

AI-based audio processing: implications for forensic deepfake detection

Amaia Sagasti, Finnian Kelly

Oxford Wave Research, Oxford, UK

{amaia|finnian}@oxfordwaveresearch.com

While synthetic speech is often synonymous with the threat of deepfakes, the underlying algorithms are increasingly being applied to benign audio processing tasks, such as speech enhancement, restoration or neural compression. This AI-based processing has mixed implications for the forensic and investigative speech analysis landscape: On one hand, it may introduce distributional shifts in the feature representations used for deepfake detection [1], increasing the risk that these samples are misclassified as ‘fake’. On the other hand, AI-based processing may improve the accuracy and reduce hallucinations in automatic speech recognition systems by enhancing low-quality speech signals typical of challenging, real-world scenarios [2]. In this paper, we focus on the implications of AI-based processing for deepfake detection.

We first investigate the presence of AI-based processing artifacts within the high-dimensional SSL features often employed for deepfake detection. For this purpose, we extracted BEATs (a large-scale audio model that previously demonstrated effective generalisation and competitive performance for deepfake detection [3]) embeddings for a speech dataset of real, real AI-processed, and fake samples.

We selected 100 real files from the VoxCeleb2 dataset and applied a selection of speech enhancement algorithms: *DeClick2*, *DeFilter*, *DeVerberate3* and *Extract Dialogue* from the Acon Digital Acoustica toolkit, and *Speech Enhance* from Adobe, to create several AI-processed sets. A corresponding set of 100 ‘fake’ (unprocessed) samples was generated using *Chatterbox* voice conversion from Resemble AI, to compare against the real and AI-processed samples.

After extracting BEATs embeddings, we measured the cosine distance for all samples relative to the original real samples. Figure 1 (y-axis) shows that the shift in the embedding space varies by tool. Certain processes, like *Acoustica DeClick2*, yield very small cosine distances, reflecting only small impacts on the feature space, while others, like *Adobe Speech Enhance*, introduce more substantial shifts (we note that the fake *Chatterbox* samples are minimally different from their real source files, as the voice conversion source and target files used here were identical). This diversity naturally raises a critical question: Do all AI-based enhancement algorithms pose an equivalent threat to the reliability of deepfake detectors?

As an attempt to answer this, we tested all samples in the latest version of Fauxdio [4], our deepfake detection system. Figure 1 maps the relationship between the resulting deepfake detection scores and the BEATs embedding cosine distances. As expected, the fake samples obtain the highest average score, but processed samples fall into a ‘zone of ambiguity’ between real and fake, meaning that they are likely to impact binary classifier performance.

It can also be seen in Figure 1 that the processing algorithms presenting the greatest risk of misclassification also show the largest shifts in the BEATs embedding space; this motivates the development of a dedicated AI-processing detector to act as a

gatekeeper for deepfake detection. Figure 2 shows the output of a new AI-processed sample detector within Fauxdio. These initial results show that effective discrimination between AI-processed and other samples may be a feasible solution for mitigating against the risks of processed audio in forensic deepfake detection.

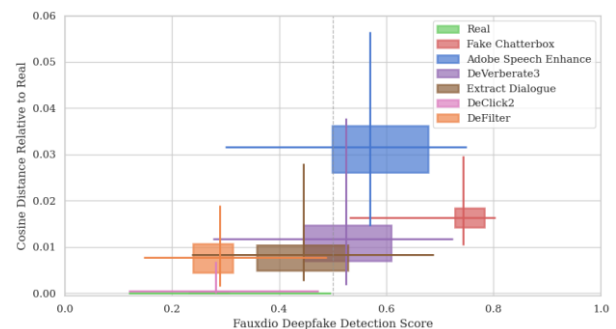


Figure 1: 2D-boxplot of BEATs embedding cosine distances relative to real samples vs. Fauxdio deepfake detection scores (where a score nearer 1 is more likely to be fake)

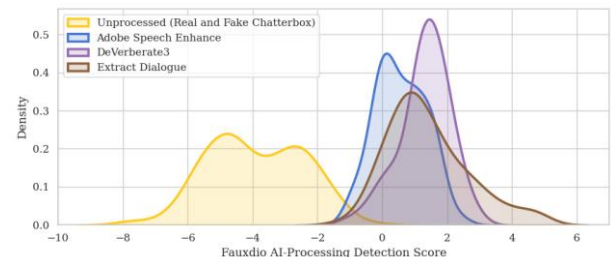


Figure 2: Kernel Density Estimate (KDE) of Fauxdio AI-processing detection scores.

Generative AI Statement

Gemini 3 assisted in writing initial drafts of this work. The authors take full responsibility for its content.

1. References

- [1] S. H. B. Satish, H. Lameris, J. Gustafson, and É. Székely, “What counts as real? Speech restoration and voice quality conversion pose new challenges to deepfake detection,” *under review at Interspeech 2026*, arXiv preprint:2603.14033, 2026.
- [2] J. Virji, F. Kelly, and A. Alexander. Speaker-specific speech transcription for forensic investigative triage, *UK and Ireland Speech (UKIS) Workshop 2025*, York, UK.
- [3] A. Pianese, D. Cozzolino, G. Poggi, and L. Versoliva, “Training-free deepfake voice recognition by leveraging large-scale pre-trained models,” *2024 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec ’24)*, Baiona, Spain, 2024.
- [4] A. Alexander, L. Gerlach, T. Coy, O. Forth, L. Loneragan, and F. Kelly, “FAUXDIO: An audio deepfake detector for law enforcement and forensics,” *Int. Assoc. Forensic Phonetics Acoust. (IAFPA) Conf. 2025*, Den Haag, Netherlands, 2025.