氏　　名	ANUWAT CHAIWONGYEN		
学位の種類	博士（情報科学）		
学位記番号	博情第 538 号		
学位授与年月日	令和 6 年 12 月 24 日		
論文題目	Deepfake speech detection using speech-pathological features		
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
	吉高　淳夫	同	准教授
	KONGPRAWECHNON Waree	タマサート大学 SIIT	准教授
	KARNJANA Jessada	タイ・NECTEC	研究員

論文の内容の要旨

There is a great concern regarding the misuse of deepfake speech technology to synthesize a real person's voice. Therefore, developing speech-security systems capable of detecting deepfake speech remains paramount in safeguarding against such misuse. Although various speech features and methods have been proposed, their potential for distinguishing between genuine and deepfake speech remains unclear. Since speech-pathological features with deep learning are widely used to assess unnaturalness in disordered voices associated with voice-production mechanisms, investigated the potential of speech-pathological features for distinguishing between genuine and deepfake speech.

In this work, two categories of pathological speech features were investigated: perceptual and acoustic features. For perceptual features, eight characteristics were examined: depth, sharpness, booming, hardness, brightness, roughness, warmth, and reverberation. The acoustic features analyzed included jitter (three types), shimmer (four types), harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio, normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE). The proposed method was evaluated on four datasets: Automatic Speaker Verification Spoofing and Countermeasures Challenges (ASVspoof) 2019 and 2021, and Audio Deep Synthesis Detection (ADD) 2022 and 2023.

In the first step, two types of speech-pathological features, perceptual and acoustic, are investigated. The data from the feature extraction for each type of feature were averaged. These averaged features were then fed into a multi-layer perceptron neural network for training and evaluating the performance of the model.

After investigation, it was found that acoustic speech-pathological features and perceptual speech-pathological features could effectively detect deepfake speech, except for HNR. To improve the efficiency of the proposed features, the important features from both acoustic and perceptual speech-pathological features were selected. The results indicate that when the important speech-pathological features are combined, the efficiency of the proposed features is improved.

Consequently, aimed to enhance the efficiency of the acoustic speech-pathological features by using

segmental frames of analysis. This approach extends the dimension of the features beyond a simple average. The results indicated that using segmental frames of analysis significantly improved the efficiency of the acoustic speech-pathological features.

Therefore, in this work, proposes a method for detecting deepfake speech by using segmental frames of analysis of speech-pathological features. These features include jitter (*local*), jitter (*PPQ3*), jitter (*PPQ5*), shimmer (*local*), shimmer (*APQ3*), shimmer (*APQ5*), shimmer (*APQ11*), GNE, NNE, CHNR. These features are fed into a ResNet-18 for classification, and the results demonstrate that incorporating these ten features with ResNet-18 significantly improves the efficiency of detecting fake speech.

Moreover, this paper proposes a method of combining two models on the basis of two different dimensions of speech-pathological features to greatly improve the effectiveness of deepfake speech detection, along with mel-spectrogram features, to enhance detection efficiency. The proposed method is evaluated on the ASVspoof 2019, 2021, ADD 2022, and ADD 2023 datasets. It consistently outperforms the baselines in terms of accuracy, recall, F1-score, and F2-score across these datasets. However, the equal error rate for the ADD 2022 test set remains relatively high. Overall, the method demonstrates high performance and effectiveness in deepfake speech detection.

Keywords: Deepfake speech detection, speech-pathological features, acoustical features, perceptual features, and neural network.

論文審査の結果の要旨

近年，サイバーフィジカル空間における音声コンテンツの利用は，スマートフォンの普及やAIスピーカ等の登場とともに急激な伸びを示している．このような急激な需要拡大に対して，音声情報を安心・安全に利用するための技術革新や法整備は相当な遅れをとっており，音声コンテンツの不正利用やなりすまし，音声改ざんといった問題も招いている．サイバーフィジカル空間において，ディジタル表現された音声メディア情報を安心・安全に利用するために，話者情報の真偽性を検証する基盤技術を確立する必要がある．特に，生成AIの登場により，本物の人間の声を合成できるような懸念も高まりつつあり，音声セキュリティ上，大きな問題になっている．その防御策・解決策として，ディープフェイク音声検出の基盤技術の確立が喫緊の課題となっている．

本研究では，上述した基盤技術を確立する上で重要となる，ディープフェイク音声の検出に特化した音響特徴表現を検討している．従来，音響特徴量としてMFCCやメル対数スペクトルなどが利用されてきたが，これらはなりすまし検出やフェイク音声検出といったタスクに対し，適切であるかどうか議論されずに利用されている．本研究では，ディープフェイク音声でみられる不自然さを，言語聴覚士が音声知覚で利用する音声病理学的特徴（Speech pathological features）を利用して判別できるのではないかという着想に至り，この音声病理的特徴に関連する音響特徴量を利用した．これらの特徴には，韻律やジッター・シマー，高調波対雑音比といった音声生成に関連する情報だけでなく，明るさや粗さ，鋭さ，響きといった音色に関わる聴知覚属性の情報も含まれた．

本研究では，自動話者認識におけるなりすまし対策の研究用音声データセット（ASVspoof2019,

2021）とディープフェイク音声検出の研究用音声データベース（ADD2022, 2023）を利用した．提案法では，音声病理学的特徴をベースに 2 種類の識別器（多層パーセプトロンと ResNet18）を利用し，長時間あるいはフレームベース型で特徴抽出する体系でシステムが構成された．これらについて大規模評価を行った結果，ベースライン法よりも大幅に検出性能を向上させ，精度，再現率，F1 スコア，EER 等でよい結果を出すことに成功した．以上，本論文は，言語聴覚士が病理音声の判断に利用する特徴を考慮した深層学習ベースの識別器を構成することで，高性能のディープフェイク音声検出法を確立した．この着眼点は高い新規性と独創性を持ち，聴知覚メカニズムに基づいた検出法として，ディープフェイク音声の検出の他にも音声の異常さの推定といった課題へと応用範囲が広く，貢献度も高い．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．