Forensic speech science needs forensic phoneticians

Paul Foulkes
Department of Language and Linguistic Science, University of York

Forensic speech science (FSS) is a rapidly evolving field. The most prominent goal of FSS is to establish how individual voices differ from one another. This is of vital importance in forensic speaker comparison cases, which typically require analysis of a voice recorded during a crime (e.g. bomb threats) and comparison with the voice of a suspect (usually recorded in police custody). Comparison of voices is far from straightforward, as anyone trained in phonetics will understand: the voice is highly plastic, affected by many factors relating to the individual (e.g. speech style, health, emotion) and the environment (e.g. background noise, telephone transmission). In short, no two speech samples are ever identical. Voice is therefore an imperfect biometric; voices are certainly distinctive and act as a marker of identity, but, unlike DNA or fingerprints, there is nothing in a voice that is indelible or immutable.

One consequence of this fact is that there is no single method for voice analysis that is universally adopted. Methods of conducting voice analysis have evolved via two largely separate traditions. In Europe FSS is largely grounded on phonetics. Standard analytic methods such as vowel formant and f0 analysis are applied to forensic recordings. In the US, by contrast, the approach is largely informed by engineering and computer science, from which have developed automatic speaker recognition (ASR) systems (similar technology is used for non-forensic purposes e.g. in speech operated tools such as Alexa). ASR technology is making rapid advances, as is evident from its increasing presence in our everyday lives. In controlled experiments ASR systems can now attain almost perfect performance in classifying voice samples by speaker.

However, ASR is not a perfect solution to FSS problems. In my view it never will be. There remain a number of difficulties in using ASR in forensic cases. These include:

1. Applicability. ASR is driven by commercial interests, not forensic ones. Even the best systems struggle to handle non-standard speakers and speech, especially the kinds of materials typically available in forensic cases. These are often of poor technical quality, involve speakers who are stressed or emotional, and there may be considerable variability within a speech sample. Recordings may also be very short. These would therefore likely be rejected as unsafe for analysis in ASR systems.
2. Transparency. ASR systems remain ‘black boxes’ – exceptional in performance, but with no one quite sure why. That is, it is not clear how the features extracted via ASR systems map onto the concrete vocal and linguistic features understood by phoneticians. What is it about the voice that is being picked out and classified as similar and unusual? This is an issue of critical importance in the delivery of justice: it is a principle that evidence in a legal case must be fully transparent.

3. Evaluation. A key step in forensic voice analysis, both by ASR and phoneticians, is that we must judge the typicality of the observed features against a background population to assess whether the voice is unusual or unremarkable. This background population is defined by the characteristics of the offender, i.e. the person whose identity is in question. We therefore face an inevitable paradox: if the speaker's identity is unknown, so is the population from which he or she is drawn. But ASR systems tend to work with populations defined simply by place and language (e.g. 'Swedish', 'UK English').

The two separate traditions of FSS have so far made only limited moves towards integration. In this talk I hope to show why that integration is beneficial. Drawing on real case materials and empirical research I will argue that phonetic methods offer an essential complement to ASR systems to address these three issues.

1. Applicability. Phonetic methods can be used profitably with short or poor quality materials, and to identify the sociolinguistic factors that need to be addressed to better train ASR systems.
2. Transparency. Phonetics can help clarify how ASR results map to vocal features. Ongoing research explores the relationship between (i) ASR features and phonetic features, and (ii) relative performance of ASR and phonetic analysis on the same materials.
3. Evaluation. Phonetic analysis is crucial in establishing the background population, including uncertainty over how to delimit it.

In summary, I aim to show that, in an era of tremendous technical advance, forensic speech science still needs forensic phoneticians.