Abstract

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 mecha- nisms, 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, war- mth, 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 se- lected. 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-pathologi- cal 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 fea- tures 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.