Developing a forensic methodology for deepfake audio detection

Anna Bartle[1], *Matthew Boyle*[1], *Ben Gibb-Reid*[2], *Ronan Given*[1], *Amelia Gully*[2], *Philip Harrison*[2], *Vincent Hughes*[2], *Lucy Morville*[2], *Luca Resti*[3], *Andrew Robinson*[2] *and Jessica Wormald*[2]

[1]*Metropolitan Police Service, Audio Lab, UK.*

anna.bartle@met.police.uk

[2]*Department of Language and Linguistic Science, University of York, UK.*

[3]*Department of Computer Science, University of York, UK.*

In recent years, there has been huge growth in the performance and availability of AI-based systems capable of producing realistic-sounding synthetic voices. These systems can be used to generate synthetic speech from a non-specific individual or, perhaps more worryingly, be used to mimic the voice of a real person (via voice cloning or voice conversion) to produce speech that person never said. Deepfake audio therefore presents an emerging threat within law enforcement. In the UK, the Office of the Policing Chief Scientific Adviser and the National Police Chiefs' Council have identified deepfakes as a primary concern in ensuring the integrity of evidence passing through the criminal justice system. However, at present, there is no established and validated framework for forensic speech and audio experts to address questions about deepfakes.

In this paper, we describe a collaborative project between the Metropolitan Police Service (MPS) and the University of York to develop human-based audio deepfake detection methods to complement AI-based automatic detection tools. The aim is to effectively integrate human expertise into deepfake detection to improve reliability and provide explainability. The project utilises a large set of genuine and deepfake recordings made by MPS. The deepfakes were generated using a range of state-of-the-art commercial text-to-speech (i.e. voice cloning) and speech-to-speech (i.e. voice conversion) systems. The project consists of two strands of method development:

(1) Auditory and acoustic phonetic analysis: This involves adapting established forensic voice comparison methods (see ENFSI BPM 2021) to the task of deepfake detection in cases where there is a reference recording of a known speaker. Initially, we conducted auditory analysis across a cross-section of real and deepfake voices to identify potentially useful features, such as consonant and vowel realisations, long-term pitch and voice quality, as well as non-linguistic features, such as hesitations, pauses and breathing patterns. More systematic, data-driven, analysis will then be conducted on a larger set of recordings in order to quantify the performance of different features.

(2) Signal processing analysis: This involves taking methods from the signal processing domain to identify properties of the signal that contribute the most to deepfake detection. In this strand, we will focus on bi-spectrum analysis (detecting internal signal correlations), sub-band entropy (measuring signal smoothness), spectral feature trajectories (assessing the variability within a recording) and the envelope modulation spectrum (detecting periodicity in noise).

Once relevant features have been determined through the two strands of work, testing will be conducted with a set of previously unseen real and deepfake samples to assess the overall efficacy of the methods. These results will then be combined with the output of AI-based deepfake detection

systems being tested in-house by MPS to produce a three-pronged analysis framework integrating phonetic, automatic, and signal processing methods.

References

ENFSI (2021) Best Practice Manual for the Methodology of Forensic Speaker Comparison. ENFSI-BPM-FSC-01.