**Assessing the speaker recognition performance of humans and machines: implications for forensic voice evidence**

Vincent Hughes (vincent.hughes@york.ac.uk)
*Department of Language and Linguistic Science*

Forensic voice comparison (FVC; [1,2]) typically involves the analysis of recordings of unknown (often a criminal e.g. in a bugged car or threatening telephone call) and known speech samples (usually from a police interview), to help establish whether they contain the voice of the same individual or not. Such analysis is considered a form of forensic evidence, requiring expertise beyond that which can reasonably be expected from lay people. Around the world, this expert analysis is typically conducted using linguistic-phonetic methods (i.e. detailed auditory transcription combined with acoustic measurements) and/or automatic speaker recognition systems (i.e. a piece of software that processes and compares voices statistically within relatively little human input).

However, the speaker recognition ability of non-expert listeners in the FVC context is also an important issue for two key reasons. Firstly, lay people in many jurisdictions sit on juries which act as the trier-of-fact in criminal trials. It is their responsibility to make a judgement about the overall innocence or guilt of the accused. In cases involving FVC evidence, samples are often played to the jury and so jurors are necessarily making judgements about the voice. Juries may also be presented with expert FVC evidence either in the form of oral testimony or a written report, but little is known about how juror judgements override or interact with this information, or how this expert evidence is understood. Secondly, there is growing focus within the field of FVC on the complementarity of different methods, better understand what speaker-specific information is being captured. Much of this work has focused on the combination of automatic and linguistic methods [3,4,5], but little is known about how these methods compare with listener judgements (see [6,7] although these are not forensic in focus).

In this talk I will introduce a new project at the University of York called *Humans and Machines: Novel Methods for Testing Speaker Recognition Performance* (UK Arts and Humanities Research Council funded; AH/T012978/1). This is some of the first work within the field of FVC to elicit statistical responses from human listeners which can then be compared and combined with the results of an automatic speaker recognition system. The project also examines how listener judgements are affected by sources of cognitive bias commonly encountered in FVC; namely the introduction of expert opinion and other types of forensic evidence. I will discuss the considerable challenges of eliciting such statistical responses from human listeners and describe a bespoke game-based experiment which positions participants as members of a jury as a way of understanding listener behaviour in the courtroom. I will also present some initial results comparing human and machine performance using forensically realistic, channel-mismatched speech samples.

[1] Foulkes, P. and French, J. P. (2012) Forensic speaker comparison: a linguistic-acoustic perspective. In P. Tiersma and L. Solan (eds.) *Oxford Handbook of Language and Law*. Oxford: OUP. pp. 557-572.
[2] Jessen, M. (2019) Forensic voice comparison. In M. Rathert and J. Visconti (eds.) *Handbook of Communication in the Legal Sphere*. Berlin: de Gruyter Mouton. pp. 219-255.
[3] Enzinger, E., Zhang, C. & Morrison, G. S. (2012). Voice source features for forensic voice comparison - an evaluation of the GLOTTEX software package. *Proceedings of Odyssey: The Speaker and Language Recognition*

*Workshop,* Singapore, pp. 78-85.
[4] González-Rodríguez, J., Gil, J., Pérez, R. & Franco-Pedroso, J. (2014). What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, pp. 33-40.
[5] Hughes, V., Harrison, P., Foulkes, P., French, J. P., Kavanagh, C. and San Segundo, E. (2017) Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of Interspeech – 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 3892-3896.
[6] Park, S. J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P. A. and Alwan, A. (2018) Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of difference speech styles. *Journal of the Acoustical Society of America* 144(1): 375-386.
[7] Afshan, A., Kreiman, J. and Alwan, A. (2020) Speaker discrimination in humans and machines: effects of speaking style variability. *Proceedings of Interspeech – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, pp. 3136-3140.