Legal context biases listeners toward hearing voice pairs as more similar

Vincent Hughes, Carmen Llamas, and Thomas Kettig

{vincent.hughes|carmen.llamas|thomas.kettig}@york.ac.uk

NWAV50
San Jose, CA
October 2022

digital creativity labs

UKRI Arts and Humanities Research Council

Joe Cutting

Daniel Slawson

digital creativity labs

Arts and Humanities Research Council

Humans & Machines: Novel methods for assessing speaker recognition performance

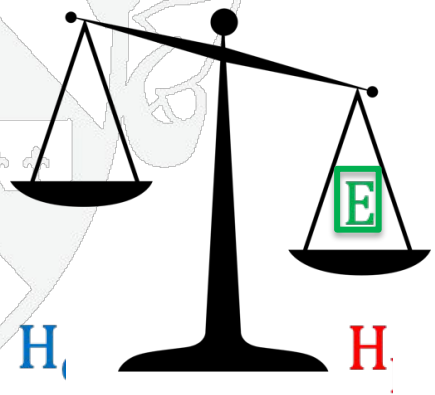
(AH/T012978/1)

https://sites.google.com/york.ac.uk/humans-machines/

Speaker comparison

- Forensic phoneticians express the strength of voice evidence via a Likelihood Ratio (LR):

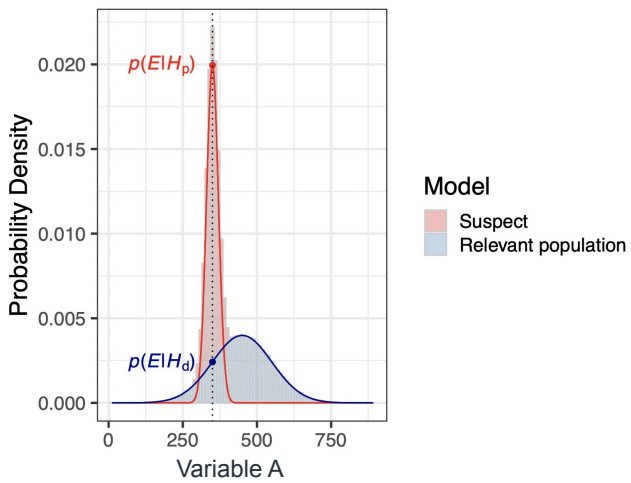
$$\frac{\textbf{Similarity} - \text{“Prosecution hypothesis”}}{\textbf{Typicality} - \text{“Defense hypothesis”}}$$



$$\frac{p(E|H_p,}{p(E|H_d,}$$

Speaker matching

- ASR systems compare these measures using available data



$$LR = \frac{0.02}{0.0025}$$

$$= 8$$

We want to test how this might work with human listeners

RQs for project:

(1) Human performance vs. ASR performance

(2) Effect of listener group/biases on human performance

(3) Effect of sample type on human performance/ASR performance

(4) Effect of contextual information on human performance

Method: Immersive jury game

- Participants encounter pairs of sound samples
  - Typicality rating elicited for first stimulus
  - Similarity and sameness ratings elicited after second stimulus presented
  - First gave us self-declared accent familiarity ratings / demographic info
- Two levels discussed here
  - Tutorial level without context
  - Jury/legal context

Stimuli

- Samples from **Standard Southern British English** (DyViS) & **Newcastle** and **Middlesbrough** men (TUULS)
- Forensically-realistic quality
  - First sample = landline phone quality (actual or noise/filter added)
  - Second sample = HQ, taken from mock police interviews
  - Short (10-11s)
- Tagged for accentedness and voice quality
- Normed for guilt/suspiciousness of sample content

Stimuli

- 120 pairs created
  - 30 SSBE pairs (15 DS, 15 SS)
  - 30 Middlesbrough pairs (15 DS, 15 SS)
  - 30 Newcastle pairs (15 DS, 15 SS)
  - 30 mixed Middlesbrough/Newcastle pairs (30 DS)
- Distributed into 15 blocks containing 8 pairs each (5 DS, 3 SS)
- Stimuli within blocks internally randomized
- 1 block presented per level, counterbalanced
- 896 participants

Tutorial

Follow the instructions below

Level 1: Comparison 1 of 1

Listen to the clip and answer using the slider below

AUDIO CLIP 1

USER OPINION

This is a Middlesbrough speaker. How typical is this voice relative to other speakers of the same accent?

60%

0% 100%

Submit

Tutorial

Follow the instructions below

Level 1: Comparison 1 of 1

Listen to both clips and answer using the sliders below

AUDIO CLIP 1

USER OPINION

How similar are the two voices?

72%

0% 100%

AUDIO CLIP 2

USER OPINION

Do these voices belong to the same speaker?

21%

0% 100%

Submit

JURY OF THE FUTURE

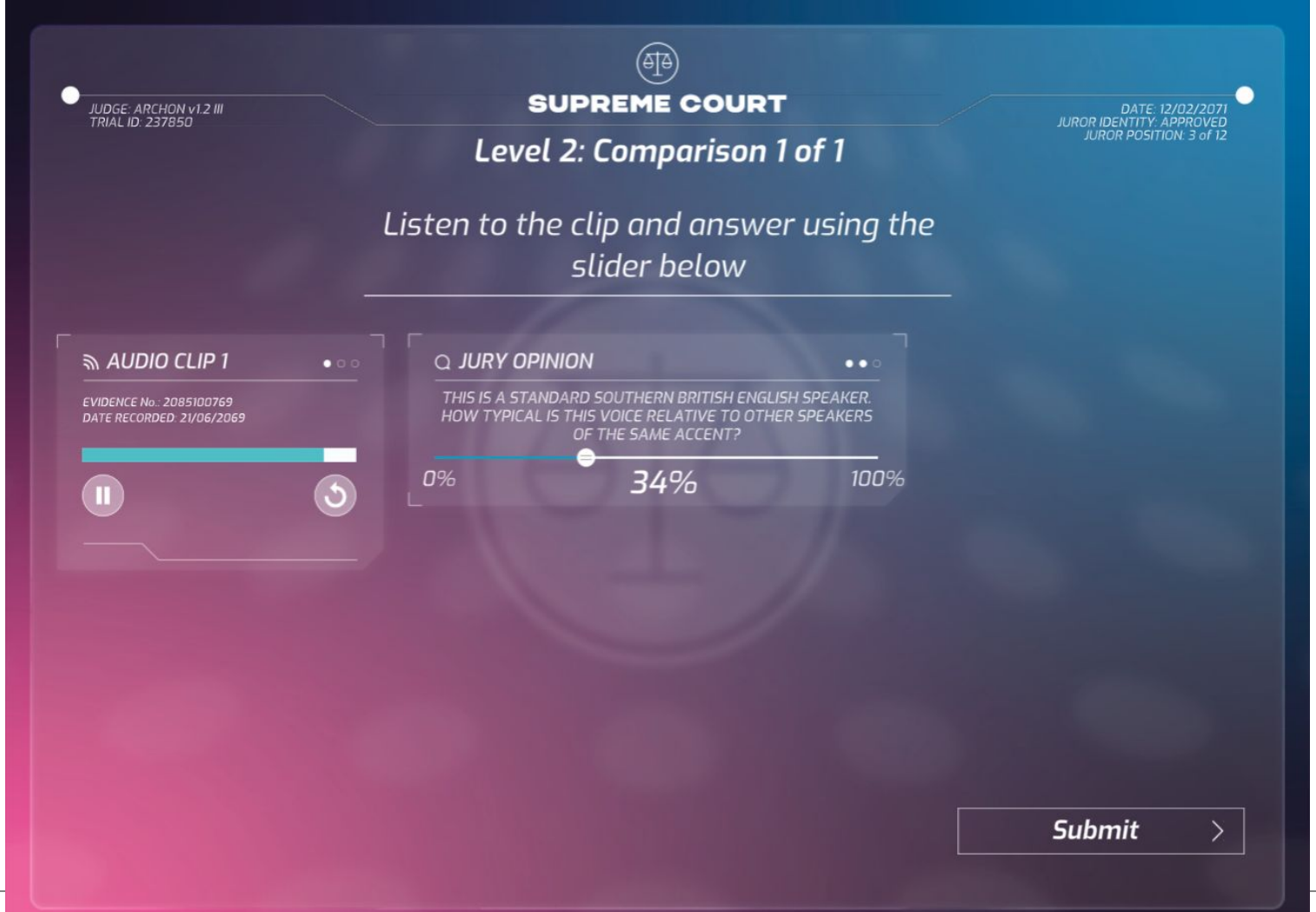
It's 2071. The justice system is now delivered by machines.

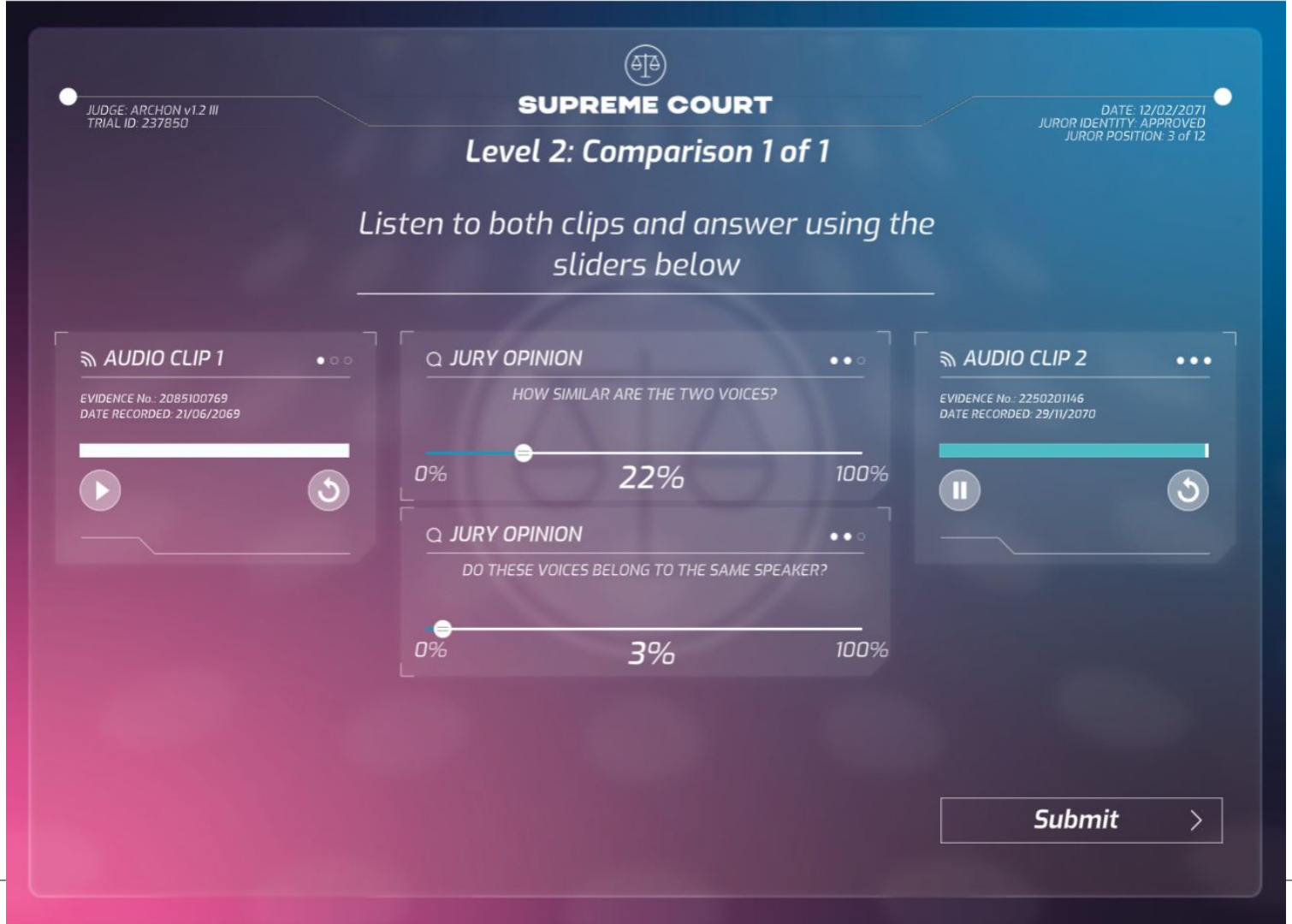
The robots are highly efficient, but errors have recently come to light.

A new trial is being held to test bringing back human juries.

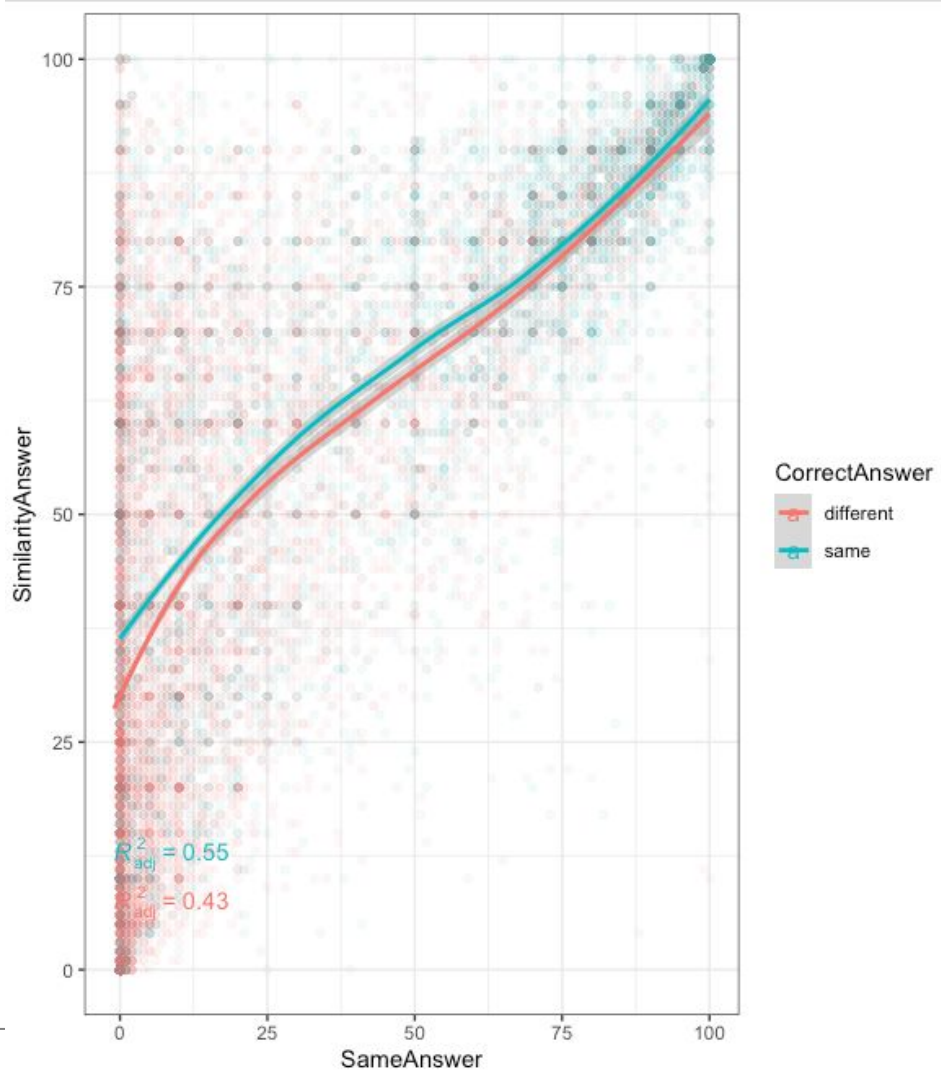
Can you beat the machine and prove that justice belongs in human hands?


09:01
LIVE
LONDON
EXCLUSIVE
HUMANS v MACHINES?
This groundbreaking trial could be the pathway back to human-only juries following several high-profile robot jury mistakes.
Next ▶
Tuesday 9th March 2071


SUPREME COURT
JUDGE: ARCHON v1.2 III
TRIAL ID: 237850
DATE: 12/02/2071
JUROR IDENTITY: APPROVED
JUROR POSITION: 3 of 12
Level 2: Comparison 1 of 1
Listen to the clip and answer using the slider below
AUDIO CLIP 1
EVIDENCE No.: 2085100769
DATE RECORDED: 21/06/2069
JURY OPINION
THIS IS A STANDARD SOUTHERN BRITISH ENGLISH SPEAKER. HOW TYPICAL IS THIS VOICE RELATIVE TO OTHER SPEAKERS OF THE SAME ACCENT?
0%
34%
100%
Submit

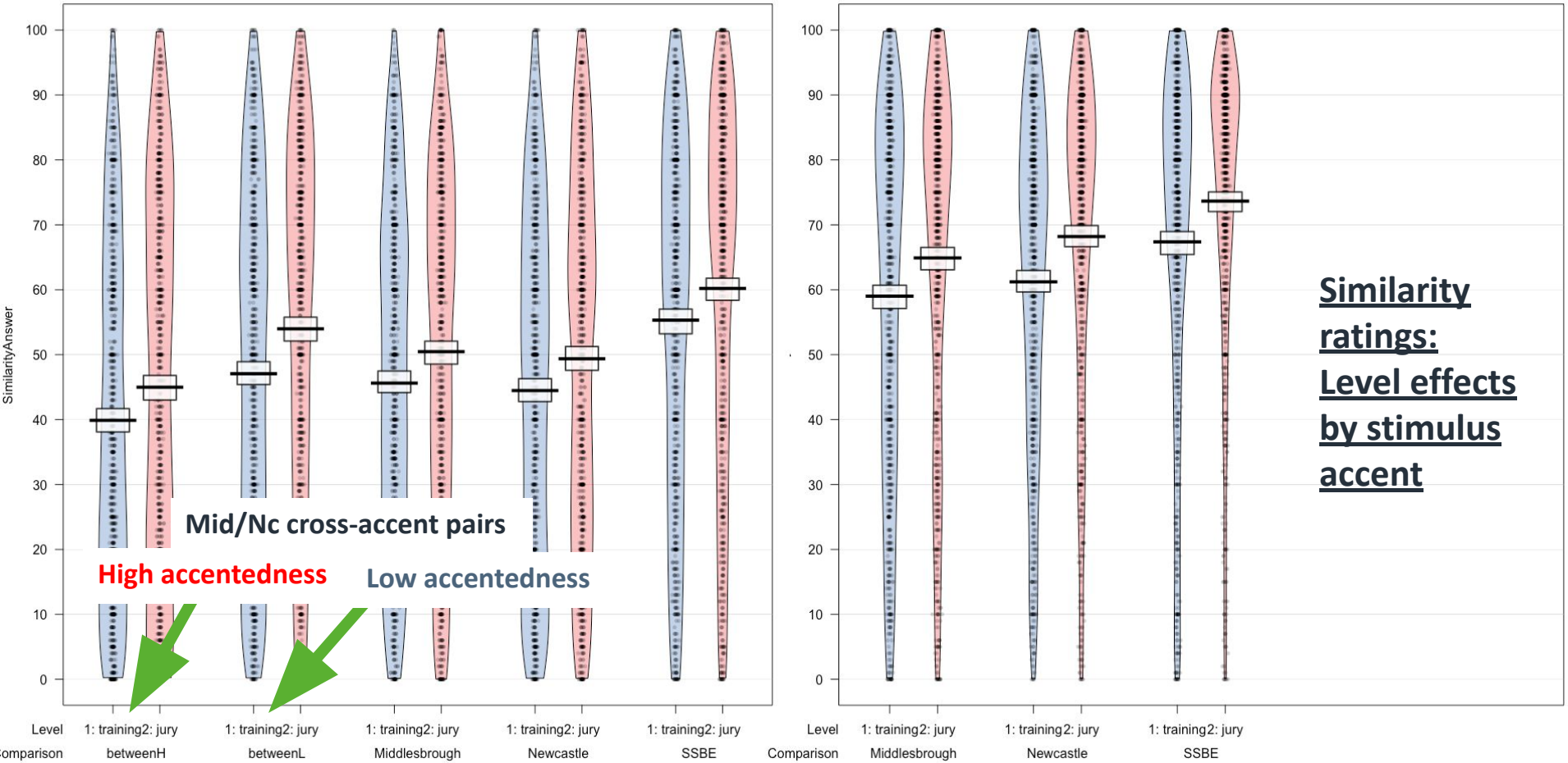

SUPREME COURT
JUDGE: ARCHON v1.2 III
TRIAL ID: 237850
DATE: 12/02/2071
JUROR IDENTITY: APPROVED
JUROR POSITION: 3 of 12
Level 2: Comparison 1 of 1
Listen to both clips and answer using the sliders below
AUDIO CLIP 1
EVIDENCE No.: 2085100769
DATE RECORDED: 21/06/2069
JURY OPINION
HOW SIMILAR ARE THE TWO VOICES?
0%
22%
100%
JURY OPINION
DO THESE VOICES BELONG TO THE SAME SPEAKER?
0%
3%
100%
AUDIO CLIP 2
EVIDENCE No.: 2250201146
DATE RECORDED: 29/11/2070
Submit

Sameness ratings vs. similarity ratings



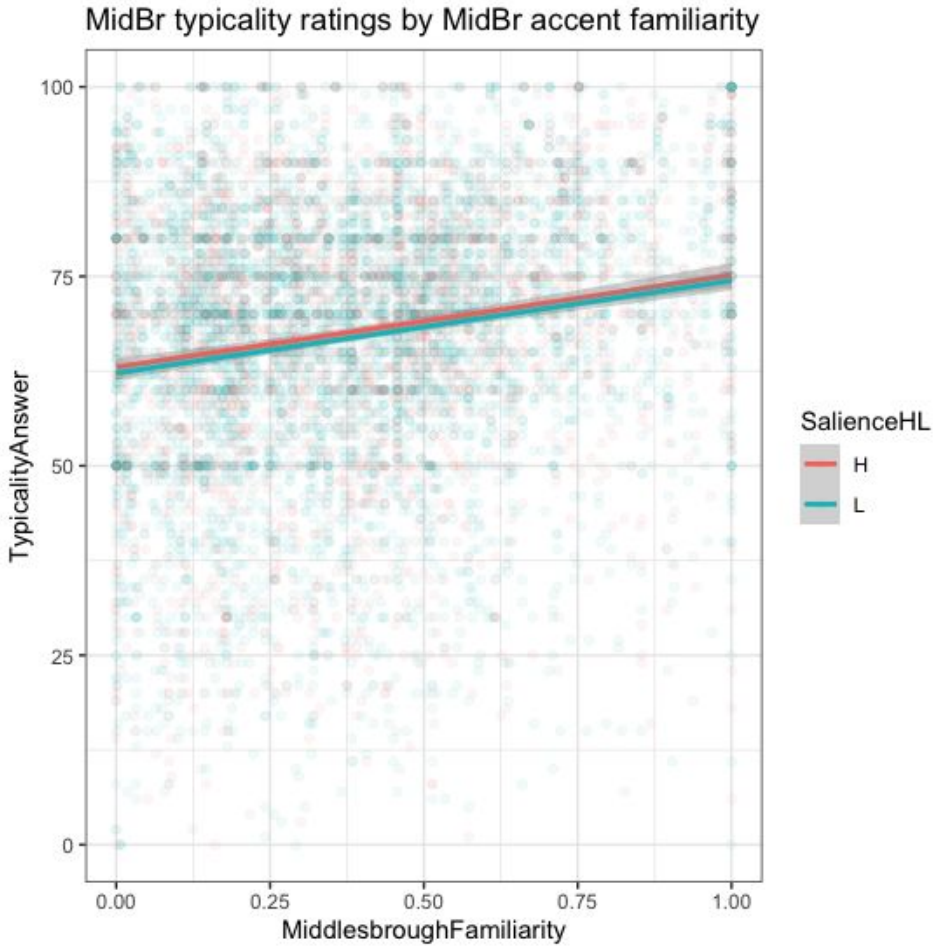
Different-speaker pairs

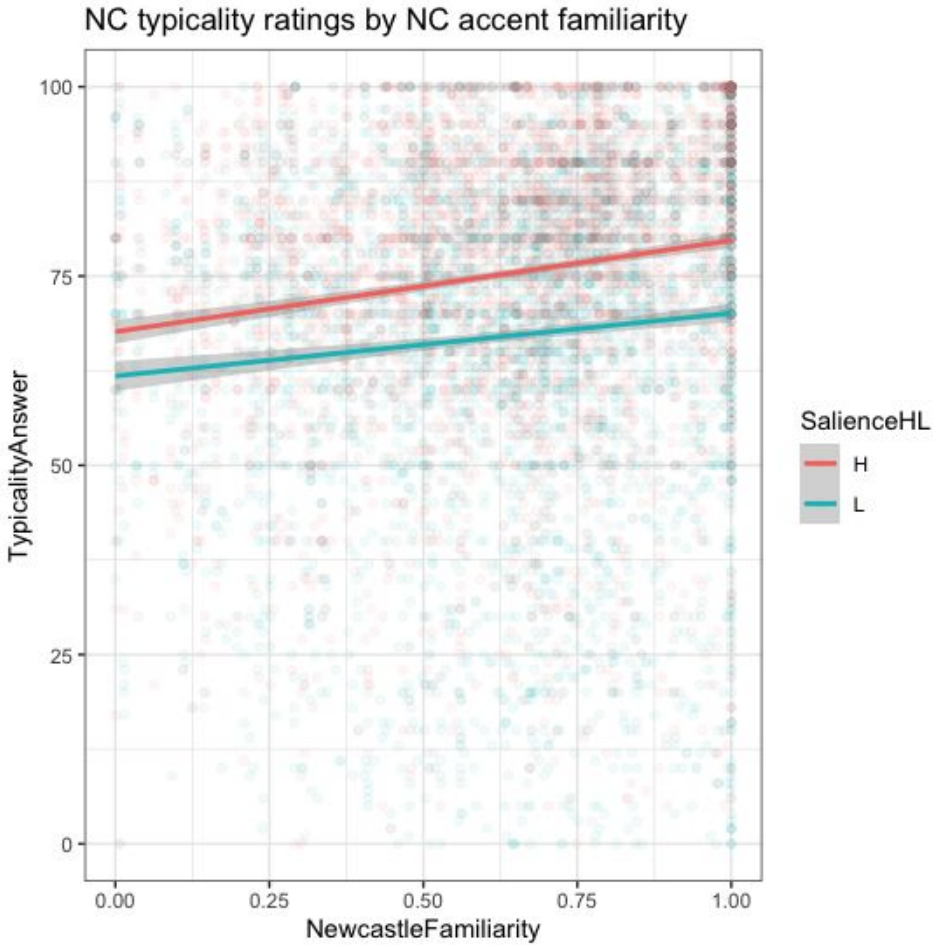
Same-speaker pairs



Similarity ratings: Level effects by stimulus accent

Typicality ratings by Northern accent familiarity and accentendess





Conclusions

- On both sides of the similarity/typicality likelihood ratio equation, potential consequences for jury decisions and forensic applications
- Potential of game-based immersive experimental methods vs. ‘vanilla’ Qualtrics-type ones usually used

digital creativity labs

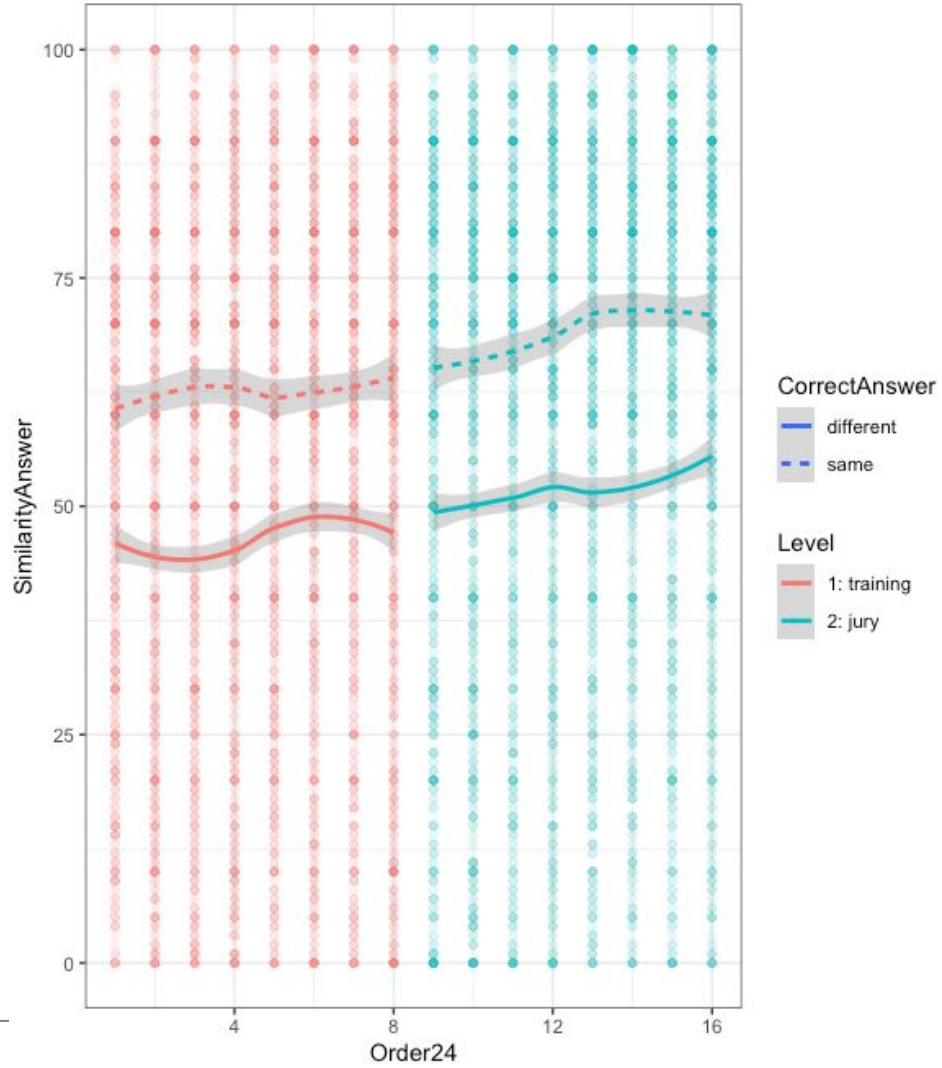
UKRI Arts and Humanities Research Council

UNIVERSITY of York

Thank you!
Questions?

{vincent.hughes|carmen.llamas|thomas.kettig}@york.ac.uk

@VinceH_Forensic @TKettig


100
75
50
25
0
SimilarityAnswer
4
8
12
16
Order24
CorrectAnswer
different
same
Level
1: training
2: jury