Abstract

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.