Differences between human- and machine-based audio deepfake detection – analysis on the ASVspoof 2019 database

Junichi Yamagashi

National Institute of Informatics, Japan

jyamagis@nii.ac.jp

To automatically detect audio deepfake and prevent spoofing attacks, we have built a large corpus, ASVSpoof2019, which pairs natural human speech with speech waveforms generated by several types of synthesis algorithms. The speech synthesis methods are diverse and include text-to-speech synthesis and voice conversion.

In this talk, we will first present the results of large-scale listening tests conducted on this database to discriminate between natural and synthetic human speech. In the test, the subjects were asked to conduct two role-playing tasks. In one task, they were asked to judge whether the utterance was produced by a human or machine, given an imagined scenario where they must detect abnormal telephone calls in the customer service center of a commercial bank. In the other task, the subjects listened to two utterances and were asked to judge whether they sounded like the same person's voice.

Next, the results of several automatic detection algorithms for similar tasks on the same database are presented. Finally, the differences between human- and machine-based audio deepfake detection are discussed.