Liu, Ning Xu, Meng Yi, and Kaiwei Cai. Attacks and Design of Image Recognition CAPTCHAs. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 187–200. ACM, 2010.

[66] Elias Athanasopoulos and Spiros Antonatos. *Enhanced CAPTCHAs: Using Animation to Tell Humans and Computers Apart*, pages 97–108. Springer Berlin Heidelberg, 2006.

[67] Sajad Shirali-Shahreza and Mohammad Shirali-Shahreza. A New Human Interactive Proofs System for Deaf Persons. In *Fifth International Conference on Information Technology: New Generations (ITNG 2008)*, pages 807–810, 2008.

[68] Manar Mohamed, Niharika Sachdeva, Michael Georgescu, Song Gao, Nitesh Saxena, Chengcui Zhang, Ponnurangam Kumaraguru, Paul C. van Oorschot, and Wei-Bang Chen. A Three-way Investigation of a game-CAPTCHA: Automated Attacks, Relay Attacks and Usability. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, ASIA CCS '14, pages 195–206. ACM, 2014.

[69] Song Gao, Manar Mohamed, Nitesh Saxena, and Chengcui Zhang. Emerging Image Game CAPTCHAs for Resisting Automated and Human-Solver Relay Attacks. In *Proceedings of the 31st Annual Computer Security Applications Conference*, ACSAC '15, pages 11–20. ACM, 2015.

[70] Mauro Conti, Claudio Guarisco, and Riccardo Spolaor. *CAPTCHaStar! A Novel CAPTCHA Based on Interactive Shape Discovery*, pages 611–628. Springer International Publishing, 2016.

[71] E. Colin Cherry. 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, 1953.

[72] 赤木 正人. カクテルパーティー効果とそのモデル化. **電子情報通信学会誌**, 78(5):450–453, may 1995.

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

[74] Greg Kochanski, Daniel P. Lopresti, and Chilin Shih. A reverse turing test using speech. In *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH 2002*, 2002.

[75] 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*, Assets '07, pages 245–246. ACM, 2007.

[76] 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.

[77] H. Gao, H. Liu, D. Yao, X. Liu, and U. Aickelin. An Audio CAPTCHA to Distinguish Humans from Computers. In *Electronic Commerce and Security (ISECS), 2010 Third International Symposium*, pages 265–269, Jul 2010.

[78] Yannis Soupionis and Dimitris Gritzalis. Audio CAPTCHA: Existing Solutions Assessment and a New Implementation for VoIP Telephony. *Computers and Security*, 29(5):603–618, Jul 2010.

[79] 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.

[80] Matthew Davidson, Karen Renaud, and Shujun Li. *jCAPTCHA: Accessible Human Validation*, pages 129–136. Springer International Publishing, 2014.

[81] Jonathan Lazar, Jinjuan Feng, Tim Brooks, Genna Melamed, Brian Wentz, Jon Holman, Abiodun Olalere, 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*, CHI '12, pages 2267–2276. ACM, 2012.

[82] Abiodun Olalere, Jinjuan Heidi Feng, Jonathan Lazar, and Tim Brooks. Investigating the Effects of Sound Masking on the Use of Audio CAPTCHAs. *Behaviour and Information Technology*, 33(9):919–928, 2014.

[83] J. C. R. Licklider George A. Miller. The Intelligibility of Interrupted Speech.

[84] P.K. Dick. *Do Androids Dream of Electric Sheep? (resale)*. Ballantine Books, 1996.

[85] 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 of *LNCS*, pages 605–615. Springer-Verlag, 2006.

[86] *SemCAPTCHA—User-friendly Alternative for OCR-based CAPTCHA System*, 2008.

[87] G. Mori and J. Malik. Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I134–141, Jun 2003.

[88] Kumar Chellapilla and Patrice Y. Simard. Using Machine Learning to Break Visual Human Interaction Proofs (HIPs). In *Advances in Neural Information Processing Systems 17*, pages 265–272. MIT Press, 2004.

[89] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large Margin Methods for Structured and Interdependent Output Variables. *J. Mach. Learn. Res.*, 6:1453–1484, Dec 2005.

[90] Kumar Chellapilla, Kevin Larson, Patrice Simard, and Mary Czerwinski. Computers Beat Humans at Single Character Recognition in Reading based Human Interaction Proofs (HIPs). In *CEAS 2005, Conference on Email and Anti-Spam*, Jan 2005.

[91] Jeff Yan, Ahmad Salah, and El Ahmad. Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms. In *Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007)*, pages 279–291, Dec 2007.

[92] Jeff Yan and Ahmad Salah El Ahmad. A Low-cost Attack on a Microsoft Captcha. In *Proceedings of the 15th ACM Conference on Computer and Communications Security*, CCS '08, pages 543–554. ACM, 2008.

[93] Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay D. Shet. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. *CoRR*, abs/1312.6082, 2013.

[94] Ahmad S El Ahmad, Jeff Yan, and Mohamad Tayara. The Robustness of Google CAPTCHAs. Technical report, School of Computer Science, Newcastle University, UK, May 2011.

[95] Paul Baecher, Niklas Buscher, Marc Fischlin, and Benjamin Milde. Breaking reCAPTCHA: A Holistic Approach via Shape Recognition. In *SEC*, volume 354 of *IFIP Advances in Information and Communication Technology*, pages 56–67. Springer, 2011.

[96] Philippe Golle. Machine Learning Attacks against the Asirra CAPTCHA. In *Proceedings of the 15th ACM Conference on Computer and Communications Security*, CCS '08, pages 535–542. ACM, 2008.

[97] How we broke the NuCaptcha video scheme and what we propose to fix it, `https://www.elie.net/blog/security/how-we-broke-the-nucaptcha-video-scheme-and-what-we-propose-to-fix-it` (retrieved March 5, 2016).

[98] Jennifer Tam, Jiri Simsa, Sean Hyde, and Luis von Ahn. *Breaking Audio CAPTCHAs*. Advances in Neural Information Processing Systems 21. MIT Press, 2008.

[99] Jennifer Tam, Jiri Simsa, David Huggins-daines, Luis Von Ahn, and Manuel Blum. Improving Audio CAPTCHAs. In *In Proceedings of the 4th Symposium on Usability, Privacy and Security (SOUPS '08)*, 2008.

[100] Yoav Freund and Robert E. Schapire. Experiments with a New Boosting Algorithm. In *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96)*, pages 148–156, 1996.

[101] B.V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. NN Norms : NN Pattern Classification Techniques. IEEE Computer Society Press, 1991.

[102] Hynek Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[103] Vibha Tiwari. MFCC and its Applications in Speaker Recognition. *International Journal on Emerging Technologies*, 2009.

[104] Defeating Audio (Voice) CAPTCHA, `http://vorm.net/captchas` (retrieved May 1, 2013).

[105] Elie Bursztein and Steven Bethard. Decaptcha: Breaking 75% of eBay Audio CAPTCHAs. In *Proceedings of the 3rd USENIX Conference on Offensive Technologies*, WOOT'09, pages 8–8. USENIX Association, 2009.

[106] David A. Ferrucci. IBM's Watson/DeepQA. *SIGARCH omputer Architecture News*, 39(3):–, Jun 2011.

[107] Robert M. French. Moving beyond the Turing Test. *Communications of the ACM*, 55(12):74–77, Dec 2012.

[108] 森本 浩介, 片瀬 弘晶, and 山名 早人. *N*-gram と離散型共起表現を用いたワードサラダ型スパム検出手法の提案. **研究報告データベースシステム**（*DBS*）, 2009(24):1–8, Jul 2009.

[109] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting Spam Web Pages Through Content Analysis. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 83–92. ACM, 2006.

[110] Gilchan Park, Lauren M. Stuart, ulia M. Taylor, and Victor Raskin. Comparing Machine and Human Ability to Detect Phishing Emails. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2322–2327. IEEE, Oct 2014.

[111] Podec Trojan Can Trick CAPTCHA Into Thinking It Is Human, http://www.techweekeurope.co.uk/security/authentification/podec-malware-captcha-kaspersky-labs-164284 (retrieved August 15, 2015).

[112] Trojan.Captchar.A — Symantec, https://www.symantec.com/security_response/writeup.jsp?docid=2007-103012-0328-99 (retrieved May 1, 2013).

[113] Manuel Egele, Leyla Bilge, Engin Kirda, and Christopher Kruegel. CAPTCHA Smuggling : Hijacking Web Browsing Sessions to Create CAPTCHA farms. In *SAC 2010, 25th ACM Symposium On Applied Computing, March 22-26, 2010, Sierre, Switzerland*, Mar 2010.

[114] Graig Sauer, Harry Hochheiser, Jinjuan Feng, and Jonathan Lazar. Towards a Universally Usable CAPTCHA. In *In Proceedings the 4th Symposium On Usable Privacy and Security (SOUPS '08)*, 2008.

[115] Marios Belk, Christos Fidas, Panagiotis Germanakos, and George Samaras. Do Human Cognitive Differences in Information Processing Affect Preference and Performance of CAPTCHA? *International Journal of Human-Computer Studies*, 84(1):1–18, 2015.

[116] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1 of *ACL '98*, pages 86–90. Association for Computational Linguistics, 1998.

[117] Welcome to FrameNet, https://framenet.icsi.berkeley.edu/fndrupal/ (retrieved February 5, 2014).

[118] 青空文庫　Aozora Bunko, http://www.aozora.gr.jp/ (retrieved May 1, 2013).

[119] 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** *The transactions of Human Interface Society*, 15(1):337–352, Nov 2013.

[120] Yuji Matsumoto and Taku Kudo. Japanese Dependency Analysis using Cascaded Chunking. In *In Proceedings of the 6th Conference on Natural Language Learning*, pages 63–69, 2002.

[121] ウィキペディア, `https://ja.wikipedia.org/wiki` (2016 年 11 月 25 日に参照).

[122] Wikipedia, `https://en.wikipedia.org/wiki` (retrieved November 25, 2016).

[123] S.J. Young and Sj Young. The HTK Hidden Markov Model Toolkit: Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44, 1994.

[124] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[125] L. E. Baum and J. A. Eagon. An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology. *Bull. Am. Math. Soc.*, 73:360–363, 1967.

[126] Hugo Steinhaus. Sur la division des corps matériels en parties (French). *Bull. Acad. Polon. Sci. Cl. III. 4*, pages 801–804, 1956.

[127] 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, Jun 2001.

[128] 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.

[129] Preslav Nakov, Francisco Guzman, and Stephan Vogel. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *COLING'12*, pages 1979–1994, 2012.

謝辞

本研究を遂行し学位論文をまとめるにあたり，多くのご支援とご指導を賜りました．

指導教官である明治大学大学院先端数理学研究科研究科長の菊池浩明教授に，心より感謝いたします．終始笑顔を絶やさず，しかし時に厳しくご指導頂いたことを通して，私自身の至らなさや研究者としての心構えを学ぶことができたことは，今後の努力の糧になるものであります．

チームフェローである明治大学大学院先端数理学研究科の杉原厚吉教授と二宮広和教授に，心より感謝いたします．それぞれの専門分野を背景としたご指導により，研究に関する広い視点の必要性を学ぶことができました．

博士課程への進学および研究全般に渡るご支援を賜りました筑波技術大学保健科学部情報システム学科の岡本健准教授に，心より感謝致します．博士課程前期でのご指導を通して，研究への取り組みの基礎を学ぶことができました．

本研究分野を始めるにあたりご指導を賜りました産業技術総合研究所人工知能研究センター知識情報研究チーム長である中田亨博士に，心より感謝いたします．熱心なご指導や議論を通して，新しい研究分野に取り組む姿勢を学ぶことができました．

萩原一郎所長を始めとする明治大学先端数理科学インスティテュート (MIMS) の所員・研究員各位に，心より感謝いたします．本研究科の充実した研究環境とご支援のおかげで，自身の研究に集中することができました．

最後に，研究に関わりました関係者各位と家族の支援に深く感謝いたします．

研究業績

学術論文誌

1. 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**，情報処理学会，56(9):1834–1845，Sep 2015.

2. 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** *The transactions of Human Interface Society*，ヒューマンインタフェース学会，15(1):337–352，Nov 2013.

紀要

1. 岡本 健，山口 通智，三宅 輝久，石塚 和重，野口 栄太郎，大越 教夫．“視覚に障害をもつ医療系学生に適する情報セキュリティ技術”．**筑波技術大学紀要**，筑波技術大学，21(2)：17–22，Mar 2014.

2. 山口 通智，岡本 健．“文意や文脈の解釈問題を用いた視覚障害者向け CAPTCHA とその評価”．**筑波技術大学紀要**，筑波技術大学，21(2)：12–16，Mar 2014.

国際会議投稿論文

1. Michitomo Yamaguchi, Takeshi Okamoto and Hiroaki Kikuchi. “CAPTCHA System by Differentiating the Awkwardness of Objects”. *Proceedings of The 18th International Conference on Network-Based Information Systems 2015*, IEEE Conference Publishing Service, pp. 257-263, 2015.

2. Michitomo Yamaguchi, Toru Nakata, Hajime Watanabe, Takeshi Okamoto and Hiroaki Kikuchi. “Vulnerability of the Conventional Accessible CAPTCHA used by the White House and an Alternative Approach for Visually Impaired People”. *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, pp. 3946–3951, 2014.

3. Michitomo Yamaguchi, Toru Nakata, Takeshi Okamoto and Hiroaki Kikuchi. “An Accessible CAPTCHA system for People with Visual Disability — Generation of Human/Computer Distinguish Test with Documents on the Net”. *Proceedings of the 16th International Conference on Human-Computer Interaction*, Springer-Verlag, pp. 119–130, 2014.

国際学会ポスター発表

1. Michitomo Yamaguchi and Hiroaki Kikuchi. “Audio-CAPTCHA with Distinction between Random Phoneme Sequences and Words Spoken by Multi-speaker”. *International Conference on Mathematical Modeling and Applications*, 2016.

国内研究会投稿論文

1. 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”．**コンピュータセキュリティシンポジウム** *2016* **論文集**，2016(2):363–370，2016.

2. 山口 通智，岡本 健，菊池 浩明．“不自然さの識別問題を用いた CAPTCHA に関する研究”，**情報通信システムセキュリティ研究会**，114(489):55–60，2015.

付 録A

A.1　1つの被験者群で異なる2つの実験条件での有意性の検定方法

5.4 節の実験結果から子音交替適用前と適用後の有意性を判定するために，本論文で使用した検定方法を示す.

- 有意水準は，心理実験で慣習である 5% とする.
- 被験者 i ごとに，子音交替適用前の作問に対する正答数 A_p と，適用後の作問に対する正答数 A_a を調べる．$A_p > A_a$ ならば $D_i = 1$，$A_p < A_a$ ならば $D_i = -1$，$A_p = A_a$ ならば $D_i = 0$ とする．D_i について，12 人の被験者について和を求め，$D = \Sigma_{i=0}^{11} D_i$ を得る.
- 帰無仮説として，D_i が正負ランダム，すなわち二項分布すると仮定する.
- 二項分布に従い D の値を調べると，$|D| \geq 8$ となる確率は 3.2%，$|D| \geq 6$ では 10.6% になる．よって，$|D| \geq 8$ ならば有意水準 5% を満たすので，帰無仮説を棄却する．すなわち，有意差ありと判定する.

A.2　機械翻訳とその性能

本論文では，機械翻訳として Excite 翻訳[1]や Weblio 翻訳[2]を作問プログラムに組み込んでいる．これらの翻訳精度は，機械翻訳文識別テストの作問精度に影響を及ぼすが，オンラインサービスであるが故に，バージョン管理による再現実験には不向きである.

そこで，これらの翻訳性能について簡単に評価すると共に，本節での簡易評価で同程度の翻訳性能を持つオフライン動作可能なフリーの翻訳ソフト Yamato 英和.NET Lite Ver.1.08[3]を紹介する.

翻訳性能の指標には，スムージング型の BLEU+1 [129] を用い，原文とその正解となる和訳文のコーパスには，AAMT 機械翻訳文テストセット[4] を用いた．このコーパスでは，

[1] http://www.excite.co.jp/world/
[2] http://translate.weblio.jp/
[3] http://www.vector.co.jp/soft/win95/edu/se364617.html
[4] http://corpus.aamt.info/corpus/ja/corpus.html

正解文が各原文につき 1 つしかないため，BLEU の評価としては，簡易的なものでしかないことに注意されたい．

結果のみを示すと，各指標値は，Excite 翻訳 0.319，Weblio 翻訳 0.311，Yamato 英和.NET Lite Ver 1.08 が 0.314 であった．Mann・Whitney の U 検定にて，これらに有意性が無いことを確認した．