Title	聴覚知覚に関連する音響特徴を用いた人間模倣音声検出
Author(s)	KHALID, ZAMAN
Citation	
Issue Date	2025-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/20072
Rights	
Description	Supervisor: 鵜木 祐史, 先端科学技術研究科, 博士

氏名	ZAMAN Khalid		
学位の種類	博士（情報科学）		
学位記番号	博情第 554 号		
学位授与年月日	令和 7 年 9 月 24 日		
論文題目	Human-Imitated Speech Detection Using Acoustic Features Related to Auditory Perception		
論文審査委員	鵜木　祐史	北陸先端科学技術大学院大学	教授
	岡田　将吾	同	教授
	長谷川　忍	同	教授
	吉高　淳夫	同	准教授
	栗林　稔	東北大学	教授
	水町　光徳	九州工業大学	教授

論文の内容の要旨

Speech plays an important role in human communication, allowing individuals to express thoughts, emotions, and intentions. In digital communication, speech is frequently targeted by attacks that manipulate or forge audio to deceive systems and compromise speaker identity, posing serious challenges to speech authentication and privacy. Among these threats, AI-generated synthetic speech is a prominent concern, produced using text-to-speech (TTS), voice conversion (VC), and deep learning-based techniques. These forms of spoofing are commonly evaluated in challenges such as ASVspoof and Audio Deepfake Detection (ADD), which assess vulnerabilities in automatic speaker verification systems. Despite their realism, AI-generated speech typically exhibits detectable artifacts or a uniform robotic tone, allowing current detection systems to identify it more reliably. In contrast, human-imitated speech, which is produced organically by humans mimicking others, often retains natural acoustic characteristics, making it harder to detect for both human listeners and machines.

Addressing this critical threat, the study initially aims to introduce and assess detection approaches based on standard and auditory-based features with deep learning, using a custom human-imitated speech dataset tailored for machine learning analysis. To evaluate the effectiveness of these approaches, a spoof detection model was trained using standard acoustic features exclusively on ASVspoof 2019 LA synthetic speech and then evaluated on both ASVspoof 2019 LA synthetic speech and human-imitated speech. Although the model performed well in detecting synthetic speech, its accuracy declined significantly when tested on human-imitated speech, revealing a clear generalization gap. However, when trained directly on the proposed human-imitated dataset, the same model and standard features achieved improved detection performance for imitated speech. To further examine this issue, auditory-based acoustic features derived from gammatone and gammachirp filterbanks, which are designed to closely mimic the filtering characteristics of the human inner ear, were evaluated to assess whether they could better capture relevant acoustic cues and improve detection performance. Furthermore, the experiments showed that these auditory-inspired features were more effective than standard acoustic representations in capturing discriminative variations. Although these auditory-based features performed better than standard acoustic representations, they were still not sufficient to achieve reliable detection. These findings underscored that effectively addressing the challenge of human-imitated speech will require acoustic features that are closely related to auditory perception.

Based on these considerations, this research proposes a two-phase framework: the first phase examines human listener tests through listening experiments to evaluate how accurately humans can distinguish human-imitated speech, while the second phase applies machine-based analysis of acoustic features related to auditory perception combined with machine learning techniques to distinguish between genuine and imitated speech. In the first phase, a human listening test was conducted as part of a human-centered three-phase approach to evaluate the participants' ability to distinguish accurately between genuine and imitated speech. For this evaluation, a representative subset of samples was selected from the human-imitated speech dataset proposed in this study, which was specifically designed to be compatible with machine learning frameworks. In the test, listeners were asked to classify each sample as genuine or imitated and their performance was measured by the percentage of correct responses. Building on insights from the human listening test, the second phase of the study proposes two sets of acoustic features related to auditory perception: eight timbral features, including boominess, depth, brightness, warmth, etc., and spectro-temporal modulation (STM) representations.

In the human listening test, the participants performed well, indicating that with sufficient training and exposure, listeners can effectively distinguish between genuine and imitated speech. Similarly, machine-based experiments using timbral features and STM representations, which are closely related to auditory perception, demonstrated promising discriminative capacity compared to standard acoustic features and reflected trends similar to human evaluation results.

In conclusion, this study highlights the limitations of current spoof detection systems in handling human-imitated speech and demonstrates that even auditory-based features alone are not sufficient for reliable detection. To address this challenge, it begins with human listening tests to evaluate how accurately listeners can distinguish human-imitated speech and proposes an auditory perception-based detection framework supported by new benchmark datasets.

Keywords: Human-imitated speech, Acoustic features, Auditory perception, Timbral features, Auditory models, Machine learning, Deep learning.

論文審査の結果の要旨

近年，AI 音声合成の急激な進展により，本物の人間の声を合成できるまでになりつつある．その結果，ディープフェイクとして知られるような「音声なりすまし」といった音声セキュリティ上の問題が起こっている．しかしながら，現状の生成 AI では，検出可能なアーティファクトやプロソディ上の違和感が残っているため，深層学習ベースの検出技術でも正確に音声なりすましの真偽を判別できる状況である．一方で，人間自身による他者の模倣（ものまね）音声も「音声なりすまし」として解釈されており，人間だけでなく上述の検出技術でも真偽判定が相当に難しい状況である．サイバーフィジカル空間において，音声情報を安心・安全に利用するためにも，ディープフェイク音声だけでなく模倣音声を含む，話者の真偽性を検証する基盤技術を早急に確立する必要がある．

本研究では，次の 3 つのステップを踏み，模倣音声による音声なりすましの真偽判別のための検出技術の確立を試みた．まず，人間による模倣音声の真偽判定能力を調査するために，本物の音声と人間による模倣音声の弁別に関する聴取実験を行った．人間による話者識別には話者への親密度が関係するため，ここでは細心の注意を払った実験パラダイムを構築した．実験参加者には話者識別が 90%を超えるまで繰り返し学習させ，これをクリアした参加者のみ弁別実験に参加した．その結果，人間による検出精度は約 71%であることを明らかにした．次に，ディープフェイク音声検出で利用される典型的な音響特徴を利用した検出技術では，ディープフェイク音声の検出精度が約 78%であるのに対し，模倣音声の検出精度が約 50～55%しか達成できず，その難しさが改めて確認された．

最後に，聴知覚の一つとして音色指標を利用した検出技術では，模倣音声の検出精度が 64%であった．さらに，聴知覚メカニズムに基づく音響特徴（スペクトル・時間変調表現）を利用した検出技術では，模倣音声の検出精度が約 70%と人間と同程度の性能を有することを明らかにした．

以上，本論文は，人間による模倣音声の検出精度を明らかにするとともに，その精度に肉薄する聴知覚メカニズム（聴覚末梢・中枢・高次モデル）に基づく模倣音声検出法を確立した．この着眼点は新規性と独創性を持ち，その成果が国際的に定評のある学術誌に掲載されるなど学術的水準も高い．本技術は，聴知覚メカニズムの解明の一助になり，応用範囲が広く，学術的に貢献するところも大きい．よって博士（情報科学）の学位論文として十分価値あるものと認めた．