発話型属性認証の提示文章作成に向けた誤記修正特性分析

Analysis of Error-Correction Characteristic for Sentence Generation Used in Speech-Based CAPTCHA

井戸 智斗志 [1]，石田 繁巳 [1]，稲村 浩 [1]
Satoshi Ido [1], Shigemi Ishida [1], Inamura Hiroshi [1]
[1] 公立はこだて未来大学
[1] Future University Hakodate

1 はじめに

AI の進歩とともに，人間を対象とした入力インターフェイスにおいて人間が操作していることを判別する「属性認証」の技術も高度化している．一般に，属性認証には CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) 認証が用いられる．特に，テキスト入力による属性認証は広く利用されている．

多種多様なデバイス環境に対して様々な入力インタフェースが登場しており，これらに対する属性認証が今後必要となることが予想される．実際，音声認証において，合成音声によるなりすましに成功している事例が報告されている [1].

本研究では，音声入力インタフェースに対する属性認証として，人間の文字認識特性を用いた発話型属性認証を提案する．本稿では，その実現に向けて認証の際に提示する文章を実験的に模索した結果を示す．

2 関連研究

Li らは，自己教師あり事前学習モデル HuBERT に基づく偽造音声検出手法 HuRawNet_modified を提案した [2]．HuBERT モデルによって偽造音声における単語間の異常な無音を抽出し，学習データに含まれない偽造音声も検出できることを示した．しかしながら，文献 [1] などの手法によって属性認証を突破できることが示されており，新たな属性認証手法が求められている．

3 人間の文字認識特性を用いた発話型属性認証手法

本手法のキーアイデアは，人間と機械が誤記を含む文章を読み上げるときの発話的特徴に基づいて属性認証することである．人間は，普段使用する単語の一部に誤記が含まれていても無意識に修正して正しく読むことができる [3]．普段使用しない単語は正しく修正することができず，言い淀むなど人間固有の反応を示す．一方で，機械は LLM (Large Language Models) を用いて誤記を含む英文を高精度に修正できる [4]．そこで，人間と機械とで発話的特徴の差が大きくなる文章を読み上げてもらい，その修正精度に基づいて属性認証を行う．

図 1 に，提案手法の概要を示す．本手法は文章作成ブロック，音声入力ブロック，発話型属性認証ブロックの 3 つのブロックで構成されている．文章作成ブロックでは，あらかじめ用意された文章を分かち書きし，指定した位置に誤記を加えて提示する．音声入力ブロックでは，マイクから入力される音声を音声ファイルに変換して発話型属性認証ブロックに入力する．発話型属性認証ブロックでは，入力された音声から判定要素ごとに必要な情報を抽出し，人間か機械かを判定する．

4 評価

本手法の実現に向けては，機械と人間とで発話的特徴の差が大きくなる文章を提示することが重要である．人間と機械で発話的特徴に差が出る文章の作成方法を検討するため，異なる種類の誤記に対する人間と LLM の修正率を評価した．

4.1 評価方法

誤記を含む文章として，1 個の正常な文章から誤記の種類と位置が異なる 15 タイプの文章を作成した．ランダムな文章を生成して分かち書きし，分節数を 3 等分して文章を前方，中央，後方の 3 つの部分に分解した．そして，表 1 に示す 5 種類の誤記 [5] のいずれか 1 つを各部分に加えた文章を作成し，15 タイプの文章を得た．本稿では，Gemini 1.0 Pro を用いて 17 文字程度の文章を 100 個生成し，誤記を含む 100 個 × 15 タイプ の文章を準備した．

人間の発話音声データは，Web アプリケーションを実装して収集した．被験者が Web アプリケーション上の録音ボタンを押すと，誤記を含む文章が表示される．

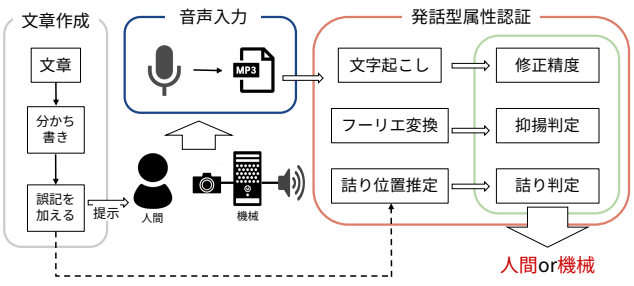

図 1: 提案手法の概要

表 1: 誤記の種類と具体例

誤記の種類	具体例：はなびらが
入替	はならびが
削除	は びらが
代用	はなひらが
結合	はびらが
挿入	はなびらはが

表 2: 人間と機械の修正率 [%]

誤記タイプ		人間	Gemini	GPT-4	GPT-4o
種類	位置				
入替	前方	34	28	21	48
	中央	41	33	30	57
	後方	39	29	28	51
削除	前方	17	32	38	54
	中央	18	36	37	50
	後方	19	38	45	59
代用	前方	56	39	48	60
	中央	58	47	42	64
	後方	50	50	48	68
結合	前方	20	31	37	52
	中央	17	41	39	53
	後方	20	39	36	50
挿入	前方	16	39	37	60
	中央	27	39	44	53
	後方	24	42	39	62

被験者は 10 秒以内にマイクに向かって文章を読み上げる．マイクは FIFINE AmpliGame A8W を用いた．被験者は 15 人である．各被験者には，100 個の文章のそれぞれに対していずれかのタイプの文章を提示した．

LLM に対しては，Gemini 1.0 Pro，GPT-4，GPT-4o それぞれの API を用いて誤記を含む 100 個 × 15 タイプの文章を修正させ，VOICEVOX(キャラクター番号：2) を用いて発話音声データに変換した．修正を指示するプロンプトは誤記タイプによらず共通のものとした．

誤記の修正率は，発話音声データを文字起こしした上で，誤記を含まない元の文章と比較することで評価した．人間と LLM から収集した発話音声データを Whisper large-v2 を用いて文字起こしし，MeCab(辞書：mecab-ipadic-NEologd) を用いてひらがなに変換した．誤記を加える前の元の文も同様にひらがなに変換し，発話音声データから得られたひらがなの文章と完全一致するかを確認した．

4.2 評価結果

表 2 に，人間と機械の修正率を示す．表の赤色・青色セルは人間・機械の修正率のそれぞれ最大値・最小値を示している．入替，代用の誤記に対しては，人間・機械ともにそれぞれ比較的低い，高い修正率であった．

人間と機械の修正率の差を確認するため，人間の修正率を基準とする人間と機械の修正率の比を計算した．表 3 に，人間と機械と修正率の比を示す．表の赤色・青色セルはそれぞれ最大値・最小値を示している．LLM Mean は 3 種類の LLM の修正率の平均値と人間の修正率の比である．表より，文章の前方に挿入の誤記を加えた文章において，人間と機械の修正率の比が大きい傾向が認められる．この結果から，人間の文字認識特性を用いた発話型属性認証に向けては，前方に挿入の誤記を加えた文章を提示することが望ましいと言える．

表 3: 人間と機械の修正率の比

誤記タイプ		Gemini	GPT-4	GPT-4o	LLM
種類	位置				Mean
入替	前方	0.82	0.62	1.41	0.95
	中央	0.80	0.73	1.39	0.98
	後方	0.74	0.72	1.31	0.92
削除	前方	1.88	2.24	3.18	2.43
	中央	2.00	2.06	2.78	2.28
	後方	2.00	2.37	3.11	2.49
代用	前方	0.70	0.86	1.07	0.88
	中央	0.81	0.72	1.10	0.88
	後方	1.00	0.96	1.36	1.11
結合	前方	1.55	1.85	2.60	2.00
	中央	2.41	2.29	3.12	2.61
	後方	1.95	1.80	2.50	2.08
挿入	前方	2.44	2.31	3.75	2.83
	中央	1.44	1.63	1.96	1.68
	後方	1.75	1.63	2.58	1.99

5 おわりに

本稿では，人間の文字認識特性を用いた発話型属性認証の実現に向けて，認証の際に提示する文章を実験的に模索した．人間と機械に誤記を含む文章を提示して発話音声データを収集した．それぞれの発話音声データを文字起こしして誤記を加える前の文章と比較し，修正率を評価した．その結果，前方に挿入の誤記を加えた文章を提示することが望ましいことが明らかになった．今後の展望として，修正率の判定閾値の設定などを行う予定である．

参考文献

[1] Kassis, A. and Hengartner, U.: Breaking Security-Critical Voice Authentication, *2023 IEEE Symposium on Security and Privacy*, pp. 951–968 (2023).

[2] Li, L., Lu, T., Ma, X. and Yuan, M.: Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT, *Applied Sciences*, Vol. 13, No. 14, p. 8488 (2023).

[3] 久保田萌々，藤川真樹，鈴木真樹史: タイポグリセミアを用いた Multi-model CAPTCHA の提案と評価，産業応用工学会論文誌， Vol. 11, No. 1, pp. 54–64 (2023).

[4] Cao, Q., Kojima, T., Matsuo, Y. and Iwasawa, Y.: Unnatural Error Correction: GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8898–8913 (2023).

[5] Marques, M., Jiang, X., Dufor, O., Berrou, C. and Kim-Dufor, D.-H.: A Connectionist Model of Reading with Error Correction Properties, *7th Language and Technology Conference, LTC 2015*, pp. 304–317 (2015).