Dr. Phil Weber

The E3 Forensic Speech Science System (E^3FS^3):

Design principles and validation of core software tools

Forensic Data Science Laboratory

Aston University, Birmingham, UK

http://forensic-data-science.net/

The E3 Forensic Speech Science System (E^3FS^3):

Design principles and validation of core software tools

1. Design principles
2. State of the art technology
3. Validation of core software tools

The E3 Forensic Speech Science System (E^3FS^3):

Design principles and validation of core software tools

1 → 2 → 3

Design principles

State of the art technology

Validation of core software tools

E³FS³ Forensic speech science **system**

Purpose: forensic voice comparison **research** and **casework**

Fundamental:

- Strength of evidence : **numeric likelihood ratio**
- Reduce potential for **cognitive bias**
- Meet **legal admissibility standards**

Basis:

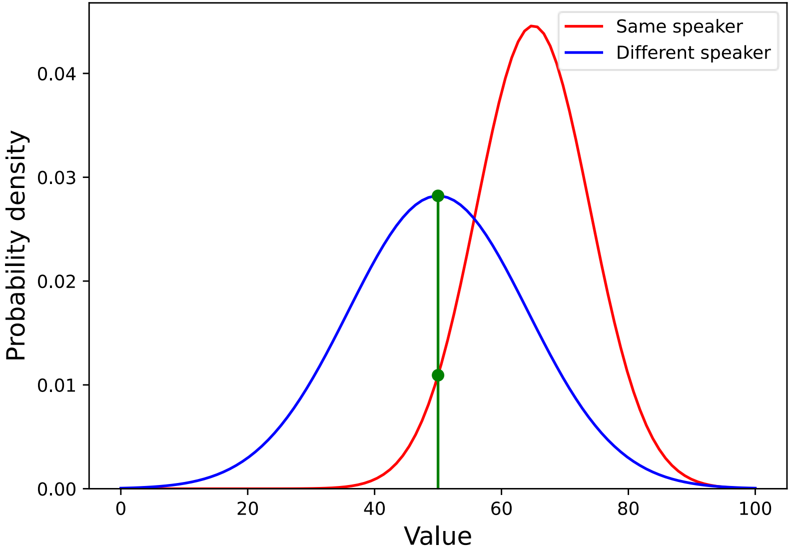
- Relevant **data**
- Quantitative **measurements**
- Statistical **models**

E^3FS^3 Forensic speech science **system**

Purpose: forensic voice comparison **research** and **casework**

Fundamental:

- Strength of evidence : **numeric likelihood ratio**
- Reduce potential for **cognitive bias**
- Meet **legal admissibility standards**



Basis:

- Relevant **data**
- Quantitative **measurements**
- Statistical **models**

E^3FS^3 Forensic speech science **system**

Purpose: forensic voice comparison **research** and **casework**

Fundamental:

- Strength of evidence : **numeric likelihood ratio**
- Reduce potential for **cognitive bias**
- Meet **legal admissibility standards**

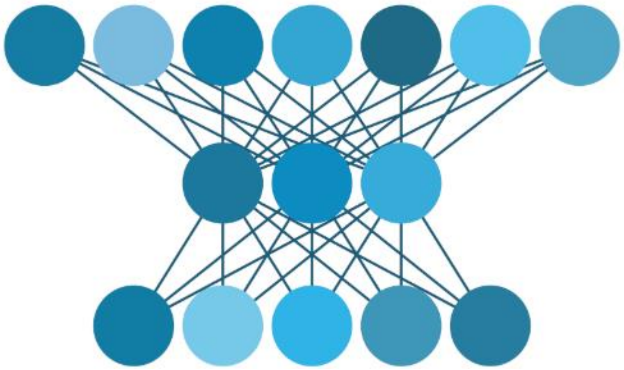
Basis:

- Relevant **data**
- Quantitative **measurements**
- Statistical **models**

E^{3}FS^3 Forensic speech science **system**

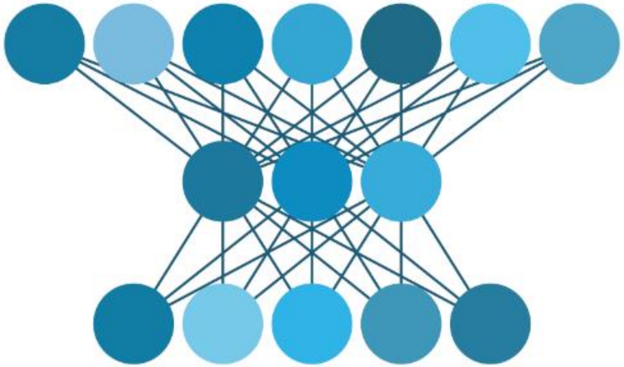
System in the broad sense.

Includes …



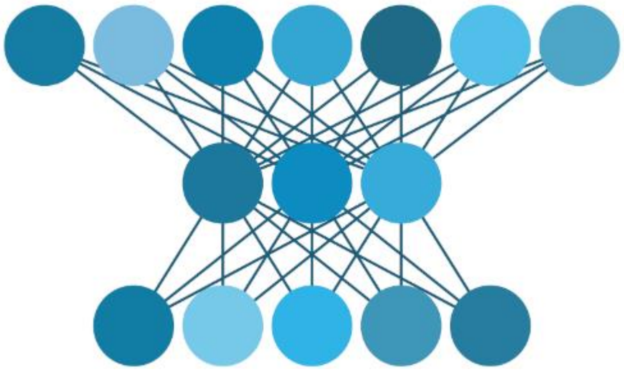
E^3FS^3 Forensic speech science **system**

open-code
software tools



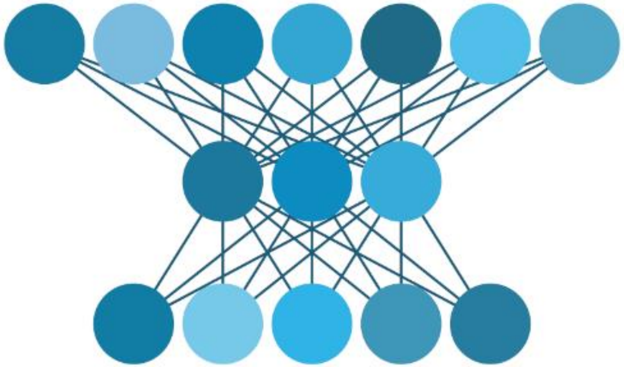
E^3FS^3 Forensic speech science **system**

data collection
protocols



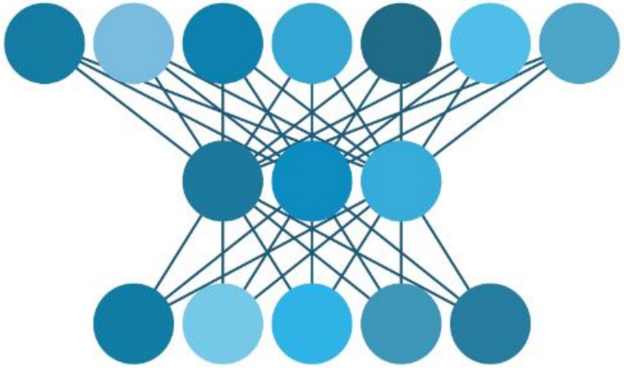
E^{3}FS^3 Forensic speech science **system**

databases
of **relevant**
populations &
conditions



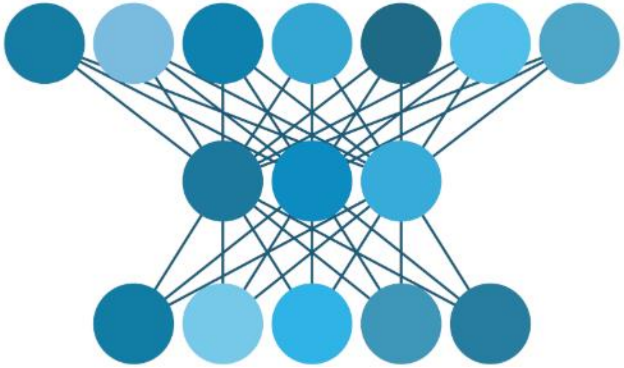
E^3FS^3 Forensic speech science **system**

standards &
guidelines



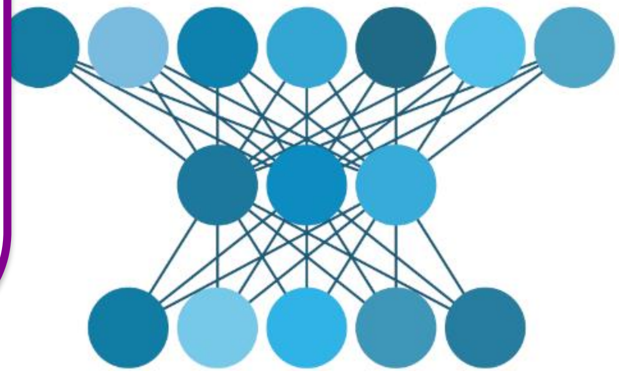
E^3FS^3 Forensic speech science **system**

standard
operating
procedures



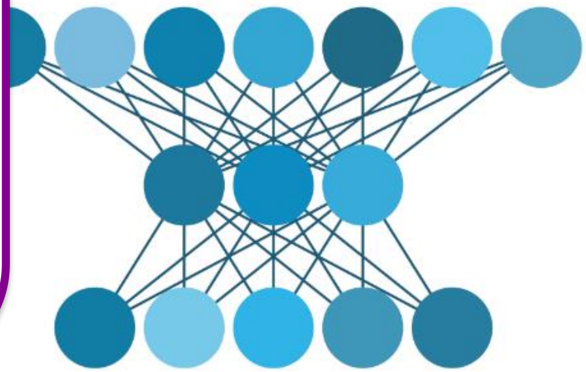
E^{3}FS^3 Forensic speech science **system**

library of validation reports



E^{3}FS^3 Forensic speech science **system**

training
for
practitioners



The E3 Forensic Speech Science System (E^{3}FS^3):

Design principles and validation of core software tools

1 → 2 → 3

Design principles

State of the art technology

Validation of core software tools

E^3FS^3 core software tools

Based on the **state-of-the-art** in Automatic Speaker Recognition:

Answer a **specific question**:

"How likely are we to obtain the **properties** on the questioned-speaker and known-speaker recordings if …

1. they were produced by the **same speaker randomly selected** from the relevant population

vs

2. they were produced by **different speakers randomly selected** from the relevant population?"

E^{3}FS^3 core software tools

Based on the **state-of-the-art** in Automatic Speaker Recognition:

Answer a **specific question**:

"How likely are we to obtain the **properties** on the questioned-speaker and known-speaker recordings if …

1. they were produced by the **same speaker randomly selected** from the relevant population

vs

2. they were produced by **different speakers randomly selected** from the relevant population?"

E^{3}FS^3 core software tools

Based on the **state-of-the-art** in Automatic Speaker Recognition:

x-vector → **fixed-length** representation of a whole recording

produced by a **deep neural network** (machine learning)

statistical modelling → to produce a **score**

calibration → to produce a (calibrated) **likelihood ratio**

"front-end" → **"back-end"** → **calibration**

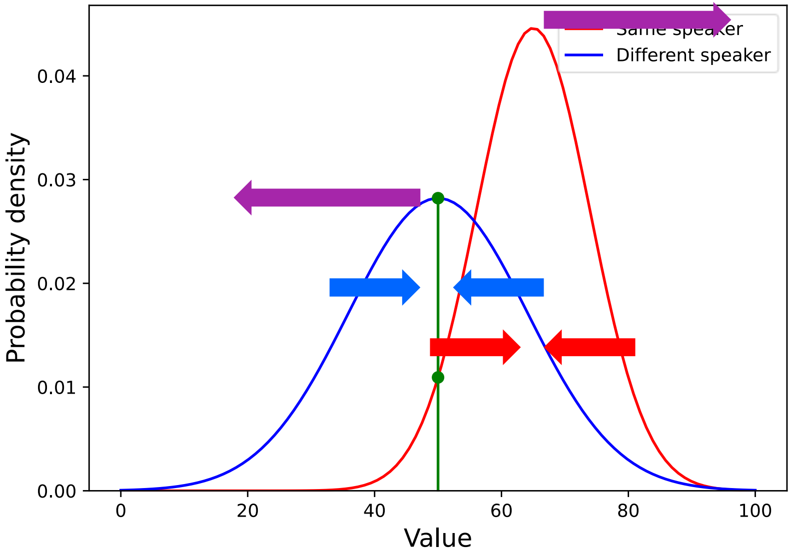
E^{3}FS^3 core software tools

Overall:

account for **between-speaker** variation

ignore **within-speaker** variation

compensate for **recording condition** variation



The E3 Forensic Speech Science System (E^3FS^3):

Design principles and validation of core software tools

1. Design principles
2. State of the art technology
3. **Validation of core software tools**

Validation principles

Morrison, G.S. et al. (2021). ***Consensus on validation of forensic voice comparison.*** *Science & Justice 61:299–309.*

> “In the **context of a case**, given the results of an **empirical validation** of a forensic-voice-comparison system, how can one decide **whether the system is good enough** for its output to be used in court?”

Validation principles

(Specific question … two mutually-exclusive propositions)

Trained on

Calibrated using

Validated on

relevant data – population and conditions (of the **case**)

Validated using **new data**

recording pairs : same-speaker and different-speaker

case conditions : questioned- and known-speaker.

Validation principles

Make subjective decisions **early** to reduce cognitive bias:

e.g.

- identifying population and conditions
- choosing **sufficiently** representative data
- choosing **sufficient** speakers
- using **separate** data (speakers) for training, calibration and validation

Validation principles

Evaluate with **standard** metric and visualisations

Metric: Cost of Log Likelihood Ratio $\boldsymbol{C_{llr}}$

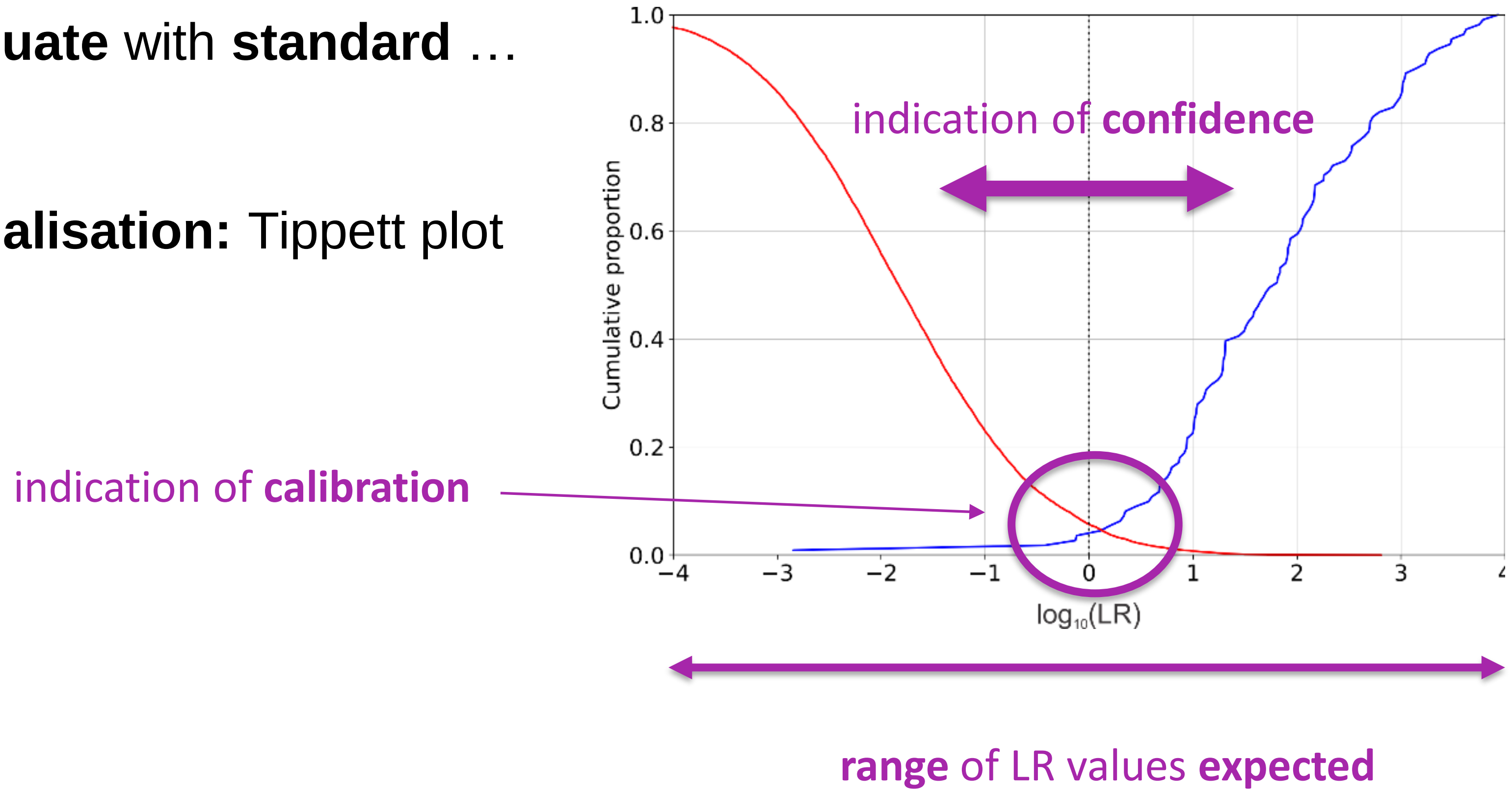
Average penalty for misleading or uninformative likelihood ratios.

$C_{llr} \rightarrow 0$: more informative system for this case
$C_{llr} \rightarrow 1$: less informative system for this case
$C_{llr} > 1$: miscalibration or error

Validation principles

Evaluate with **standard** …

Visualisation: Tippett plot



The E3 Forensic Speech Science System (E^{3}FS^3):

Design principles and validation of core software tools

1 → 2 → 3

1. Design principles
2. State of the art technology
3. **Validation of core software tools**

E^3FS^3: Three-fold Validation

E^3FS^3: Three-fold Validation

1. **Benchmark** on *forensic_eval_01*

2. Australian English **male**:
simulated case-specific conditions

3. Australian English **female**:
simulated case-specific conditions

case

1. Benchmark validation

1. **Benchmark** on *forensic_eval_01*

→ E^{3}FS^3 performs well on **known benchmark:**

2016–2019 virtual special issue of

Speech Communication (journal)

1. Benchmark validation

Data: Male speakers of Australian English.

Defined:

Questioned-speaker condition:
46s landline phone call: babble noise, compressed.

Known-speaker condition:
126s interview, reverberant room.

Training set: 423 recordings, 105 speakers.
Calibration / validation: 223 recordings, 61 speakers.

1. Benchmark validation

train **front-end:**

generic: 1000s of recordings & speakers

train **back-end:**

case-specific: *forensic_eval_01* training set

generic: **adapted** to case conditions

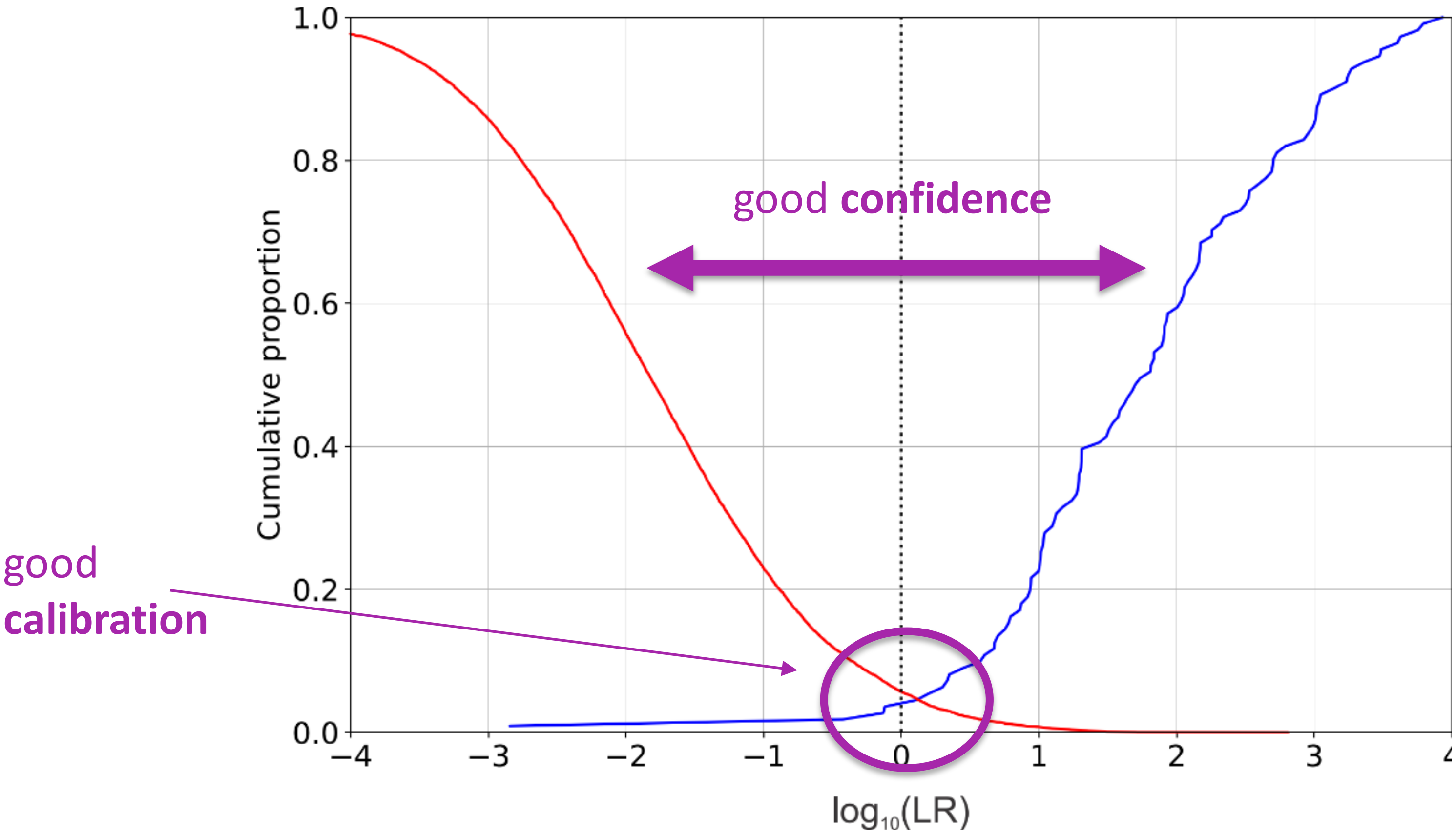
calibrate & validate:

case-specific: *forensic_eval_01* validation set

1. Benchmark validation

System Name	System Type	C_{llr}
	GMM-UBM	c 0.6
	i-vector	c 0.4
VOCALISE 2019A	x-vector	0.246
E^3FS^3 alpha	“	**0.208**
Phonexia BETA4	“	0.207

1. Benchmark validation



E^{3}FS^3: Three-fold Validation

1. **Benchmark** on *forensic_eval_01*

→ E^{3}FS^3 performs well on **known benchmark**

Equal to prior best-performing systems.

→ **Confidence to proceed** with the case-specific validation

2. & 3. Case-specific validation

The case:

Female speakers of Australian English.

Multiple known-speaker recordings

→ we used 120s sections.

Multiple questioned-speaker recordings

in different durations.

→ we ran **multiple validations**.

…

2. & 3. Case-specific validation

The case:

Telephone calls from mobile to call centre.

Manually diarized (careful process).

Three questioned-speaker conditions.

One same as the known-speaker condition.

No apparent background **noise**.

2. Male: case-specific conditions

Close to the case.

Final chance to **change the system**.

2. Australian English **male**:
   simulated case-specific conditions

2. Male: case-specific conditions

train **front-end:**

generic: 1000s of recordings & speakers

train **back-end:**

case-specific: **simulated from clean data**

generic: **adapted** to case conditions

calibrate & validate:

case-specific: **simulated from clean data**

Base data

male & female speakers (Australian English).

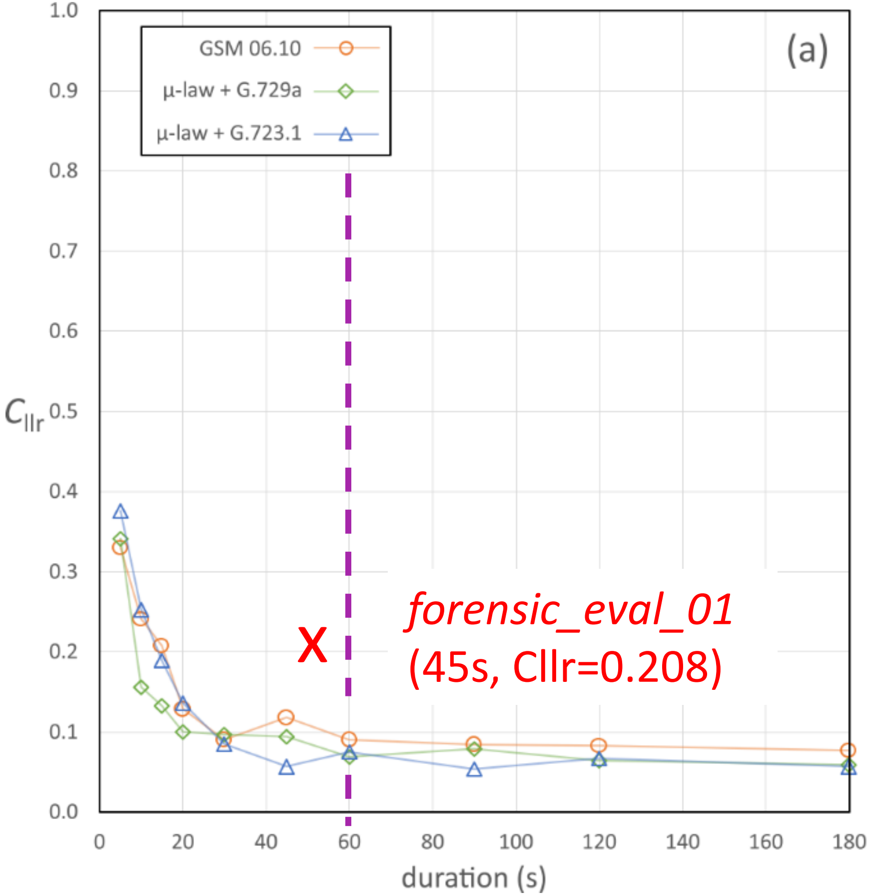
Subjective choices from:

- Multiple speaking tasks → case conditions
- Multiple recording sessions → **k** & **q** recordings
- Train / calibrate / validate **split**

Simulate from high-quality audio:

- → Apply codecs, compression, etc. **as per case**
- → Extract contiguous **duration** chunks

2. Male: results



Overall C_{llr} << *forensic_eval_01*.

No clear difference between conditions.

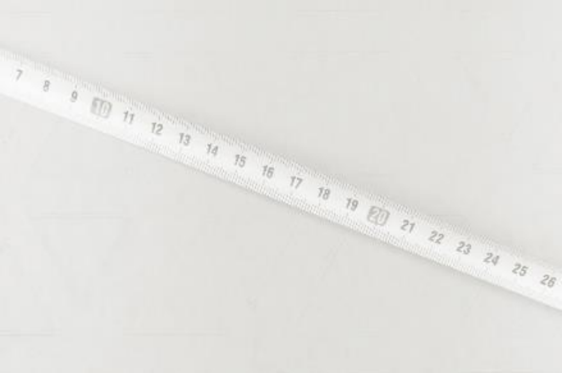
No improvement > 60 seconds.

E^{3}FS^3: Three-fold Validation

2. Australian English **male**:
   simulated case-specific conditions

   → E^{3}FS^3 performs well on a dataset **close to the case**

   → **Confidence to proceed** with the case-specific validation



3. Female: case-specific conditions

3. Australian English **female**:
   simulated case-specific conditions

3. Female: case-specific conditions

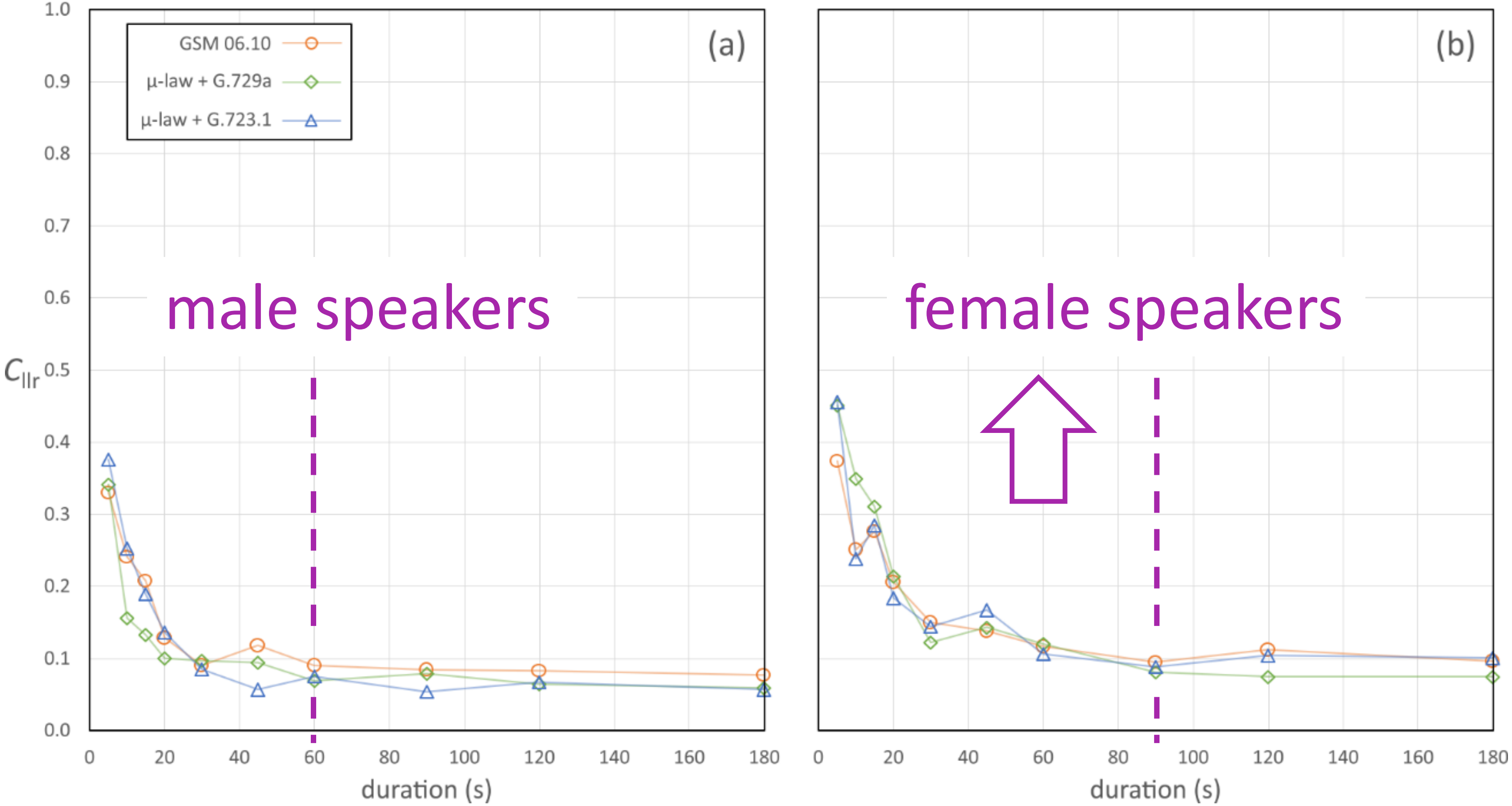
Data (AusEng 500+ database)**:**

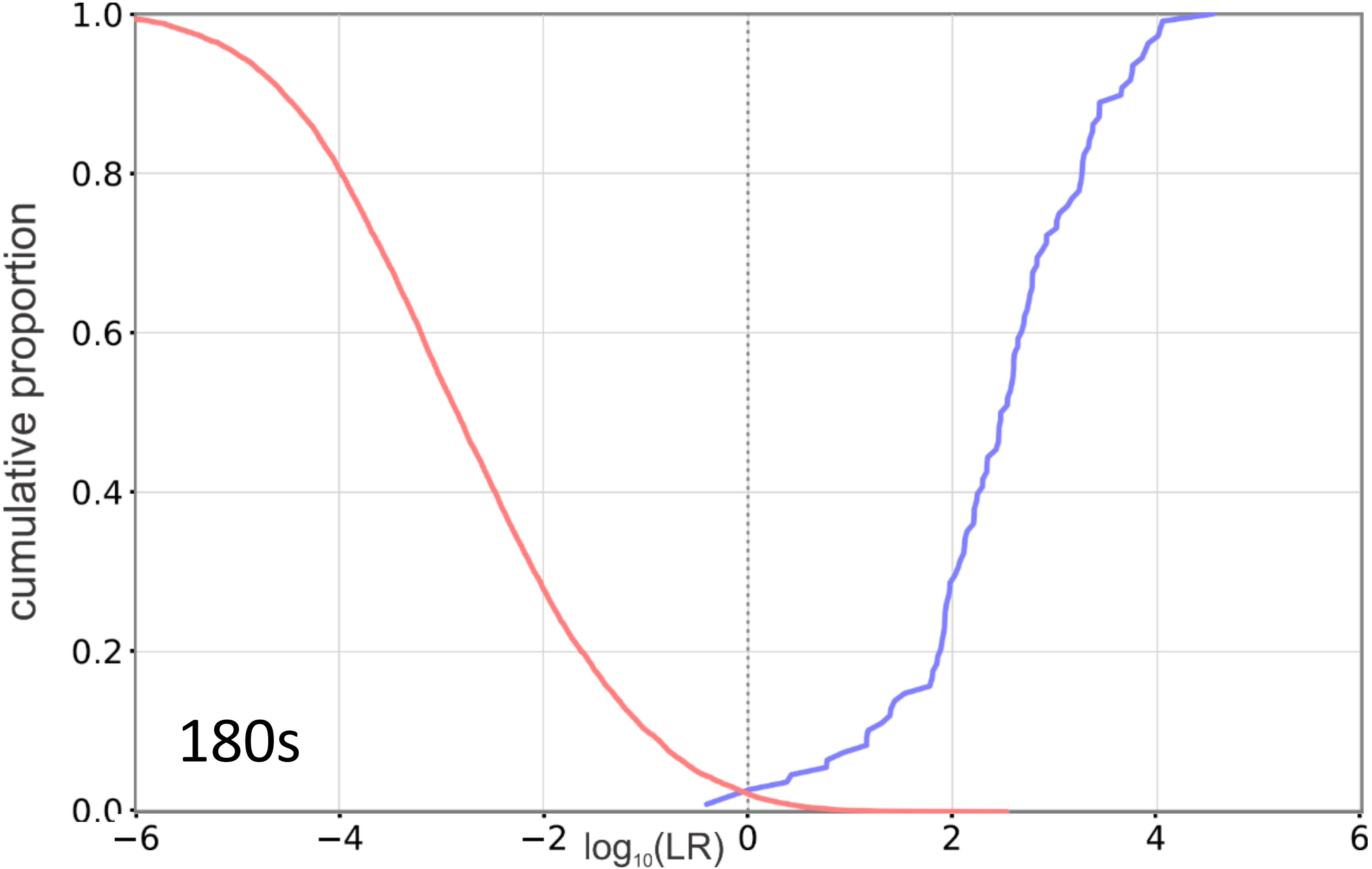
We split the dataset:

125 speakers for case-specific **training data**

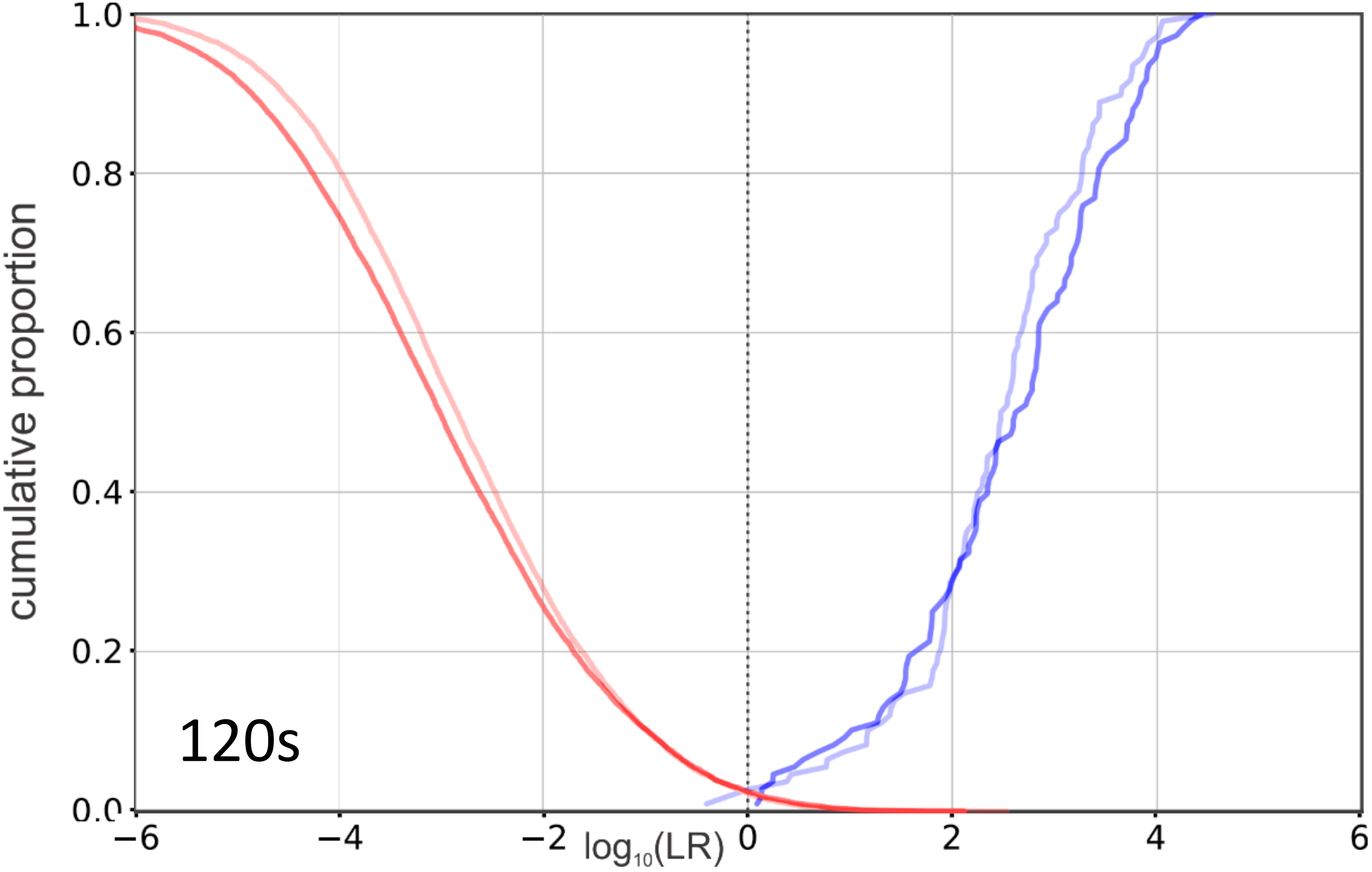
108 speakers for **calibration** and **validation**

3. Female: case-specific conditions

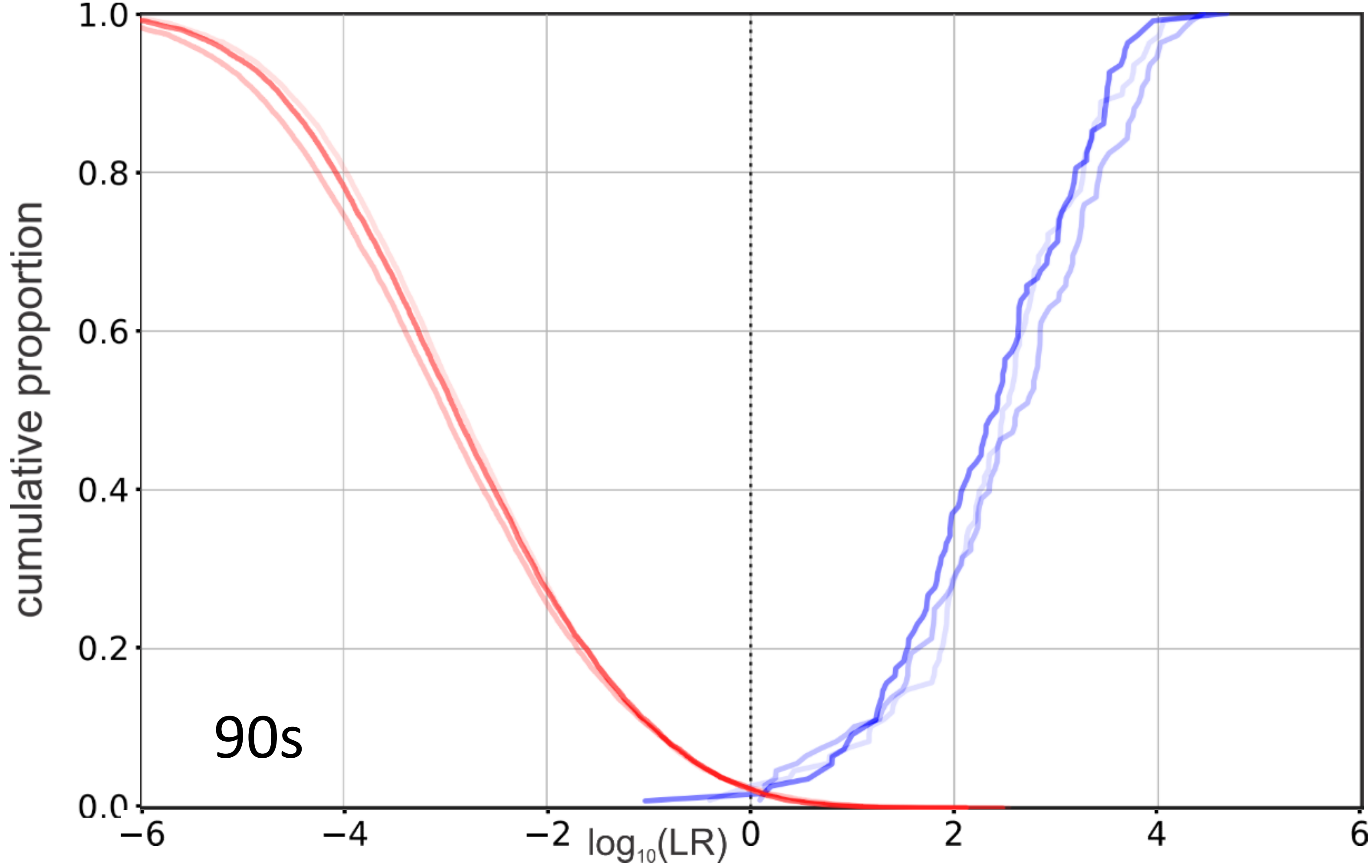




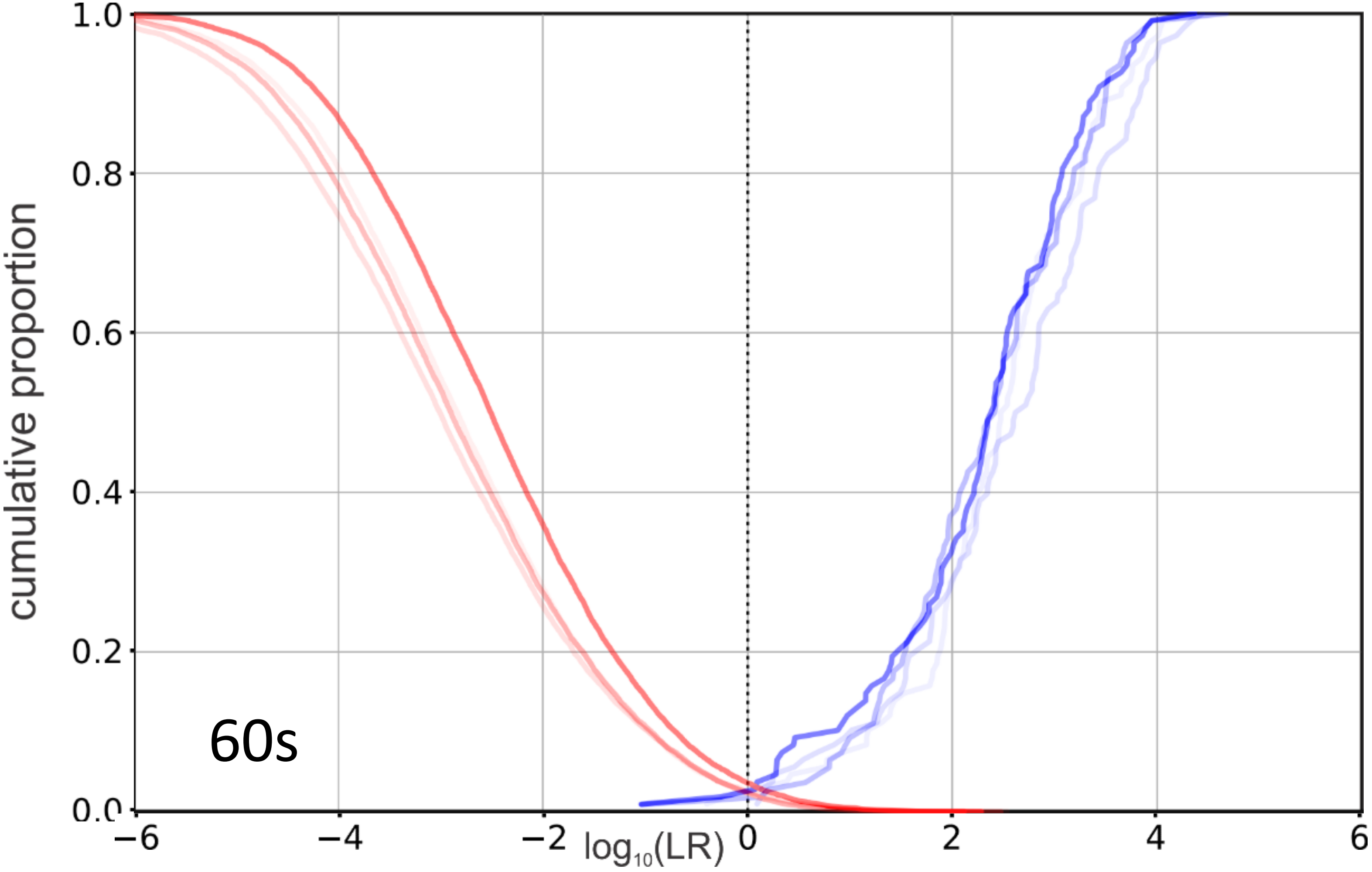
cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
180s
−6
−4
−2
log10(LR)
0
2
4
6



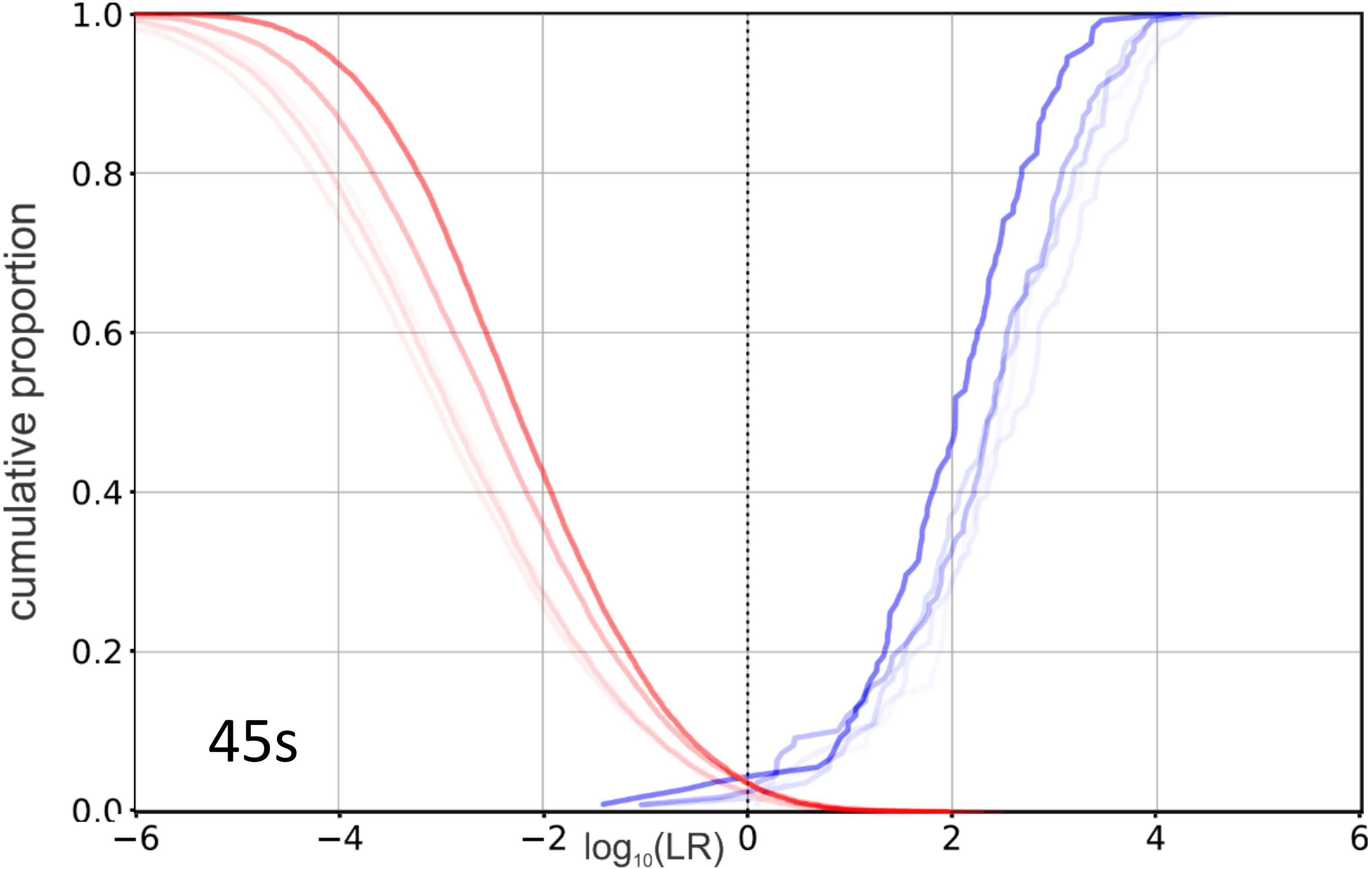
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
log$_{10}$(LR)
cumulative proportion
120s



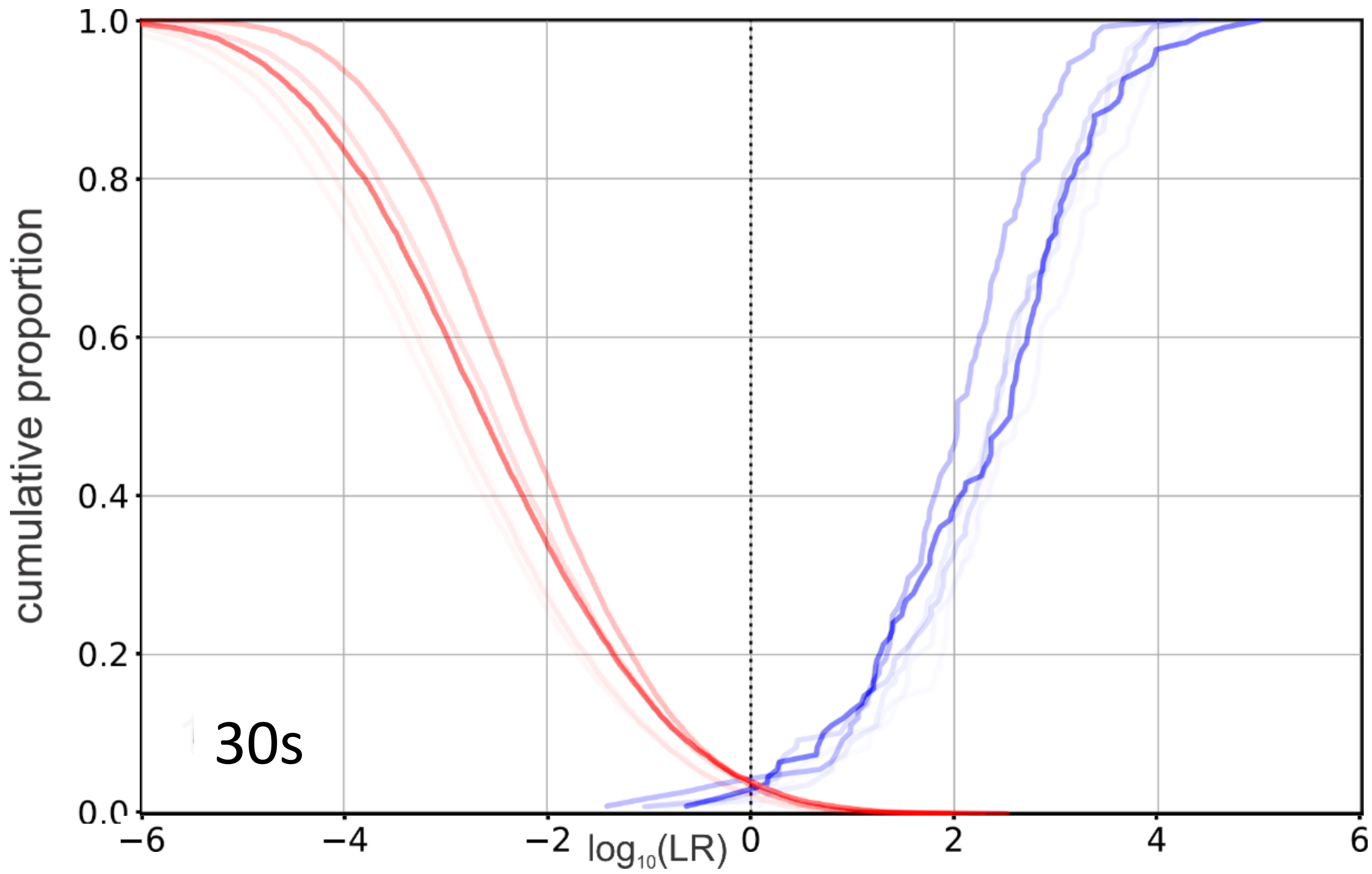
cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
log10(LR)
90s



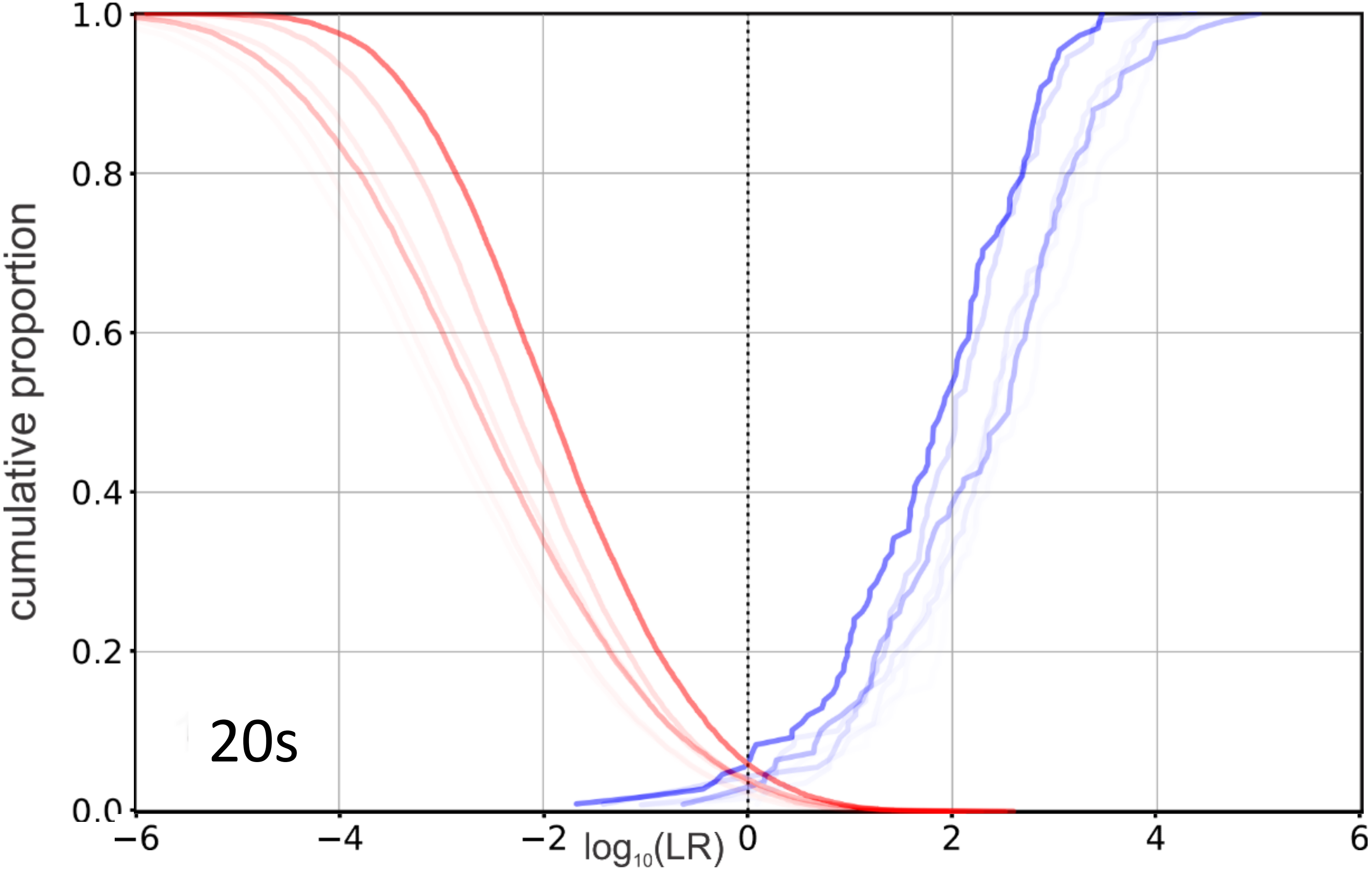
cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
log10(LR)
60s



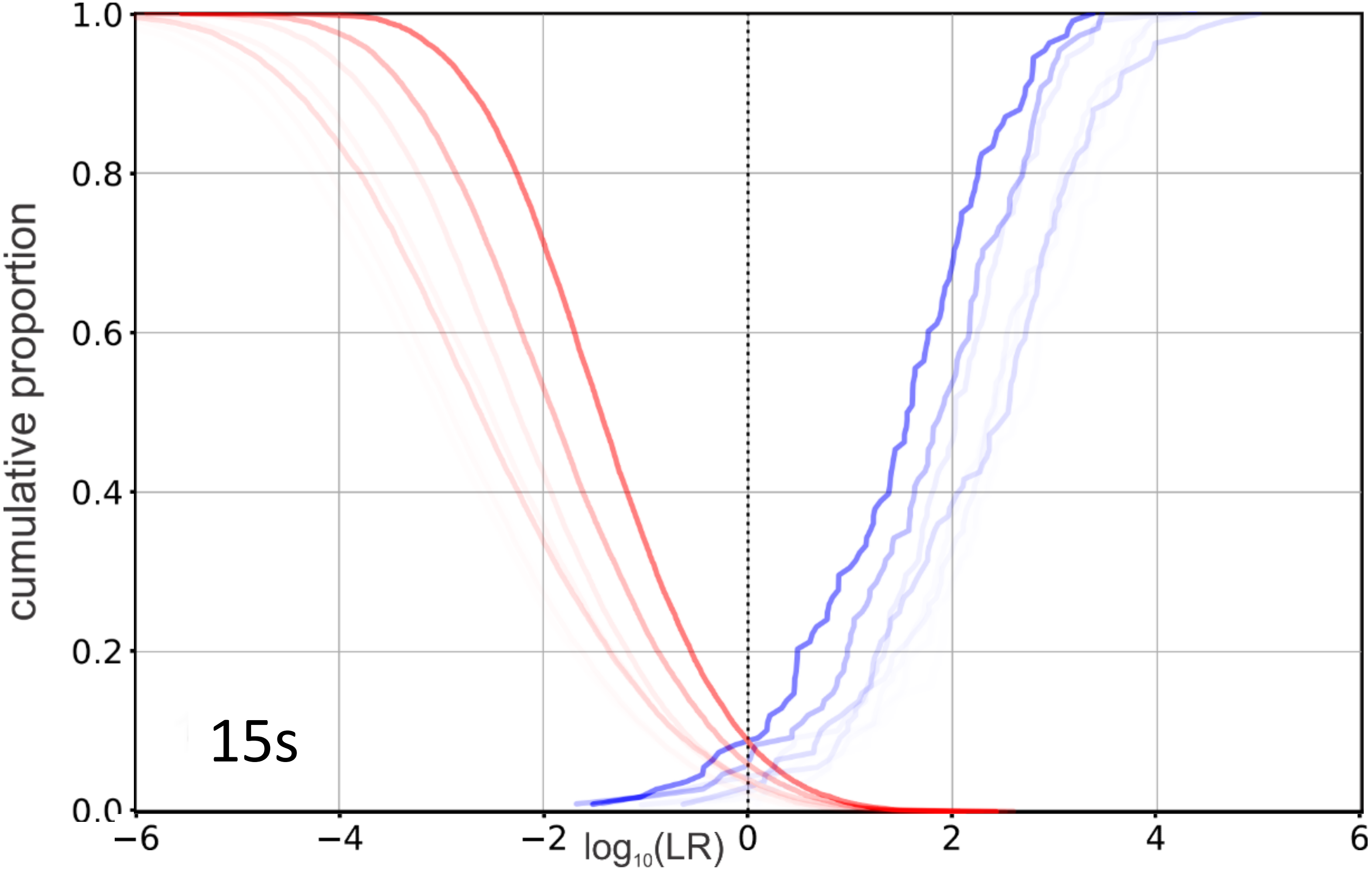
cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
log10(LR)
45s

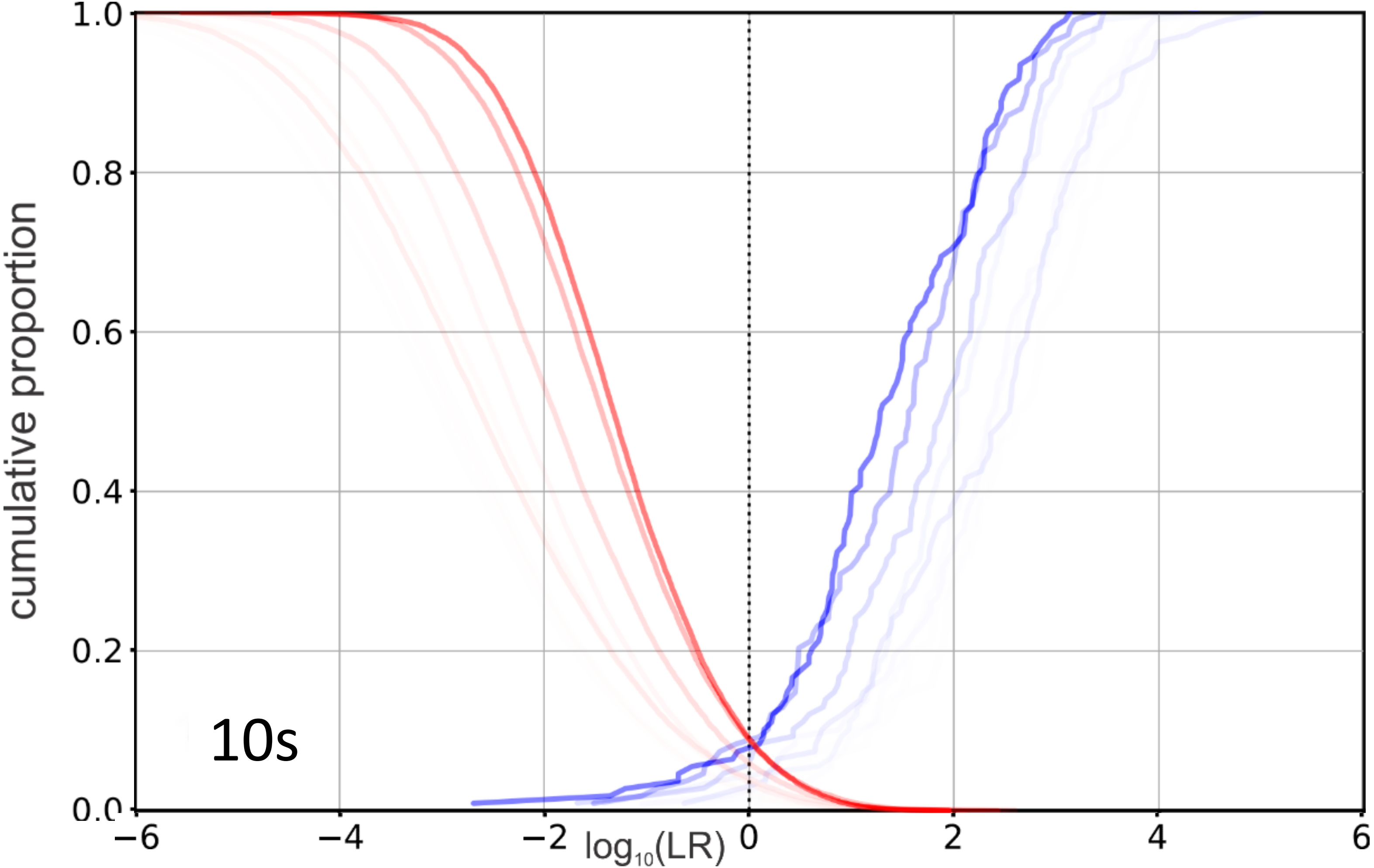


1.0
0.8
0.6
0.4
0.2
0.0
cumulative proportion
30s
−6
−4
−2
log10(LR)
0
2
4
6

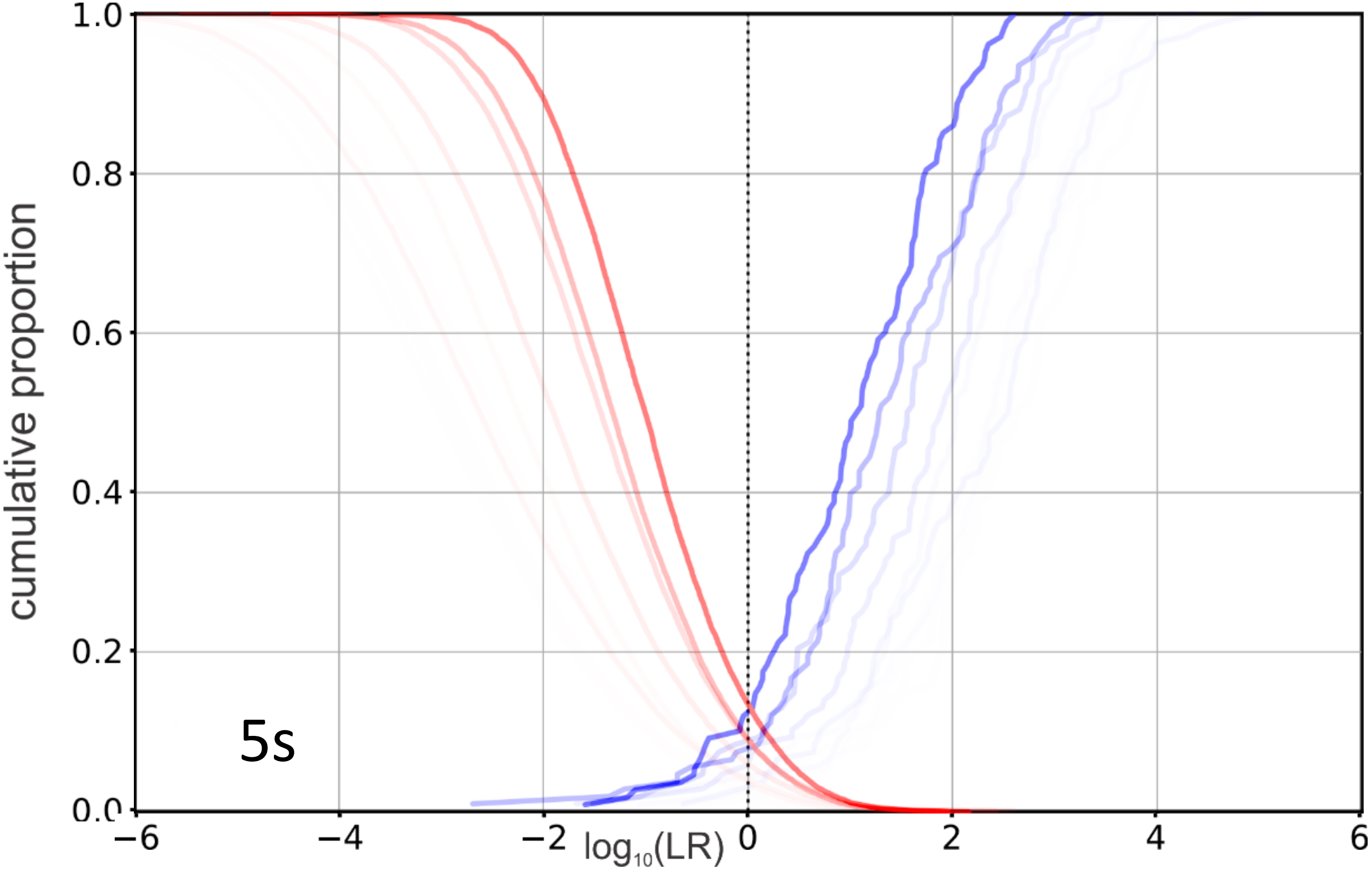


1.0
0.8
0.6
0.4
0.2
0.0
cumulative proportion
−6
−4
−2
0
2
4
6
$\log_{10}$(LR)
20s


1.0
0.8
0.6
0.4
0.2
0.0
cumulative proportion
−6
−4
−2
0
2
4
6
$\log_{10}$(LR)
15s



1.0
0.8
0.6
0.4
0.2
0.0
cumulative proportion
−6
−4
−2
0
2
4
6
log10(LR)
10s



cumulative proportion
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
log10(LR)
5s

E^3FS^3: Three-fold Validation

Well-calibrated

Support LRs into the 100s **for this case**

(for the shortest durations)

Performance (slightly) worse than for males

3. Australian English **female**:
   simulated case-specific conditions

→ **Confidence to proceed** with the case

→ case

So what?

E^3FS^3 : a forensic speech science system.

Based on **state-of-the-art** automatic-speaker-recognition algorithms.

Designed for research and casework.

Designed to be **open and transparent**.

Supported by relevant data, procedures and training.

Validated:

Leading results on the established benchmark.

"Satisfying results" on a recent case.

Dr. Phil Weber

Contact us:

p.weber1@aston.ac.uk | http://forensic-data-science.net/

Read the full paper:

Weber P., Enzinger E., Labrador B., Lozano-Díez A., Ramos D., González-Rodríguez J., Morrison G.S. (2022). **Validation of the alpha version of the E3 Forensic Speech Science System (E3FS3) core software tools.** Forensic Science International: Synergy, 4:100223.