

Towards an interpretation framework for forensic audio deepfake detection

*Finnian Kelly¹, Anil Alexander¹, Anna Bartle², Colleen Driscoll³
and Peter Milne³*

¹*Oxford Wave Research, Oxford, UK*

²*Forensic Services, Metropolitan Police, UK*

³*Royal Canadian Mounted Police, Canada*

{finnian|anil}@oxfordwaveresearch.com, anna.bartle@met.police.uk,
{Colleen.Kavanagh|Peter.Milne}@rcmp-grc.gc.ca

Deepfake audio detection systems are designed to analyse an input speech sample and produce a detection score, which can be used to inform a decision about whether the speech sample in question is *real* (has been produced by a human speaker) or *deepfake* (has been generated using a text-to-speech or voice conversion model of a specific speaker’s voice).

A key question facing the use of such systems in a forensic context is how to reliably interpret the resulting detection score in a way that is suitable to inform legal decision-making. A natural solution is to apply a likelihood ratio (LR) framework to convert the detection score into an interpretable format.

A deepfake LR could be defined as a measure of the relative strength of support for two competing hypotheses: the likelihood of the evidence (i.e. the detection score) if H_{DF} (the hypothesis that the sample is deepfake) is true, divided by the likelihood of the evidence if H_R (the hypothesis that the sample is real) is true. Evaluating such an LR requires a representative set of real and fake samples, based on how these hypotheses are defined. Drawing on the LR framework as commonly applied in speaker recognition (Drygajlo, 2015), we could consider both specific-source and common-source approaches to defining the hypotheses:

- A *common-source* approach is agnostic to the specific identity of the speaker in the questioned sample, with potential hypotheses H_{DF} : “the speech in the questioned sample is deepfake” and H_R : “the speech in the questioned sample is human”, requiring a representative set of deepfake and real samples.
- A *specific-source* approach poses hypotheses specific to the speaker in the questioned sample, with potential hypotheses H_{DF} : “the speech from the speaker in the questioned sample is deepfake” vs “the speech from the speaker in the questioned sample is human”, requiring a representative set of deepfake and real samples from the specific speaker in the questioned sample.

If there are multiple real samples available for the speaker in the questioned sample, it may be possible to generate sufficient deepfake samples to adopt a specific-source approach; however, the common-source approach is likely to be the most practical option.

With either approach, the selection of representative real and deepfake samples is of central importance to the LR. To inform this selection, it is important first to understand factors affecting deepfake detection performance, including:

- Technical: recording device, noise and compression, duration.
- Speaker: sex/gender, age, language, accent.
- Algorithmic: the specific deepfake generation method.

To begin the process of developing an interpretation framework for audio deepfake detection, we will explore the application of different hypotheses and identify some of the key factors to consider in the selection of representative data. We will demonstrate some possibilities through examples with controlled and in-the-wild data, using FAUXDIO deepfake detection software.

We will also discuss the practitioners' perspective on interpreting the output of a deepfake detector, which raises questions about handling unknowns like the deepfake generation algorithm, and in a speaker comparison case involving a questioned deepfake, whether to apply a two-stage process of deepfake detection followed by speaker comparison, or to combine the two in tandem.

References

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen J., and T. Niemi (2015), Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, *Frankfurt: Verlag für Polizeiwissenschaft*