Abstract to be presented at NWAV50 (13–15 Oct 2022)

Legal context biases listeners toward hearing voice pairs as more similar

Vincent Hughes, Carmen Llamas & Thomas Kettig

University of York

Listeners' judgements about voices have been demonstrated to be affected by information involving a legal context. For example, Lange et al. (2010) found that listeners were more likely to misinterpret ambiguous words in degraded samples as incriminating when told they came from "criminal suspects' interviews". As recordings of voices are commonly played to members of juries and used as evidence in criminal cases, understanding the effects that the context of the court has on listener judgements is of clear theoretical and applied importance.

Here we report initial results from a new project, *Humans and Machines: Novel Methods for Testing Speaker Recognition Performance*, which sets out to create a bespoke computer game to elicit listeners' judgements on the similarity of forensically-realistic pairs of very short (10 second) samples of speech. Listeners are asked to rate the pairs on a 0-100 scale for how similar the voices are as well as whether the samples belong to the same speaker. We also elicit ratings of how typical each speaker sounds as a member of the speech community in question. The samples are taken from Standard Southern British English, a variety that UK listeners are likely to be familiar with, and two varieties from the Northeast of England (Newcastle and Middlesbrough), the latter of which is unlikely to be familiar to most. Participant background and self-ratings of familiarity with each accent were also taken into account when evaluating their judgements.

The context of the game is the 'Jury of the Future' wherein the motivation is to 'beat the machine' in a situation where AI is responsible for forensic analysis in the justice system. The jury context is only used in level 2 of the game when participants are graphically and auditorily immersed in a futuristic GUI and tasked by a robot judge to conduct voice comparisons. Level 1 is run as a training level and no legal context is hinted at. An initial set of 18 sample pairs (3 same-speaker + 3 different-speaker pairs x 3 accents) have been tested in the game with 179 UK-born L1 English speakers aged 19 to 78 recruited via Prolific. Several linear regressions were fit and pairwise estimated marginal means were then calculated.

We find significant pairwise increases in ratings from training to jury context in sameness, similarity, and typicality. We also find an order effect in ratings of sameness but not similarity or typicality. We interpret these findings as indicating a legal context bias effect in which speakers' default judgements of voice similarity may shift when a comparison is contextualised in a courtroom context. A problematic consequence of this shift is that naive listeners may overestimate the similarity of unfamiliar different-speaker pairs when presented as legal evidence, in turn leading to an increase in false positives.