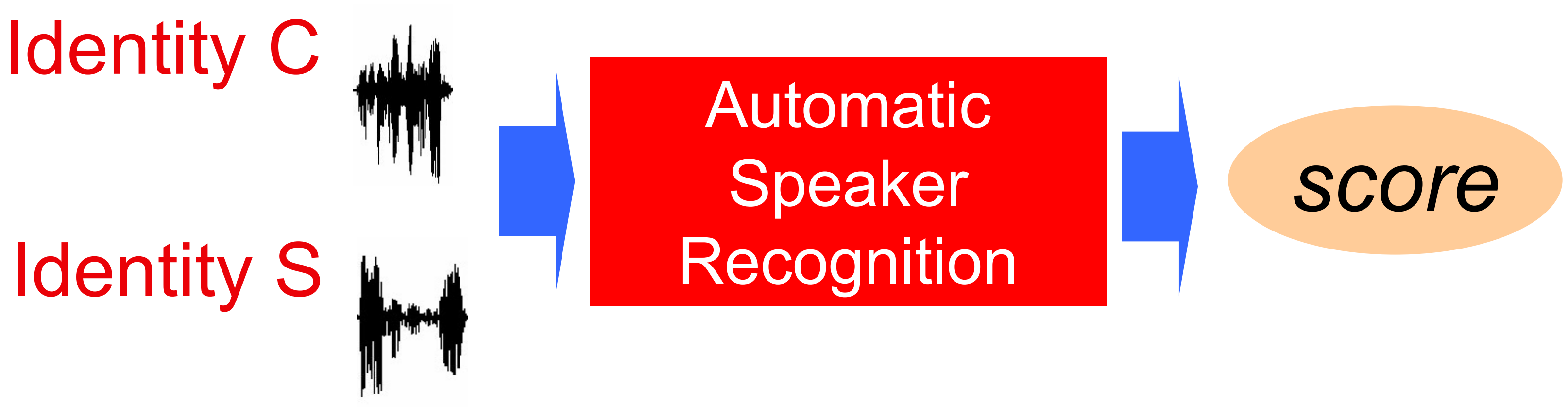
Symposium on calibration in forensic science
Aston Institute for Forensic Linguistics

Calibration in Forensic Voice Comparison

Daniel Ramos
daniel.ramos@uam.es
Audias – Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid
http://audias.ii.uam.es

Automatic Speaker Recognition

- Standard architechture: *scores*



Identity C

Identity S

Automatic Speaker Recognition

score

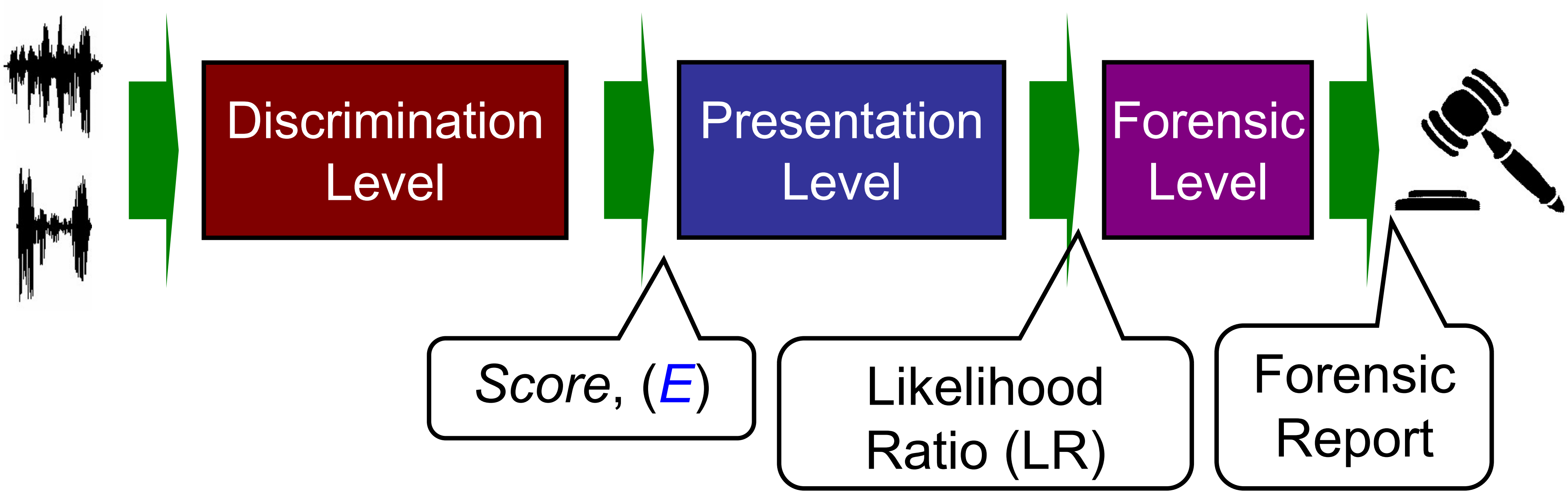
- Ideally:
  - If C y S are same identity (same-source), high***er*** score
  - If C y S are different identities (different-source), low***er*** score
- Thus, a score allows discrimination
- Not enough in forensics: a likelihood ratio (LR) is needed

ENFSI EUROPEAN NETWORK OF FORENSIC SCIENCE INSTITUTES

Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition

ENFSI GUIDELINE FOR EVALUATIVE REPORTING IN FORENSIC SCIENCE

LR with Automatic Speaker Recognition



Discrimination Level

Presentation Level

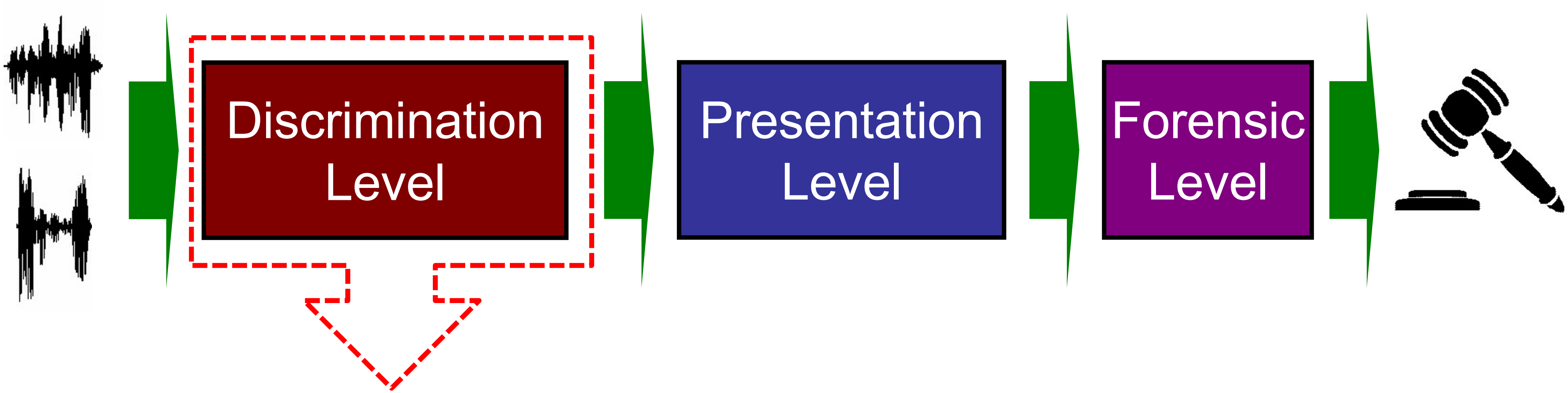
Forensic Level

Score, (*E*)

Likelihood Ratio (LR)

Forensic Report

LR with Automatic Speaker Recognition



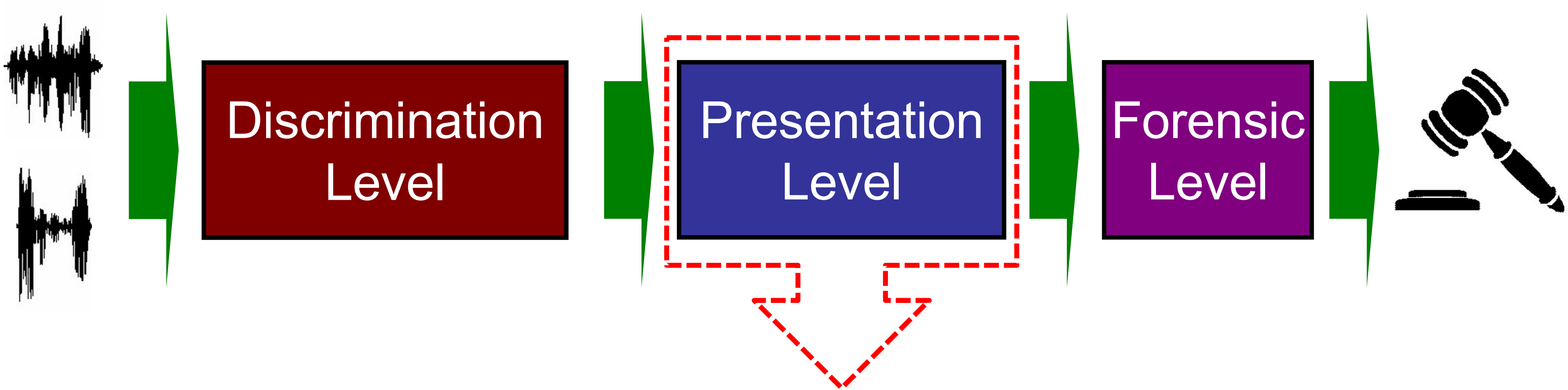
Discrimination Level

Presentation Level

Forensic Level

- Objective: discriminating scores
  - ❑ Score-based architecture
  - ❑ Improve discrimination
  - ❑ So-called automatic speaker recognition

LR with Automatic Speaker Recognition



Discrimination Level

Presentation Level

Forensic Level

- Objective: transform scores into likelihood ratios
  - Warning: LRs *must be well calibrated*

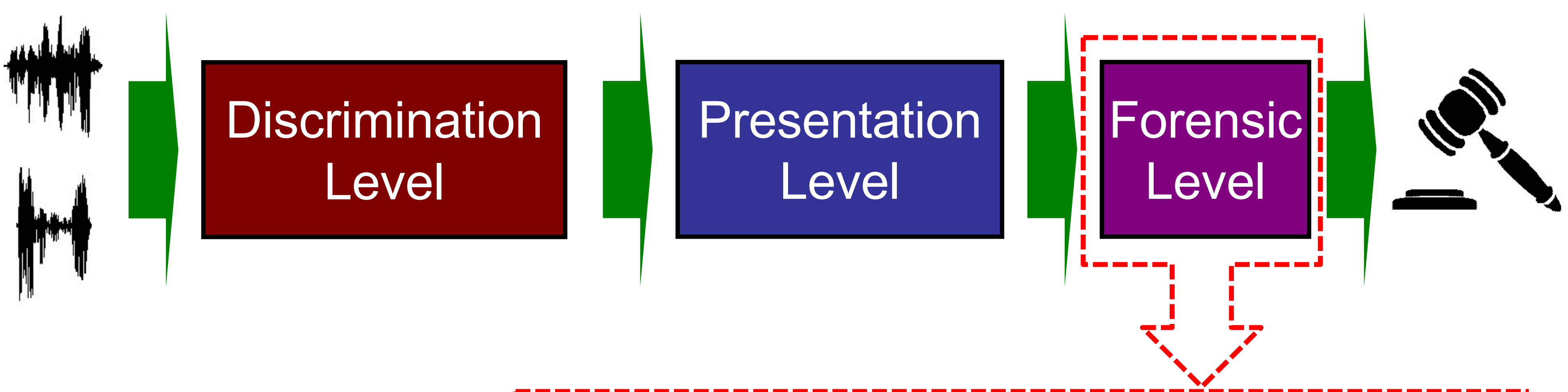
Forensic Science International

Reliable support: Measuring calibration of likelihood ratios☆

Daniel Ramos *, Joaquin Gonzalez-Rodriguez

  - C_{llr}: a popular measure of performance (the lower the better)

LR with Automatic Speaker Recognition



- Objective: adequate forensic reports
  - □ Probabilistic weight of the evidence
  - □ Following recommendations (ENFSI)
  - □ Validation
  - □ Accreditation

EUROPEAN NETWORK OF FORENSIC SCIENCE INSTITUTES

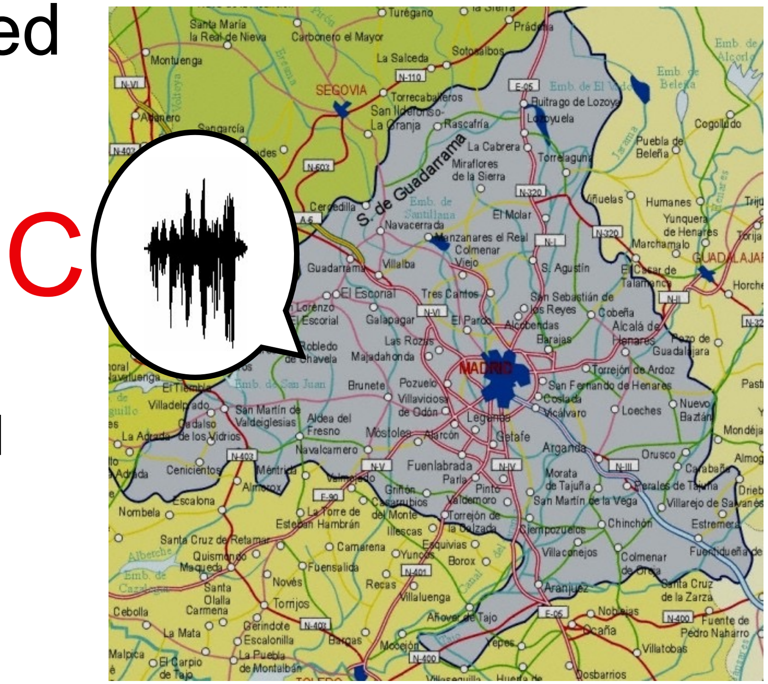
Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition

ENFSI GUIDELINE FOR EVALUATIVE REPORTING IN FORENSIC SCIENCE

A Very Simplified (Yet Illustrative) Example

Simulated Case

- Incriminating recordings wire-tapped in the Madrid region (trace)
  - ❑ Population: potential sources of the speech
    - Speakers from Madrid region, with similar characteristics with questioned speech
      - ❑ Language
      - ❑ Accent
      - ❑ ...
  - ❑ Digital Wire-Tapping (SITEL, Spanish nationwide system)

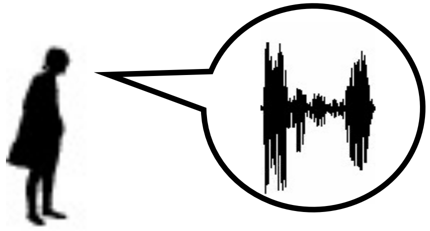


- Police investigations lead to a suspect

Simulated Case

- Recordings are taken from the suspect (reference speech)

S



- Typically, controlled recordings
  - But very different conditions as for the questioned speech
- Could be previous wire-tappings where authorship is accepted
  - Similar conditions as questioned speech

Example: Ahumada III Database (Real Cases)

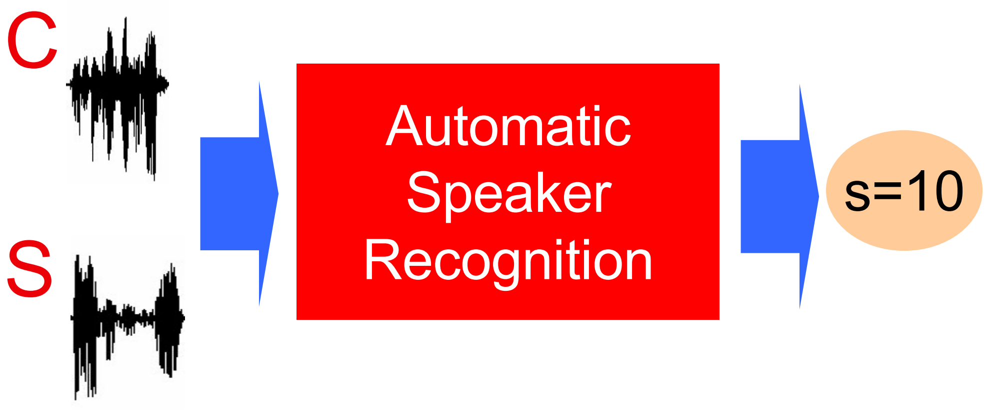
S (reference)

C (trace)

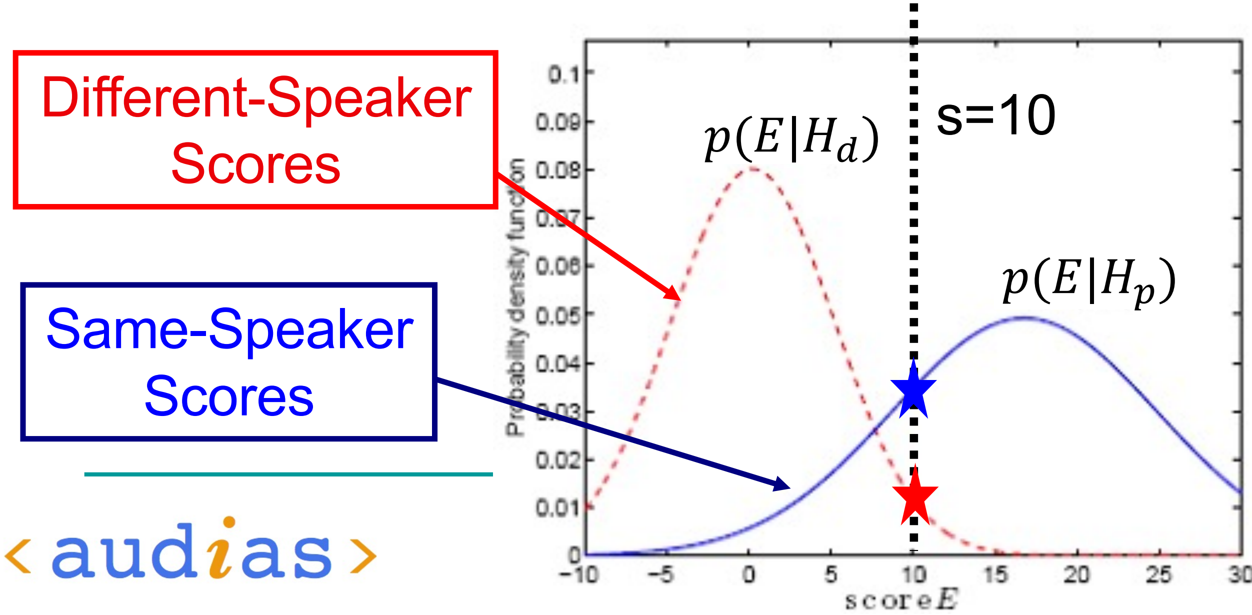
D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-case database in Spanish", in Proceedings of Interspeech 2008, pp. 1493-1496, September 2008.

LR Computation

- Step 1: the automatic system computes a score
  - No meaning on itself
    - 10 with respect to what?
  - In general, non-interpretable
    - Its range of variation is not known *a priori*



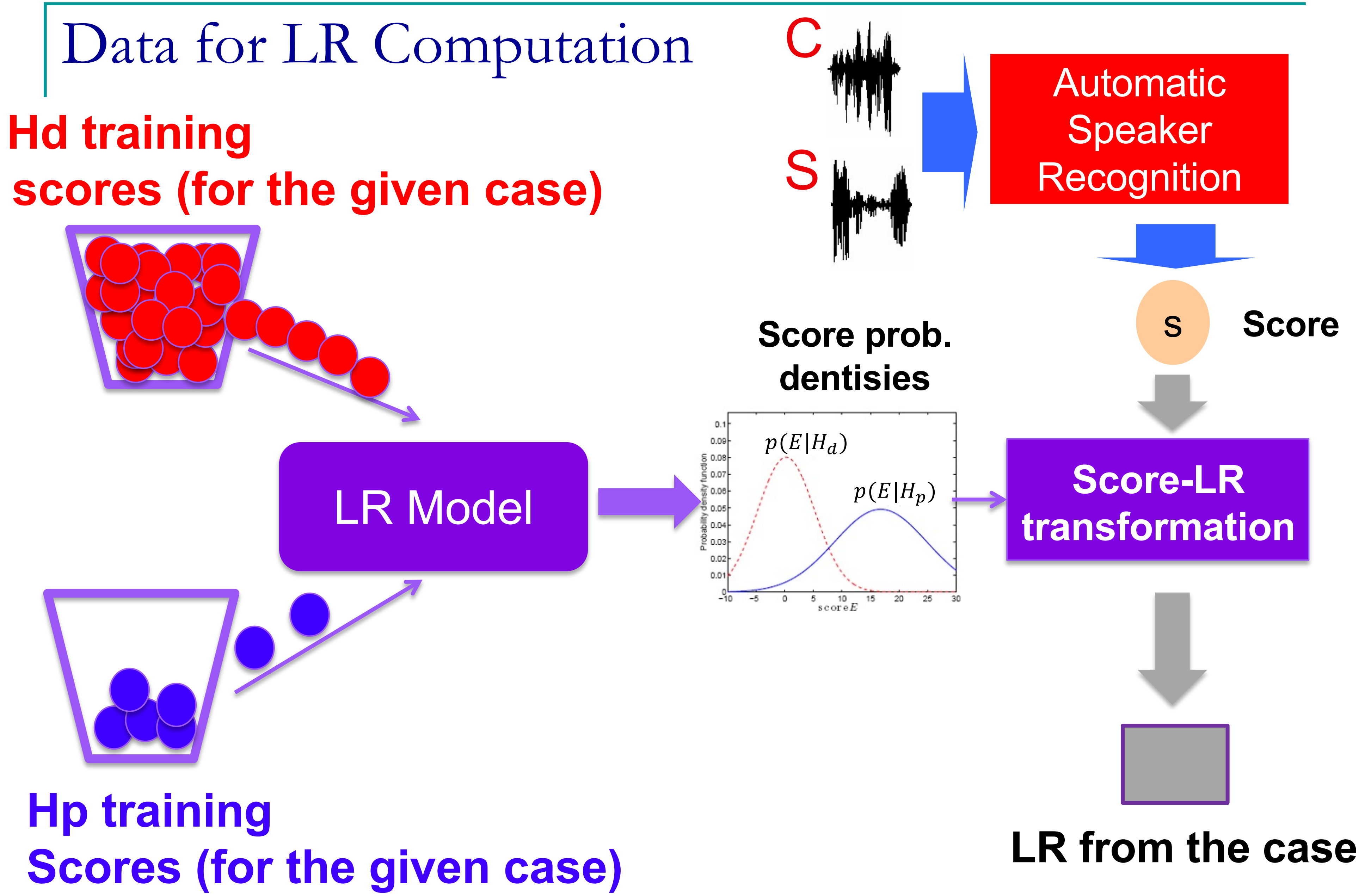
- Step 2: compute the LR
  - In this example, we use a Gaussian model



$$LR = \frac{0.35}{0.15} = 2.33$$

Weak support to prosecutor proposition ("same-source")

Data for LR Computation



C

S

Automatic Speaker Recognition

s

Score

Hd training scores (for the given case)

Hp training Scores (for the given case)

LR Model

Score prob. dentisies

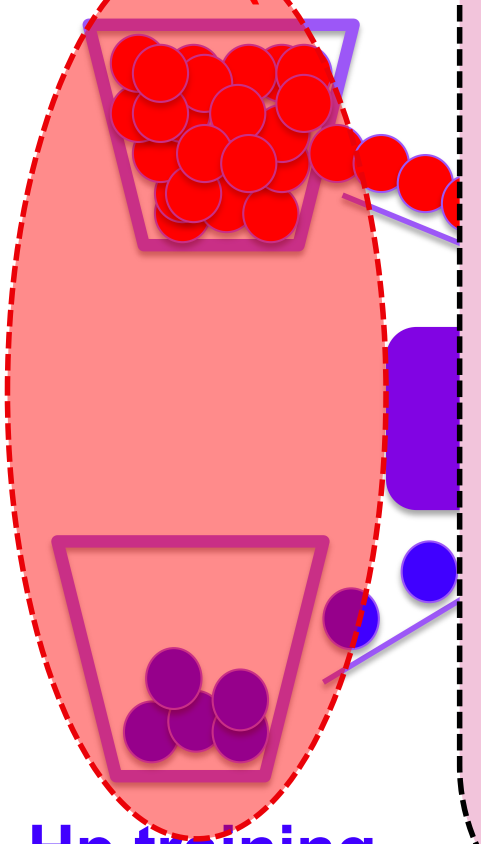
Score-LR transformation

LR from the case

Data for LR Computation

C

Hd training scores (for the

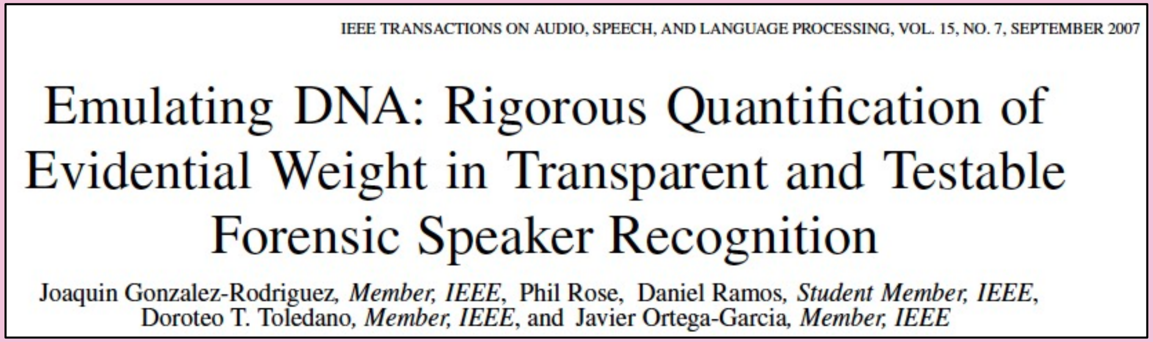


omatic ker tion

Score

-LR nation

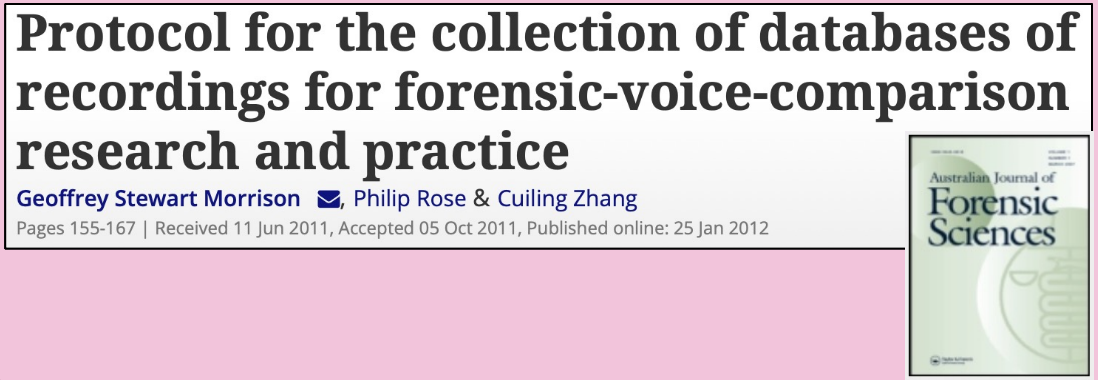
Selection of traning data is fundamental for LR computation!



IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007

Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition

Joaquin Gonzalez-Rodriguez, *Member, IEEE*, Phil Rose, Daniel Ramos, *Student Member, IEEE*, Doroteo T. Toledano, *Member, IEEE*, and Javier Ortega-Garcia, *Member, IEEE*



Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice

Geoffrey Stewart Morrison, Philip Rose & Cuiling Zhang

Pages 155-167 | Received 11 Jun 2011, Accepted 05 Oct 2011, Published online: 25 Jan 2012

Australian Journal of Forensic Sciences

Hp training Scores (for the g

n the case

A Guideline for the Validation of LR Methods (With Emphasis in Automatic Methods)

Guideline: Validation of Forensic LR Methods

Forensic Science International 276 (2017) 142–153

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation

Didier Meuwly [a,b,*], Daniel Ramos [c], Rudolf Haraksim [d]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB The Hague, The Netherlands
[b] University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands
[c] ATVS – Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, C/Francisco Tomas y Valiente 11, 28049 Madrid, Spain
[d] LTS5 – Signal Processing Laboratory, École Polytechnique Fédérale de Lausanne, Faculty of Electrical Engineering, Station 11, CH-1015 Lausanne, Switzerland

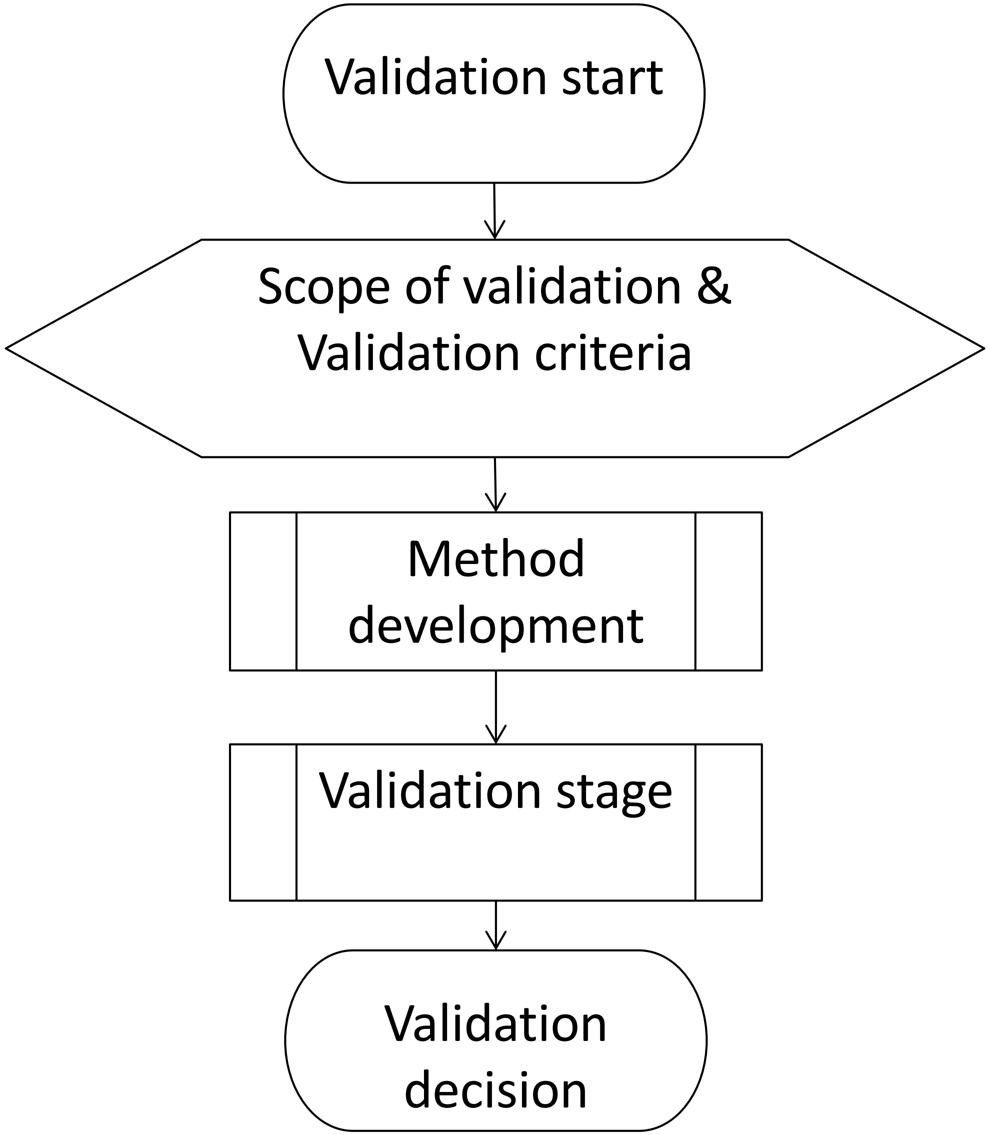
Guideline: Validation of Forensic LR Methods

- Objective
  - Determine if a LR method is valid to be used in casework
  - All the validation process should be documented for transparency
  - Towards standardization of procedures (for biometrics)
- Validation process
  - Based on Empirical Testing
    - Data: still an issue
  - Performance assessment
    - Performance *characteristics*
      - *What aspect of performance should be measured?*
    - Performance *metrics*
      - *How to measure a characteristic?*
    - Performance graphical representations
      - *Ok, show me an illustrating plot*

Emphasis on Forensic Data
* Lab (development) performance
* Followed by forensic performance

ISO
Terminology

Validation – a process



Validation start

Scope of validation & Validation criteria

Method development

Validation stage

Validation decision

Performance Characteristics and Measures

Performance Characteristic	Performance Metric	Graphical Representation
Accuracy	Cllr, EER	ECE plot DET plot
Discriminating power	Cllr^{min}	ECE^{min} plot
Calibration	Cllr^{cal}	Tippett plot
Robustness	Cllr, EER LR range	ECE plot DET plot Tippett plot
Coherence	Cllr, EER	ECE plot DET plot Tippett plot
Generalization	Cllr, EER	ECE plot DET plot

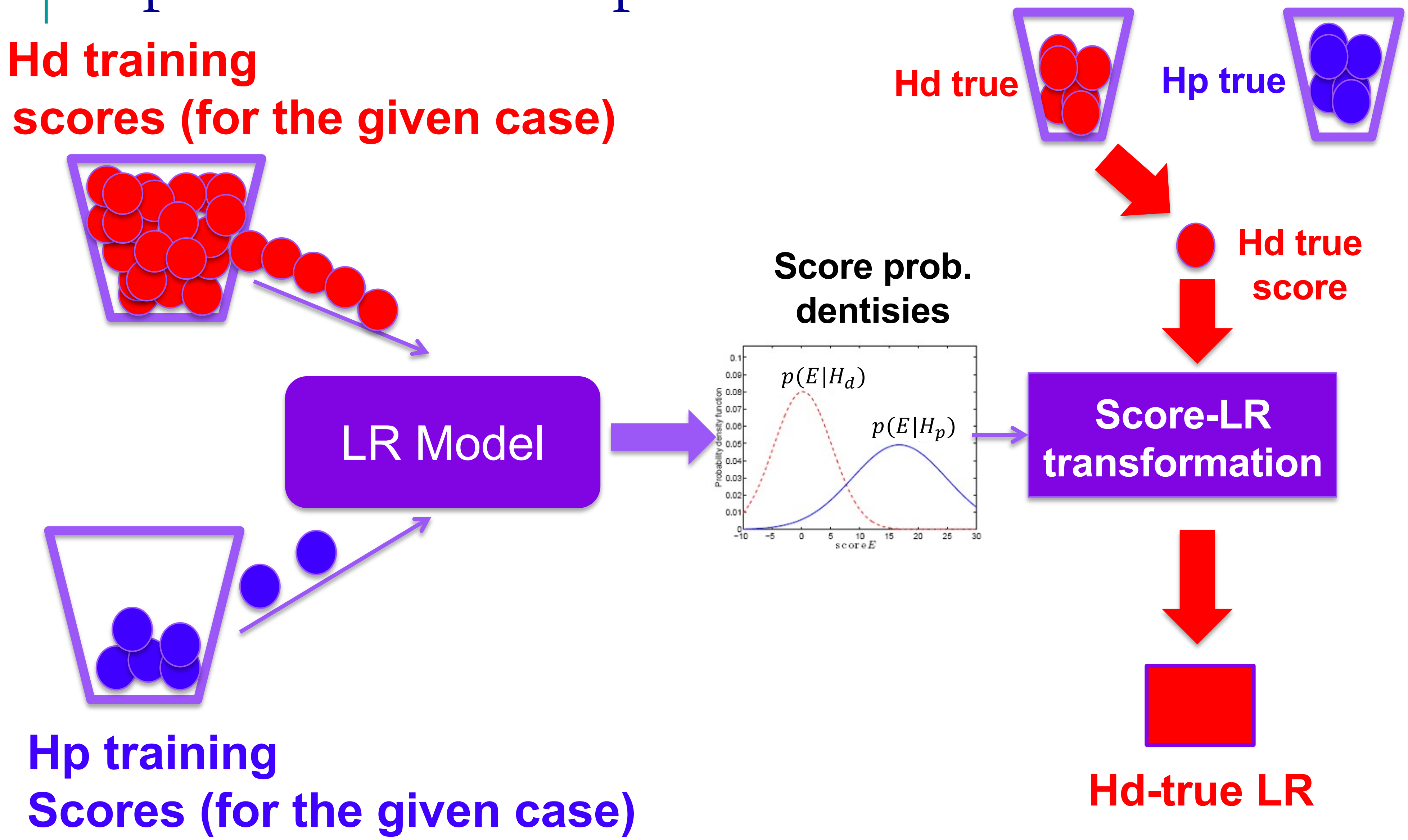
Performance Characteristics and Measures

Performance Characteristic	Performance Metric	Graphical Representation
Accura		ECE plot t
Discrim power		plot
Calibrat		plot
Robustnes	Cllr, EER	ECE plot DET plot ett plot
Coherence		plot plot ett plot
Generalization	Cllr, EER	plot DET plot

All measures require a validation set of LR values (computed automatically from a validation database)

Not restricted to these! The Guideline is thought to be open to modifications

Experimental Set-Up



Simulated Case Scores

Hd true

Hp true

Hd training scores (for the given case)

Hd true score

Score prob. dentisies

LR Model

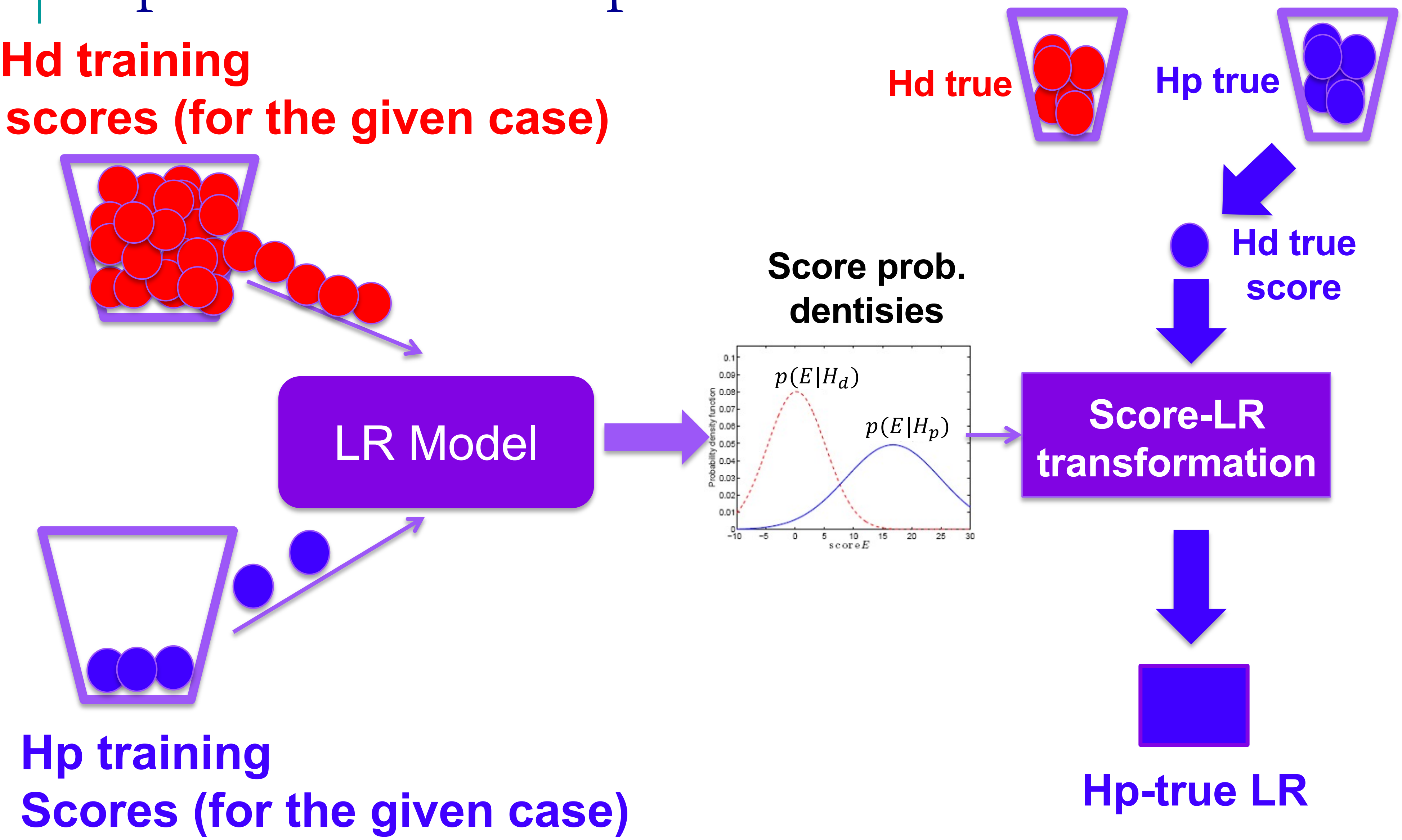
Score-LR transformation

Hp training Scores (for the given case)

Hd-true LR

Experimental Set-Up

Simulated Case Scores



Hd training scores (for the given case)

Hd true

Hp true

Hd true score

Score prob. dentisies

LR Model

Score-LR transformation

Hp training Scores (for the given case)

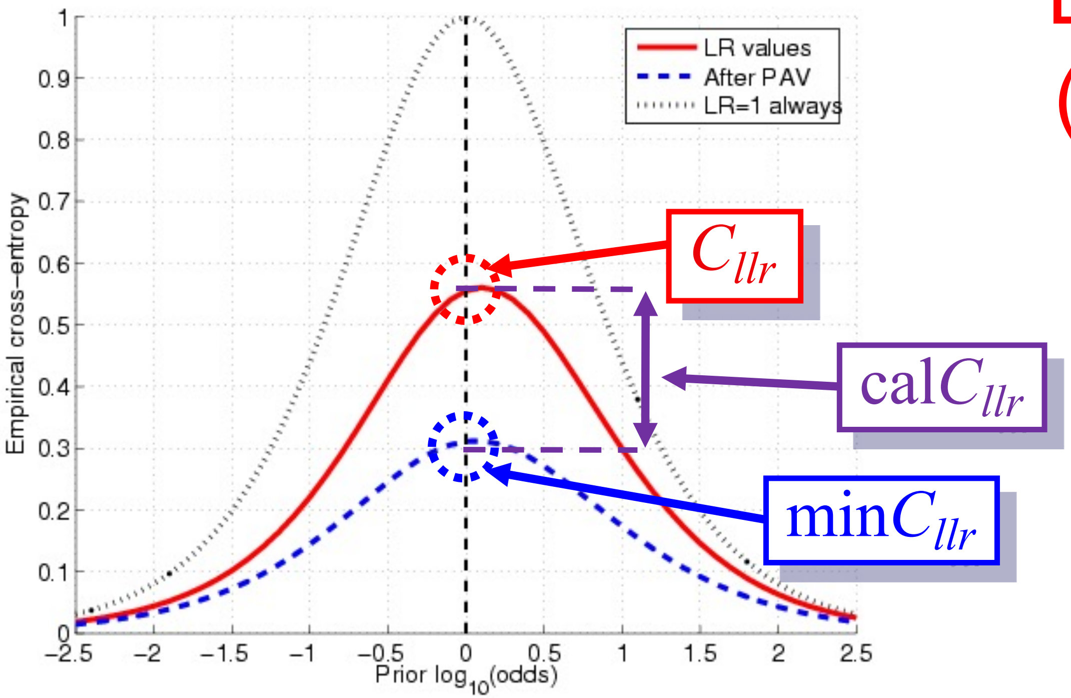
Hp-true LR

Some Performance Metrics and Representations (Included in the Guideline)

Guideline for Validation: Performance

	Performance Characteristic	Performance Metric	Graphical Representation
Primary	**Accuracy**	Cllr	ECE plot
	Discriminating power	Cllr^{min} EER	ECE^{min} plot DET plot
	Calibration	Cllr^{cal}	Tippett plot
Secondary (measure behavior of the primary)	**Robustness**	Cllr EER LR range	ECE plot DET plot Tippett plot
	Coherence	Cllr EER	ECE plot DET plot Tippett plot
	Generalization	Cllr EER	ECE plot DET plot

Empirical Cross-Entropy Plots and C_{llr}



ECE curve: Accuracy (the lower the better)

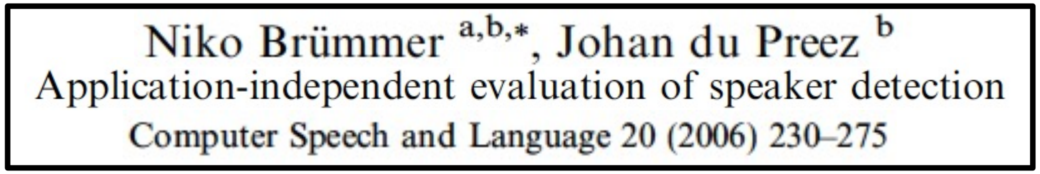
Calibration (red – blue)

Discrimination (blue curve)

D. Ramos, J. Gonzalez-Rodriguez, G. Zadora and C. Aitken. “Information-theoretical Assessment of the Performance of Likelihood Ratios”. Journal of Forensic Sciences (under minor revision)

- http://arantxa.ii.uam.es/~dramos/software.html

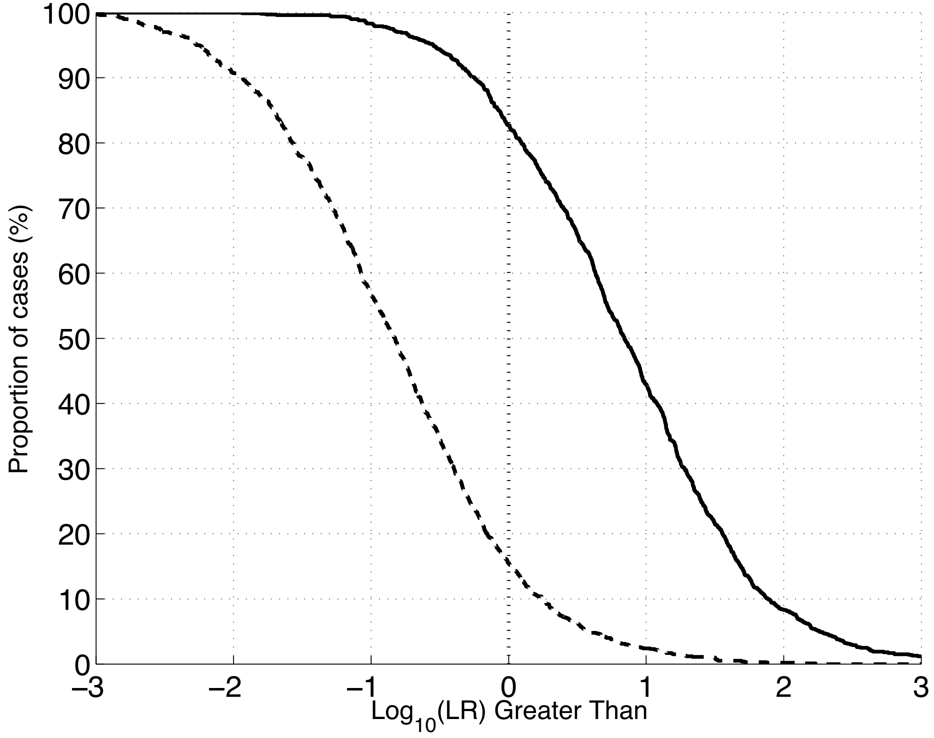
- Summarizing metric: C_{llr}

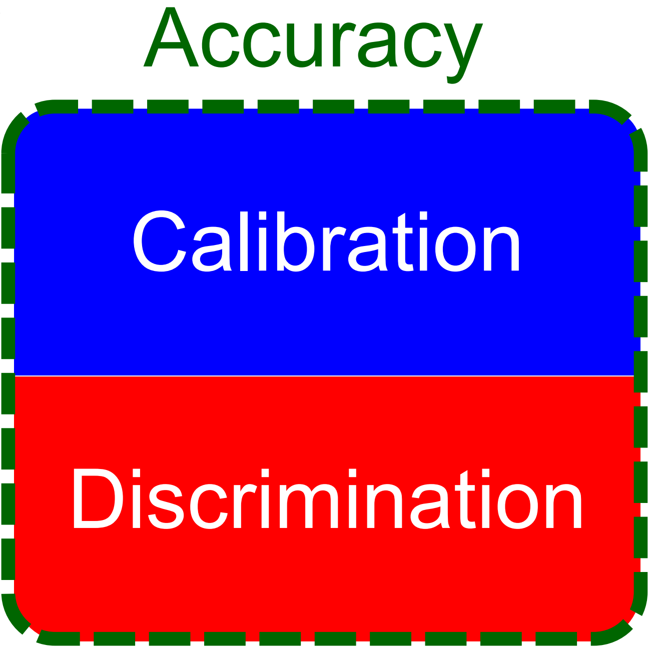


Niko Brümmer [a,b,*], Johan du Preez [b]
Application-independent evaluation of speaker detection
Computer Speech and Language 20 (2006) 230–275

Tippett Plots

- Cumulative distribution of LR values



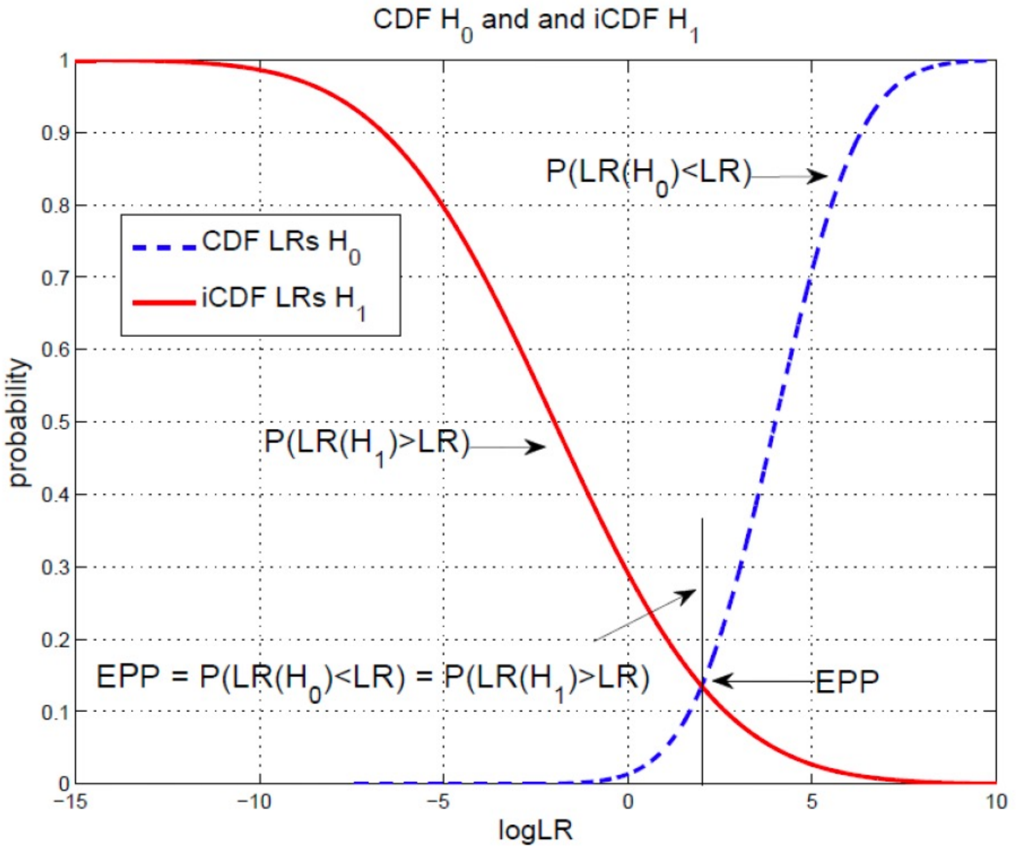


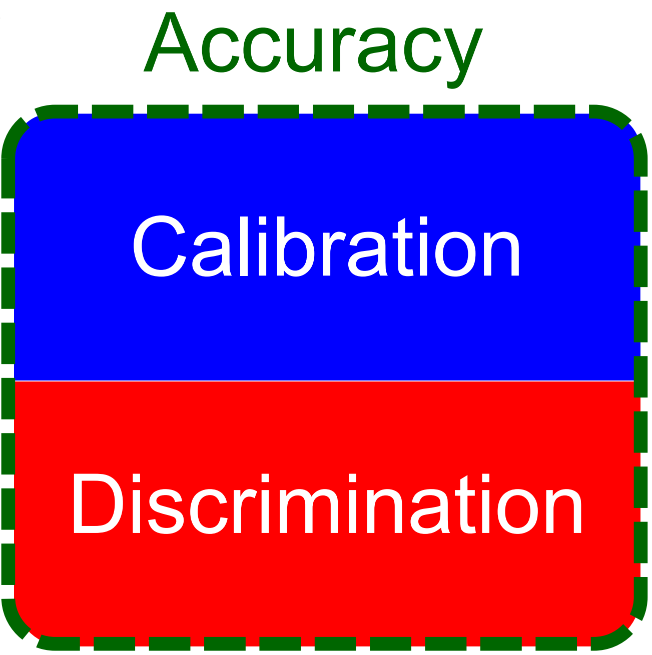
WARNING: Tippett plots do not measure them explicitly!

- Summarizing metric:
  - Rates of misleading evidence

Tippett Plots (Type II representation)

- Cumulative distribution of LR values





WARNING: Tippett plots do not measure them explicitly!

EUROPEAN NETWORK OF FORENSIC SCIENCE INSTITUTES

Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition

Acknowledgements

Acknowledgements

- People who contributed in one way or another to this presentation:
  - Didier Meuwly (NFI)
  - Rudolf Haraksim (European Comission Joint Research Center)
  - Charles Berger (NFI)