An initial empirical analysis of the effect of sampling variability in a forensic voice comparison system

Philip Weber[1,2] -- Aston University, Computer Science | Forensic Data Science Laboratory

UK SPEECH 2023 SHEFFIELD

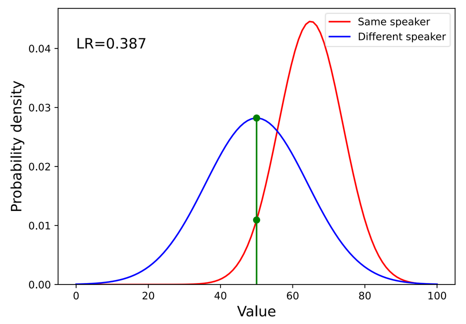
Aston University BIRMINGHAM UK

UKSpeech

Forensic voice comparison

Using **audio data (voice)** to train machine learning and statistical models to provide courts with probabilistic answers about evidence.

Objective, validated systems are **more trustworthy** in court and **perform better** than subjective human-expert reasoning.



Key Issue: specific, relevant data is required for training, adaptation, calibration and validation for each case: **cost ➔ barrier to justice.**

- costly and time-consuming to collect.
- how much data do we need?
- how “close” must it be to the case? how do we measure “close”?

Related Issue:

- validations reported at a sample point (e.g. [1, 2]).
- what are the implications of this? … (see also, e.g. [4]).

Typical process, in probabilistic terms

Presented with case recording(s) $\boldsymbol{s}$ and tasked with producing a likelihood ratio $LR(\boldsymbol{s})$ to inform court about the weight of evidence provided by $\boldsymbol{s}$.

Consider $\boldsymbol{s}$ to be **sampled** from $\boldsymbol{\mathcal{D}}$, the (hypothetical) set of all audio recordings from the relevant population $\boldsymbol{p}$ and in the recording conditions $\boldsymbol{c}$ of the case, according to **unknown distribution $\boldsymbol{Pr(\mathcal{D})}$**.

Analyst and/or system **estimates** $\boldsymbol{p'} \approx \boldsymbol{p}$ and $\boldsymbol{c'} \approx \boldsymbol{c}$, effectively **estimating** $Pr(\mathrm{D}) \approx Pr(\boldsymbol{\mathcal{D}})$. Analyst collects or **simulates** data $\mathrm{D}_{sys} \sim \Pr(\mathrm{D})$.

System is trained & validated using D_{sys} to produce likelihood ratios $LR(\cdot)$ based on assumptions governing $\Pr(\mathrm{D})$: **not** $\Pr(\boldsymbol{\mathcal{D}})$.

Final evaluation $LR(\boldsymbol{s})$ is based on $\boldsymbol{s}$ under $\Pr(\mathrm{D})$: **not** $\Pr(\boldsymbol{\mathcal{D}})$.

Sampling effects affect the goodness of fit of $\Pr(\mathrm{D})$ as a proxy for $\Pr(\boldsymbol{\mathcal{D}})$ and “appropriateness” of $LR(\boldsymbol{s})$ (e.g. accuracy and precision).

Notes: 1. This ignores questions of data use for training different parts of the pipeline, calibration and validation, and non-case-specific data used to train (e.g.) an x-vector extractor. **2.** This is not the same question discussed elsewhere about whether LRs should be reported with confidence intervals. Rather about the machine learning approach, how/what to report, and how to predict, measure and justify the accuracy of the result.

Extending the benchmark [2] and case validations [1] we investigate the effect of **resampling**

1. **speakers** chosen for training/adaptation and calibration/validation.
2. **simulation** of audio conditions for the questioned- and known-speaker recordings.
3. **sub-selection** of audio sections of given duration.

Key result: System is at a point in “sampling space”.

Benchmark *forensic_eval_01* config → optimistic C_{llr}.
Varying samples → considerably varying C_{llr}.
Additional training data → reducing mean C_{llr}, but reduced validation set → increased variance.

Similarly for the case data.

Not a problem for benchmarking, but what impact does it have on reporting findings?

Implication that even more data is needed.

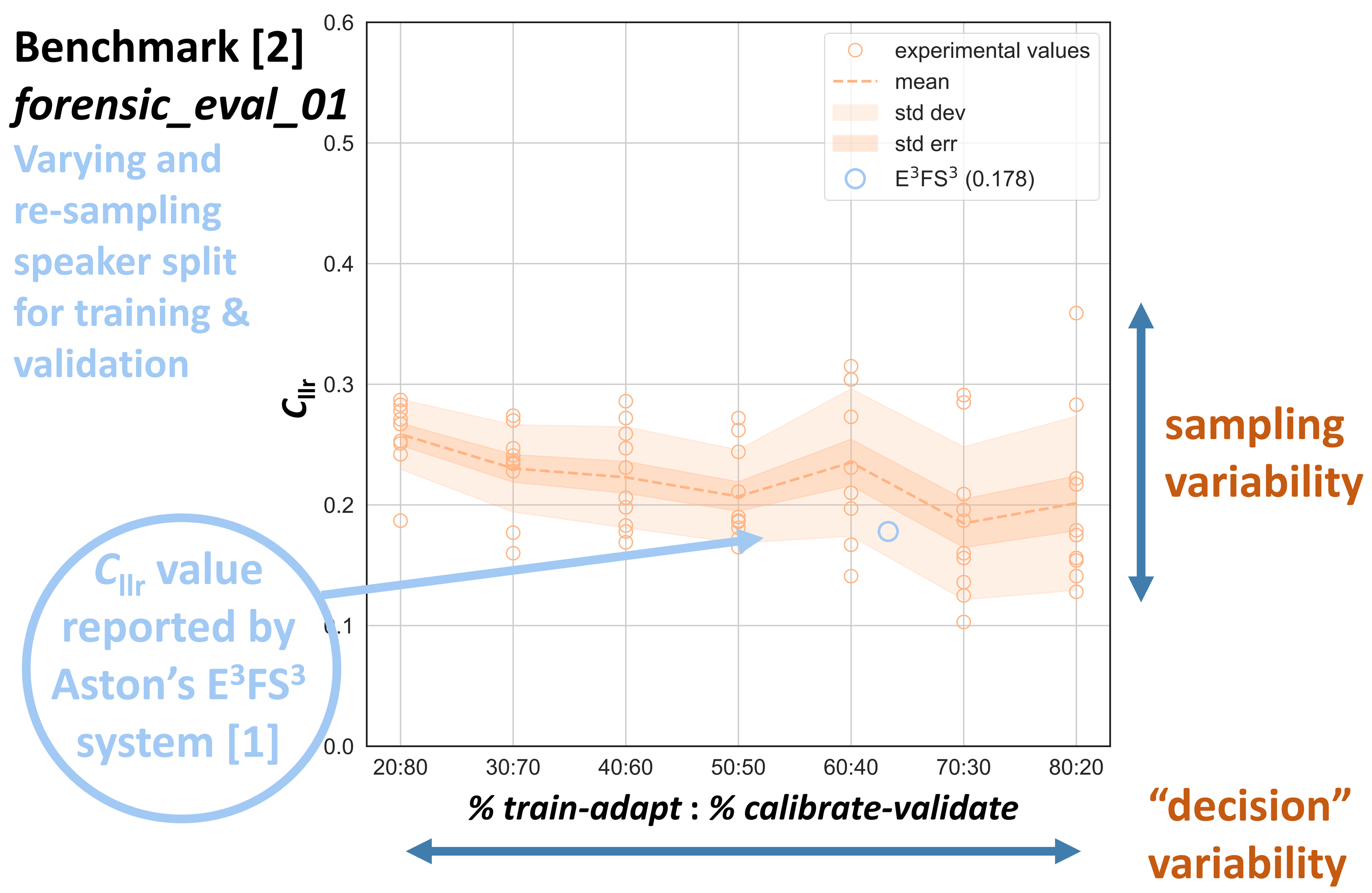
Questions:

What should we report to court? Is a single point-validation adequate? Should we conduct a more thorough validation and provide more information?

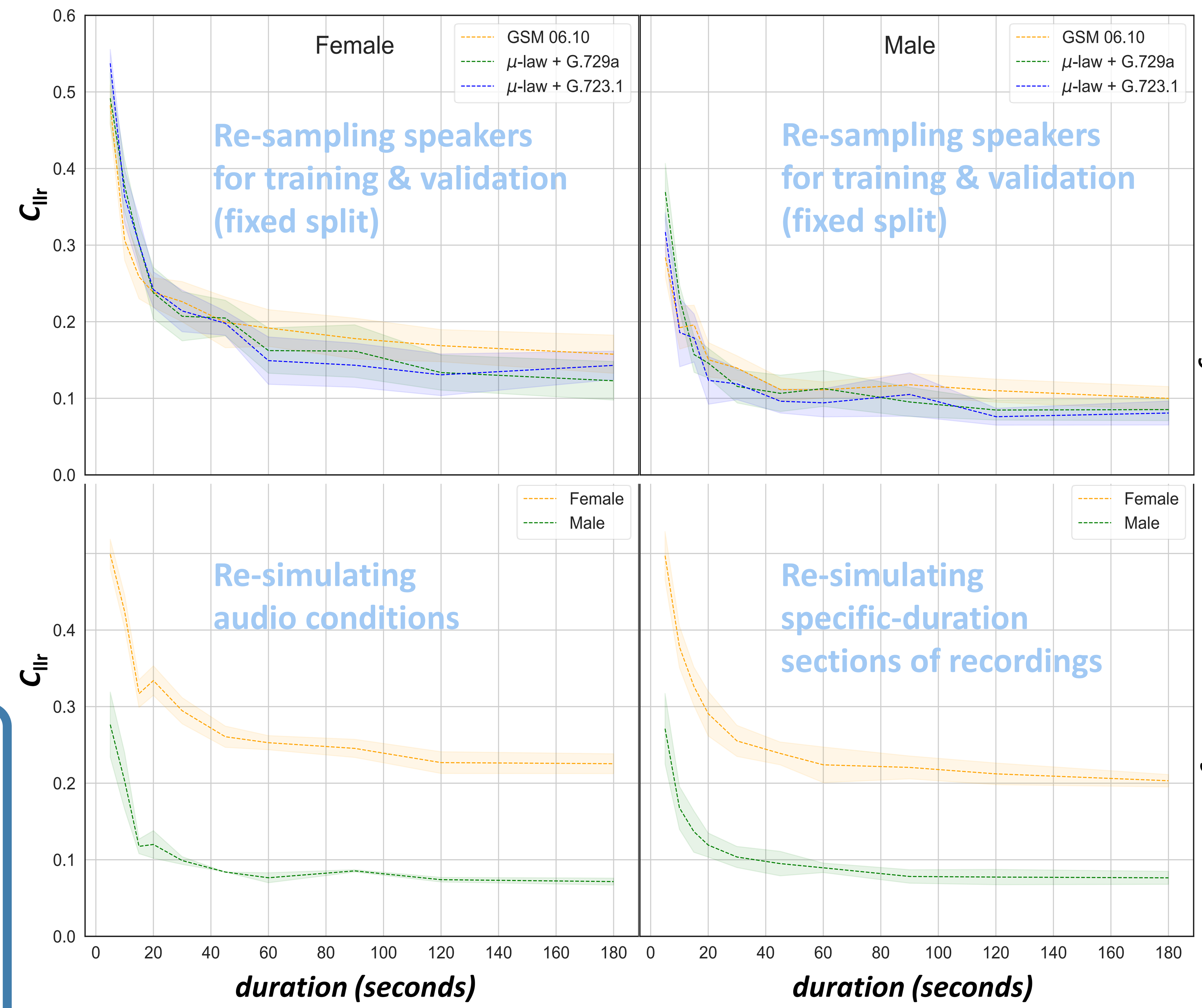
Can we devise metrics for the relevance of audio data? Pre-emptively validate & re-use systems? Predict performance by data volume/quality?

Benchmark [2]
forensic_eval_01

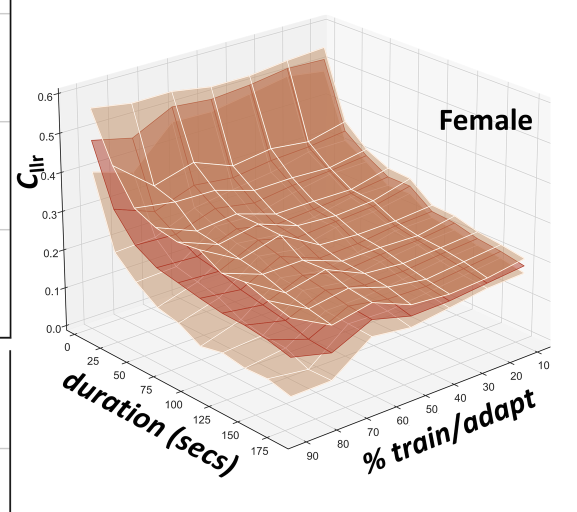
Varying and re-sampling speaker split for training & validation

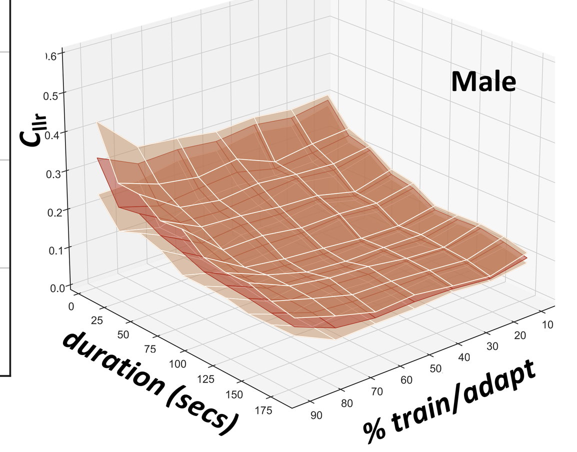


Case (based on AusEng500+)



Effect of varying and re-sampling speaker split for training & validation





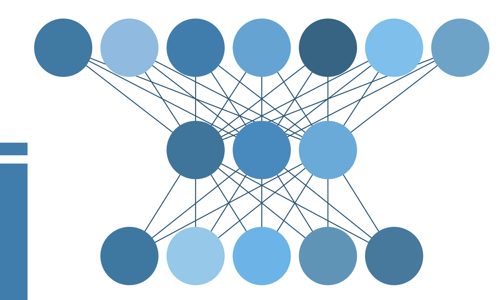
Data: *forensic_eval_01*: landline/interview 166 male Australian English (AusEng) speakers, 646 recordings. Simulated noisy landline phone call and reverberant interview. Benchmark train/validation split.

Case based on AusEng500+: 169 male & 223 female AusEng speakers. Simulated call-centre recordings, codecs & durations.

System: *E³FS³*: is a ResNet – PLDA – Logistic Regression Calibration pipeline; outputs a likelihood ratio. Based on state-of-the-art automatic speaker recognition algorithms.

Metric: C_{llr} penalises errors and lack of confidence distinguishing same-speaker or different-speaker pairs.

C_{llr} = 0: perfect; 1: uninformative; > 1: mis-calibrated.



[1] P. Weber, E. Enzinger, B. Labrador-Serrano, A. Lozano-Díez, D. Ramos, J. González-Rodríguez, G.S. Morrison (2022). **Validation of the alpha version of the E3 Forensic Speech Science System (E³FS³) core software tools.** *FSI: Synergy*, 4, 100223. DOI: 10.1016/j.fsisyn.2022.100223.

[2] G.S. Morrison, E. Enzinger (2016). **Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic eval 01) – introduction.** *Speech Communication*, 85:119–126. DOI: 10.1016/j.specom.2016.07.006.

[3] G.S. Morrison, E. Enzinger, V. Hughes, et al (2021). **Consensus on validation of forensic voice comparison**. Science & Justice, 61(3):299–309, 2021.

[4] B.X. Wang, V. Hughes, P. Foulkes (2022). **The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison**. Speech Communication, 138:38–49.

[1] **Computer Science,** Aston University, B4 7ET, UK
https://www.cs.aston.ac.uk/~weberp1/ | p.weber1@aston.ac.uk

[2] **Forensic Data Science Laboratory (FDSL)**, Aston Institute for Forensic Linguistics (AIFL)
https://forensic-data-science.net/