Attempts to understand the linguistic bases of automatic speaker recognition technology

Vincent Hughes
Department of Language and Linguistic Science, University of York

Analysis of the voice is often used as a form of forensic evidence in legal cases around the world. Most commonly, experts are asked to compare recordings of an unknown criminal and a known suspect. This is called forensic voice comparison (FVC). The issue is one of identity: is the voice of the unknown criminal the same as the known suspect? It is the role of the expert of analyse the recordings and evaluate the evidence under the competing propositions of the prosecution and defence. The trier-of-fact (judge or jury) is then responsible for weighing this, along with the other evidence in the case, to arrive at a decision about the innocence or guilt of the defendant.

In many jurisdictions, linguistic-phonetic analysis is used by experts in FVC cases. This involves the application of standard auditory and acoustic analysis techniques to a range of linguistic variables to assess the similarity between the voices, and to evaluate the typicality of the voices with reference to the wider population – strength of evidence is dependent on knowing how common or rare certain patterns are in the population; for instance, finding /h/-dropping in your two evidential recordings is unlikely to have much probative value given how ubiquitous /h/-dropping is in varieties of British English. Increasingly, automatic speaker recognition (ASR) systems, developed within the field of speech technology, are also being used in FVC casework – albeit not currently in the UK. ASR systems have many benefits: they don't rely on human interpretation, can process thousands of recordings extremely quickly, and, under certain conditions, generate extremely low error rates.

However, ASR systems work in a fundamentally different way from linguistic-phonetic analysis, and this, in part, has led to a perception that the systems are *black boxes*, where the inner workings are opaque to the user. As state-of-the-art systems begin to integrate machine learning through deep neural networks into the work flow, the *black box* nature of the systems is likely to get worse.

In this talk, I will give an overview of how current ASR systems work and provide some background on the status of ASR evidence around the world. I will then present research at the intersection of linguistics and speech technology that has attempted to address the following questions:

- What linguistic information is actually being captured by ASR systems?
- What makes a speaker easy or difficult for an ASR system?
- Are we able to predict how an ASR system will perform under specific conditions with specific types of voices?

The implications for FVC and for other applications of speaker recognition technology will also be considered.