Calibration in Speaker Verification

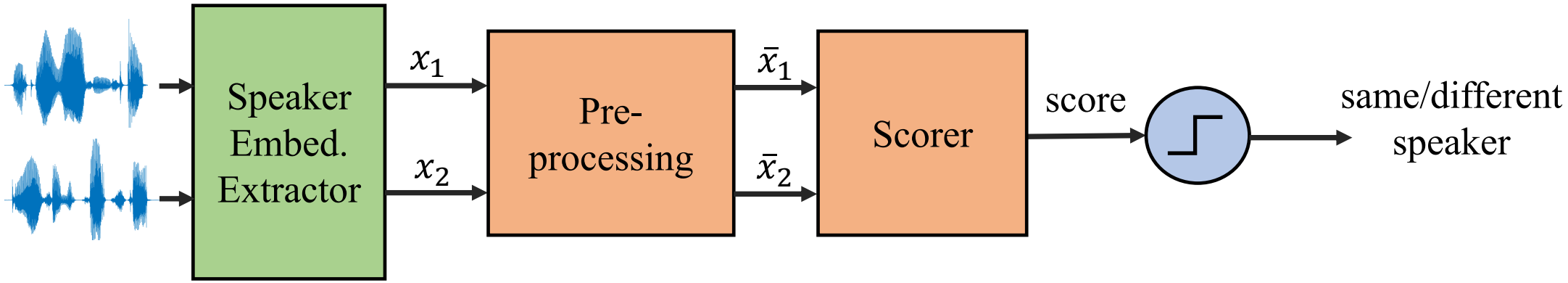
Luciana Ferrer

Symposium on Calibration, June 2021

Introduction

- Automatic speaker verification (SV) systems are often used for forensic voice comparison
- Standard SV systems are very fragile to changes in conditions
- In this talk I will describe
  - The current standard SV pipeline
  - Optimal Bayes decision theory
  - Some metrics to measure calibration
  - How we currently deal with miscalibration
  - Can we do better?

A Standard Speaker Verification Pipeline



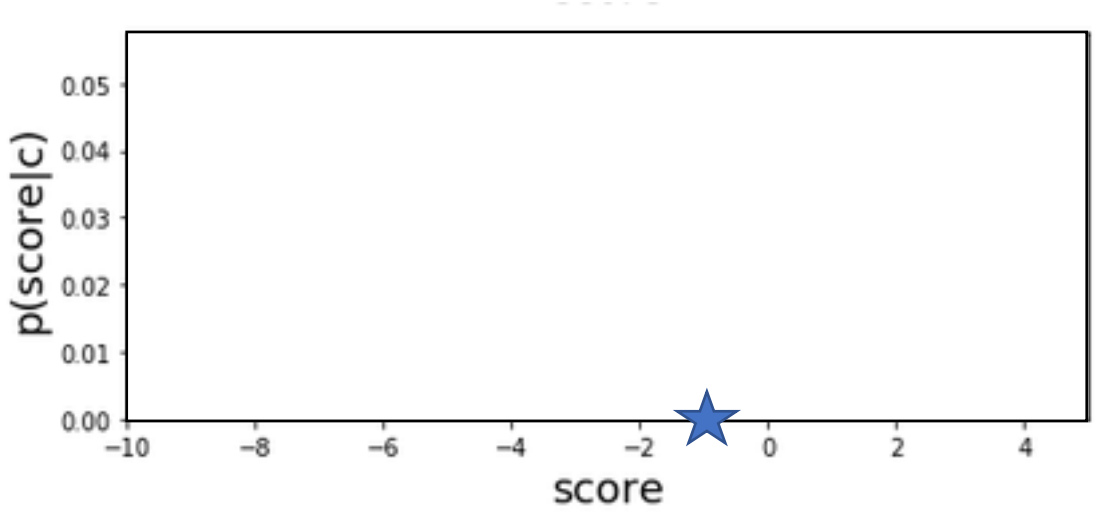
Most common scorer: Probabilistic Linear Discriminant Analysis (PLDA)

PLDA's scores are computed as log-likelihood ratios under a set of Gaussian assumptions

$$\text{score} = \log \frac{p(\bar{x}_1, \bar{x}_2 \mid \text{same speaker})}{p(\bar{x}_1, \bar{x}_2 \mid \text{diff speaker})}$$

The Calibration Problem

- In most cases, scores that come out of PLDA are misscalibrated
  - They are *not* LLRs, even though they are computed to be so
  - The cause is a mismatch between assumptions and reality
- Misscalibrated scores have no probabilistic interpretation
  - cannot be interpreted in absolute terms, only relative to each other
  - can only be thresholded optimally if we have access to their distribution



The Calibration Problem

- If scores are calibrated their value has meaning
- We say a score is calibrated if
  - For posteriors $p = P(c = 1 \mid p)$
  - For log-likelihood ratios $s = \log \frac{P(s \mid c = 1)}{P(s \mid c = 0)}$

If forecaster says 40% chance of rain, 40% of those times it rains

- Calibrated scores can be optimally thresholded using Bayes decision theory

Bayes Decision Theory

In general, we want to minimize this cost

Detected class — True class

$$\text{Cost} = C_{10}\ P(c=1)\ P(\hat{c}=0|c=1) + C_{01}\ P(c=0)\ P(\hat{c}=1|c=0)$$

Costs for deciding 0 when true class was 1

Expected prior for class 1 on **test** data

Prob. of error for class 1 on **test** data

This is minimized when $\hat{c}(d) = \begin{cases} 1 & \text{if } C_{10}\,P(c=1)\,p(d|c=1) > C_{01}\,P(c=0)\,p(d|c=0) \\ 0 & \text{otherwise} \end{cases}$

Trial's data

Section 1.5.2 from Bishop, "Pattern Recognition and Machine Learning"

Optimal Decisions

$$\hat{c}(d) = \begin{cases} 1 & \text{if } C_{10}\,P(c=1)\,p(d|c=1) > C_{01}\,P(c=0)\,p(d|c=0) \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{c}(x) = \begin{cases} 1 & \text{if } \text{LLR} > \theta \\ 0 & \text{otherwise} \end{cases}$$

- If we have LLRs we can trivially make optimal decisions for any cost function
- These decisions are good only if the system outputs are well-calibrated

Log Likelihood Ratio

$$\text{LLR} = \log \frac{p(d|c=1)}{p(d|c=0)}$$

Threshold

$$\theta = \log \frac{C_{01}\,P(c=0)}{C_{10}\,P(c=1)}$$

How do we measure calibration?

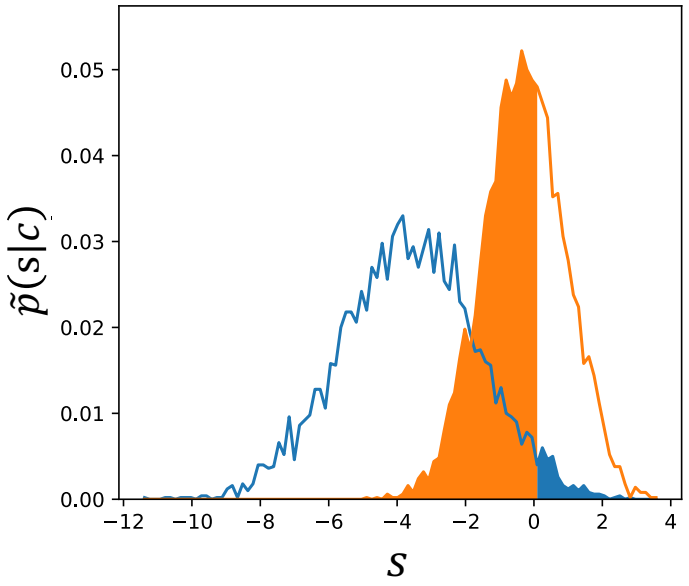
- How good will our system be at making Bayes decisions?
- Could decisions be improved by calibrating the scores?

Cost Decomposition

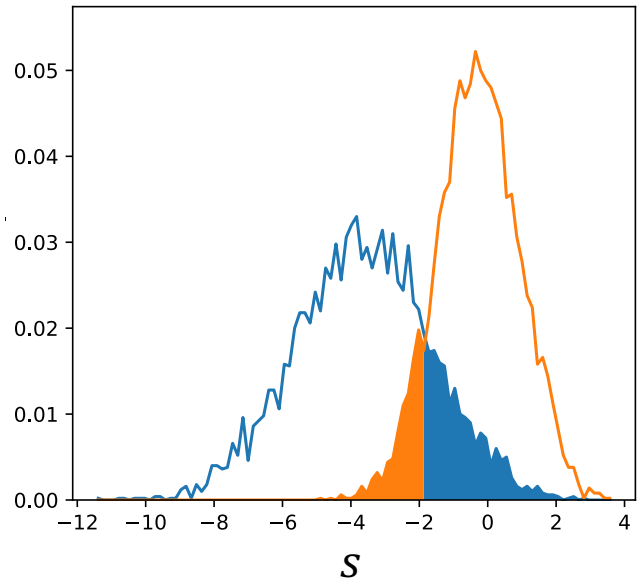
- Two sources of error compounded in cost: discrimination and calibration
  - C actual cost obtained with the theoretically optimal threshold
  - $C_{\min}$ obtained with the thr that minimizes it
  - $\Delta C = C - C_{\min}$ is a good measure of miscalibration

Taking same priors and costs
$\theta = 0$

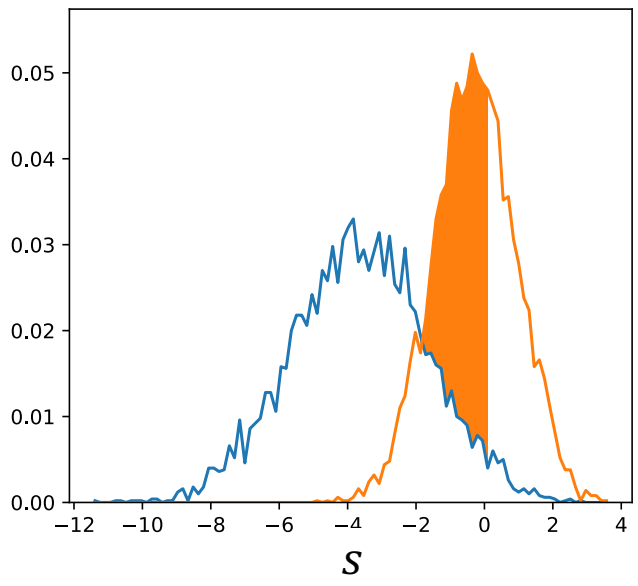
C: discrimination+calibration



$C_{\min}$: discrimination only



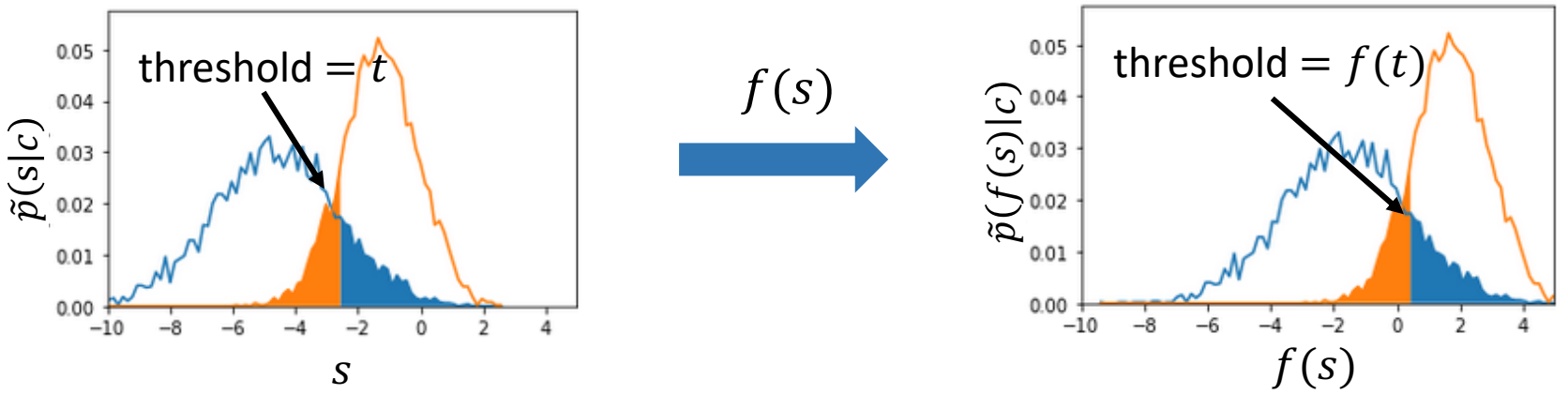
ΔC: miscalibration



Calibration vs Discrimination

- Discrimination: how well the scores separate the classes
- Calibration: whether those scores can be interpreted probabilistically

Discrimination is not changed if we transform the scores with an invertible transformation



Phil Dawid, “The well-calibrated Bayesian”, 1982
Niko Brummer, “Measuring, refining and calibrating speaker and language information extracted from speech”, 2010.

Cross-entropy as Evaluation Metric

- The cost measures performance at a single operating point
  - It evaluates the quality of hard decisions
- A more comprehensive measure is the cross-entropy $\mathrm{ECE} = -\frac{1}{N}\sum_{k} \log P(c = c_k|d_k)$
- Or its prior-weighted version:

Posteriors computed from LLR and priors using Bayes rule

$$C = \mathrm{ECE}_W = -\frac{P(c=0)}{N_0}\sum_{k|c_k=0} \log P(c=0|d_k) - \frac{P(c=1)}{N_1}\sum_{k|c_k=1} \log P(c=1|d_k)$$

Brummer and Preez, "Application-Independent Evaluation of Speaker Detection", 2006
Van Leuwen and Brummer, "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems", 2007
Ramos et al, "Deconstructing Cross-Entropy for Probabilistic Binary Classifiers", 2018

Cross-entropy as Evaluation Metric

- This cost can also be decomposed in discrimination and calibration terms

$$\Delta C = C - C_{\min}$$

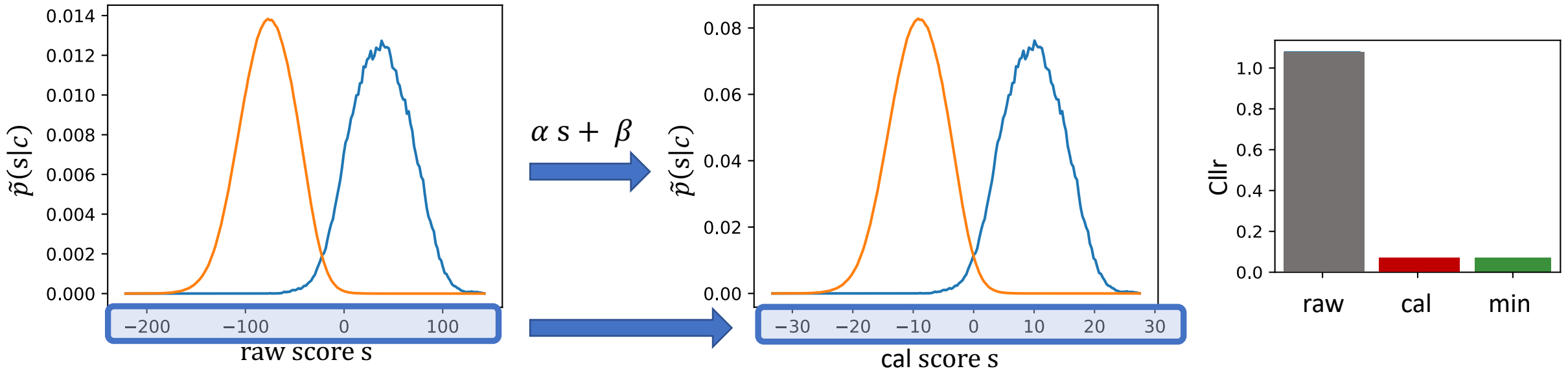
- The min is obtained by transforming the scores with the best monotonic transformation
  - Can use the pool-adjacent violators algorithm (PAV)

- The Cllr is defined as $\frac{C}{\log(2)}$ with $P(c = 0) = 0.5$
  - Property: $\text{Cllr}_{\min} < 1.0$

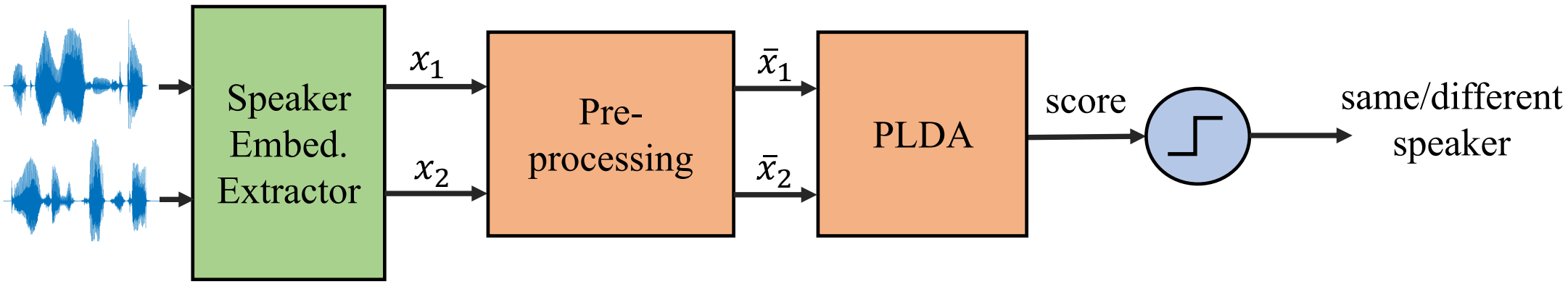
Brummer, “The PAV algorithm optimizes binary proper scoring rules”, 2013

How to Fix Bad Calibration

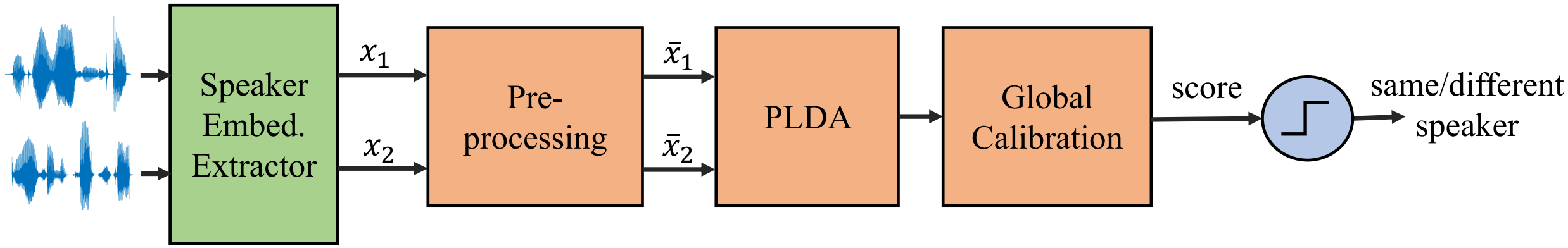
- Common approach: linear logistic regression
  - Assumes that $\mathrm{LLR} = \alpha\, \mathrm{s} + \beta$
  - Uses ECE_W as loss



A Standard Speaker Verification Pipeline

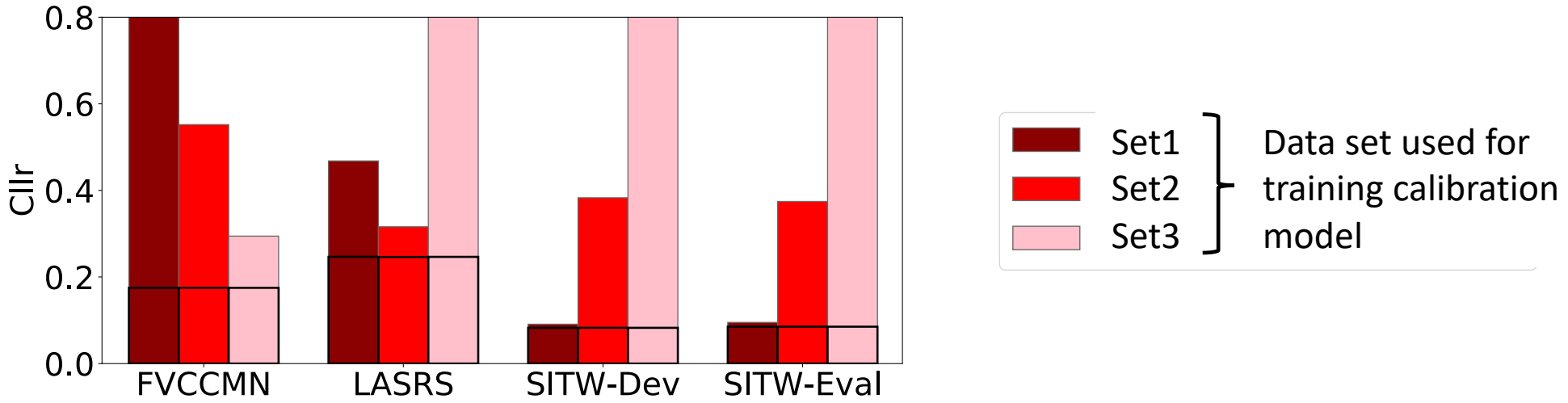


A Standard Speaker Verification Pipeline



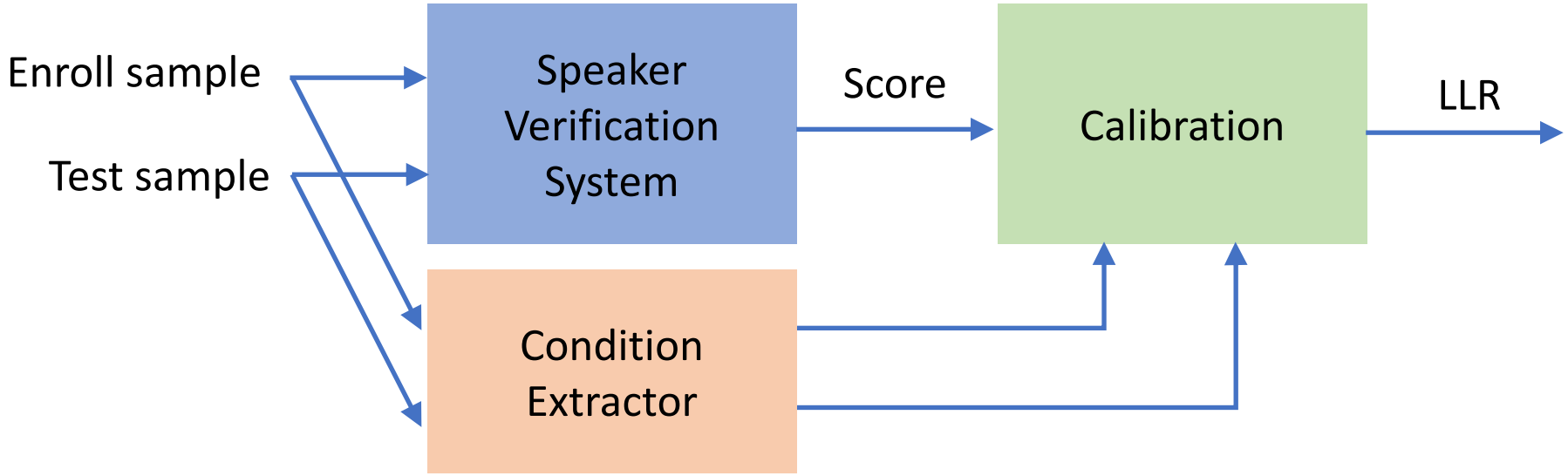
Calibration across Conditions

- Actual and min Cllr results for several datasets using different global calibration models
  - Black lines inside the bars indicate the minimum Cllr
- No model is good across the board!



Condition-Aware Calibration

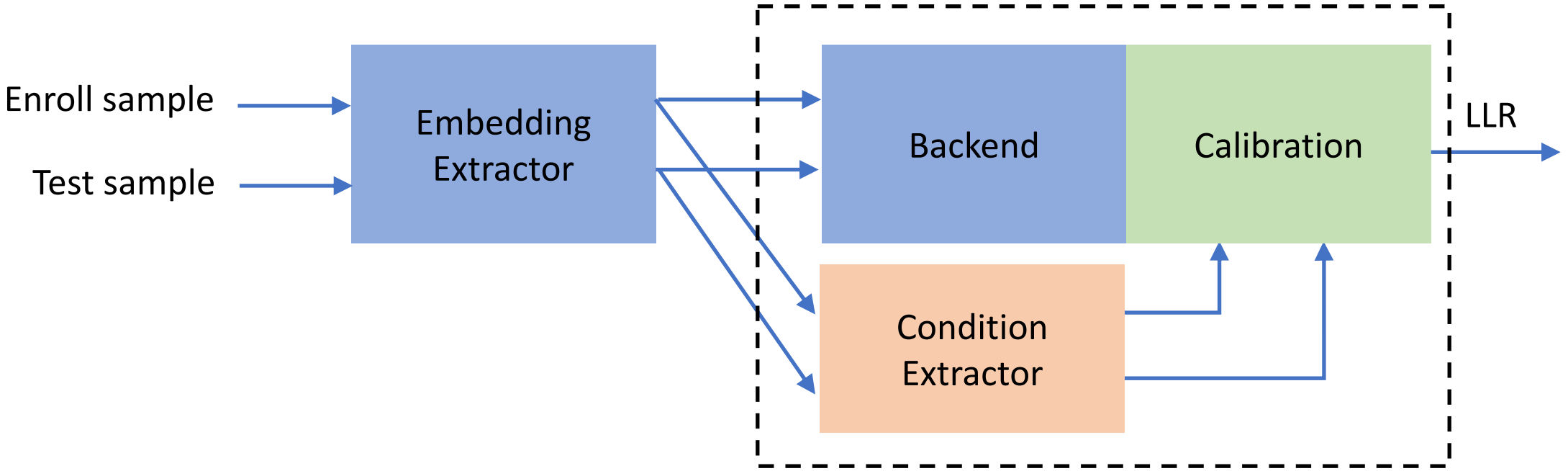
- A few approaches proposed in the literature over last two decades to solve this issue
  - Most assume an external class or vector representation (given or estimated) for the condition and used it to condition the calibration parameters



Solewicz and Koppel, “Considering speech quality in speaker verification fusion”, 2005
Mandasari et al, “Quality measures based calibration with duration and noise dependency for speaker recognition”, 2015
Nautsch et al, “Robustness of quality-based score calibration of speaker recognition systems ...”, 2016
Ferrer et al, “Toward fail-safe speaker recognition: Trial-based calibration with a reject option,” 2019

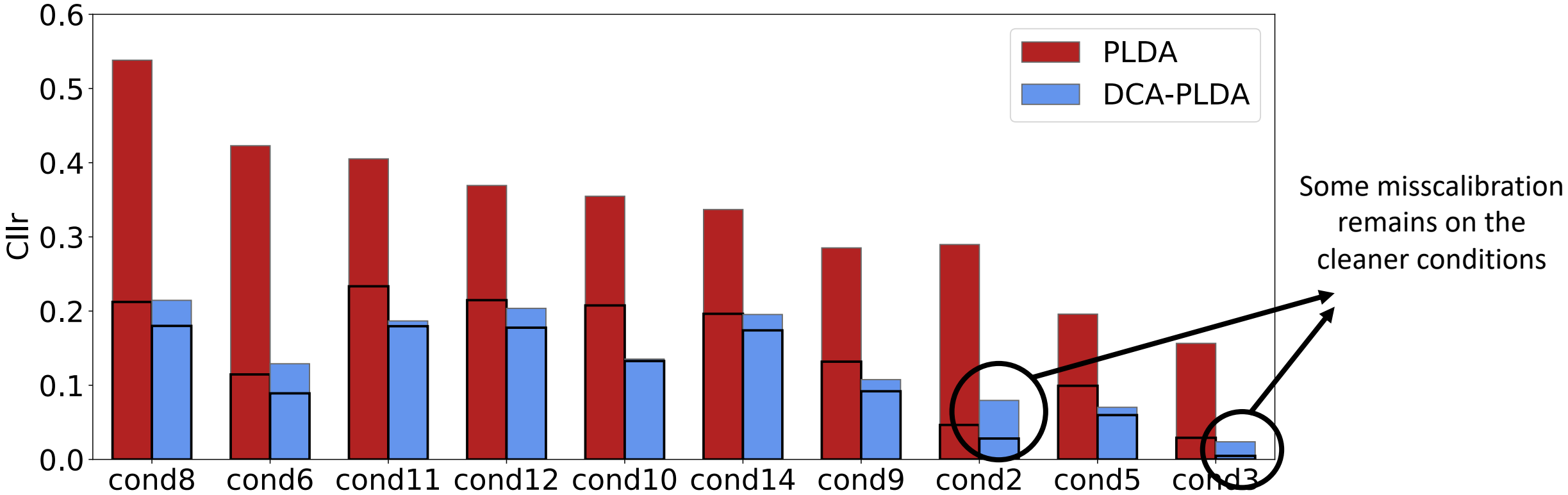
Condition-Aware Calibration

- Recently, we proposed an approach that jointly trains the backend and condition-dependent calibrator
  - Achieves excellent calibration across a wide variety of conditions



Ferrer, McLaren, Brummer, “A Speaker Verification Backend with Robust Performance across Conditions”, 2021

Condition-Aware Calibration



- Results on the FBI dataset, designed for work on calibration
- PLDA: a standard SV system with calibration stage trained on a subset of the training data
- DCA-PLDA: the discriminative condition-aware system from previous slide

Discussion

- Having calibrated scores is important
  - This is true even outside of the forensic realm!
  - SV systems are most commonly used to make hard decision
  - Calibrated scores let us make optimal decisions
- Bayes decision theory gives us a way to measure calibration
- If calibration is bad AND we have data matched to the eval scenario, we can fix it
- Alternatively, maybe we can work toward developing SV systems that do not require that extra step for every new condition
  - Some progress made in this direction
  - Yet, system is not yet ready for forensic use without proper validation

Thank you!