Forensic phonetic methods in authentic and synthetic sample comparison

Jessica Wormald, Ben Gibb-Reid & Vincent Hughes
Forensic Speech Services (FoSS) University of York; Department of Language and Linguistic Science, University of York

The task of forensic voice comparison involves contrasting the speech patterns of one or more questioned recordings with one or more recordings of a known suspect. This is with the aim of establishing the strength of the evidence assuming the samples were i) made by the same speaker, or ii) made by different speakers (Leemann et al., 2024). The auditory-acoustic and automatic methods used to achieve this task are widely discussed and increasingly are subject to validation testing (see Morrison, 2014). With the increasing development and availability of audio deepfakes (also known as 'voice clones'), a different but related question is emerging in forensic speech science: was a questioned recording produced by a **known speaker**, or was it a **synthetically produced deepfake**? Forensic laboratories currently lack robust processes to reliably detect deepfakes, especially across diverse accents and given the rapid evolution of voice-cloning technology. This project addresses the urgent need for research and the development of science-based complementary human and machine methods to provide forensic experts with robust explainable procedures and a framework for ongoing research. Most existing research on deepfake detection relies solely on machine-based detection (Almutairi & Elgibreen, 2022). The present work instead builds towards a procedure which incorporates complementary human analytical techniques, relying on existing forensic phonetic features.

We have collected a dataset to identify linguistic and signal features that are the most effective for determining whether a sample is a deepfake when a reference sample of the speaker is available. Authentic recordings of 16 speakers were made performing three short monologic tasks: a series of map tasks, a picture description task and a reading task. Four different varieties of English are assessed: southern standard British (SSB), Yorkshire, Scottish, and Hong Kong English (HK) – each with four different speakers. *ElevenLabs* was used to create voice clones and produce two types of audio deepfake recordings: text-to-speech (TTS) and voice conversion (VC). We used *Eleven Labs* as it is a widely-used, non-specialist and accessible system – making it a potential method for nefarious use. The 'target' for the synthetic voice will be each of the 16 authentic recordings. For TTS, we will provide the same written text describing the map task. For VC, we plan to perform conversions of each combination of the 16 speakers (e.g. *Voice_1* converted to *Voice_2* and *Voice_2* converted to *Voice_1* etc.). Our ultimate goal is to establish a method combining human and machine-based approaches, aligning deepfake detection with existing speaker comparison and authenticity analysis methods, and exploring qualitative and quantitative approaches to evidence presentation.

References

Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, *15*(5), 155.

Leemann, A., Perkins, R., Buker, G. S., & Foulkes, P. (2024). *An Introduction to Forensic Phonetics and Forensic Linguistics*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780367616595/introduction-forensic-phonetics-forensic-linguistics-adrian-leemann-ria-perkins-grace-sullivan-buker-paul-foulkes

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, *54*(3), 245–256.