Metrics in VoicePrivacy & ASVspoof Challenges

Andreas Nautsch

vitas.ai

2021-08-02

Disclaimer: as always, you will see the individual view of an opinionated researcher; not the view(s) of associated institutions :)

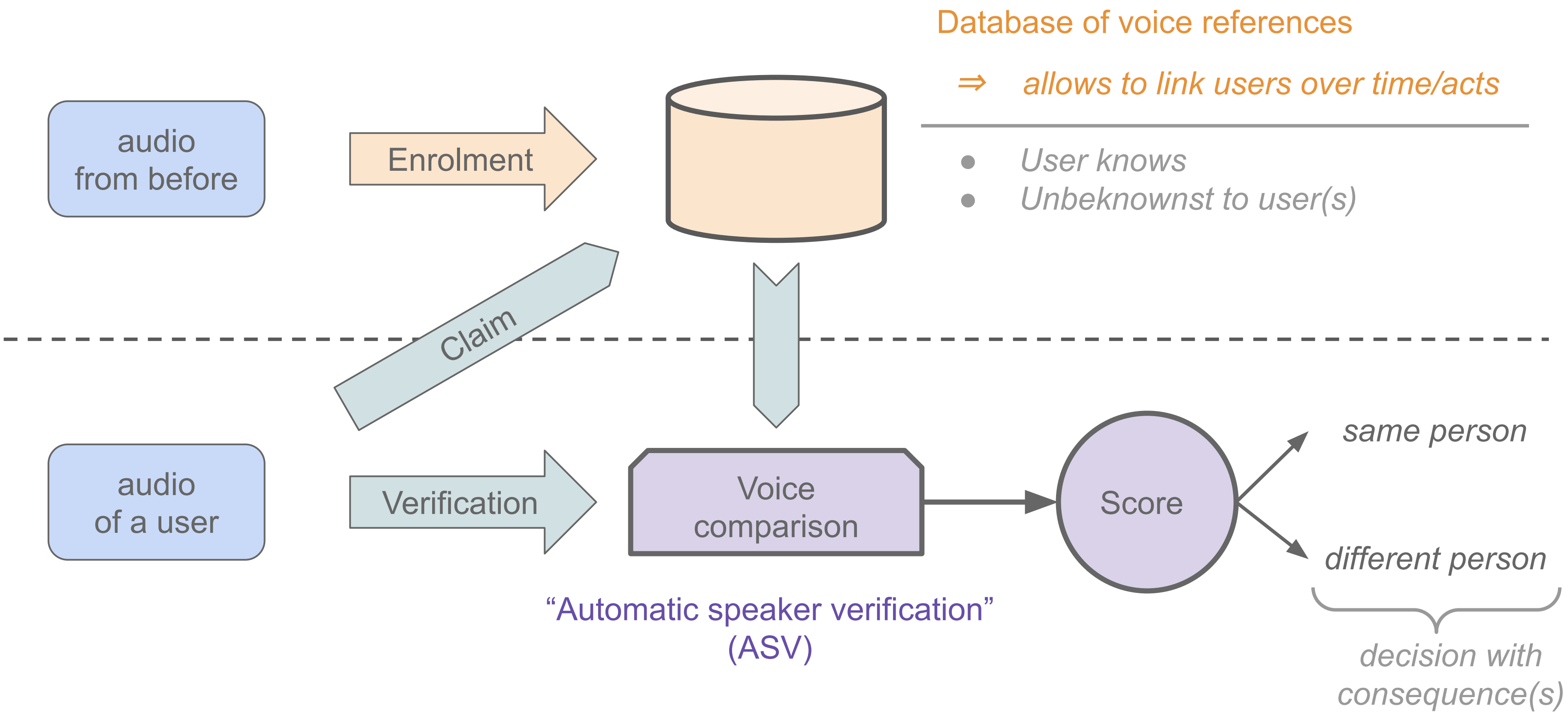
Outline

- Voice biometrics in a nutshell
- Security & privacy focus
- ASVspoof challenge: “t-DCF” metric ⇒ security in voice biometrics
- VoicePrivacy challenge: “ZEBRA” metric ⇒ privacy as <u>ANTI</u> voice biometrics

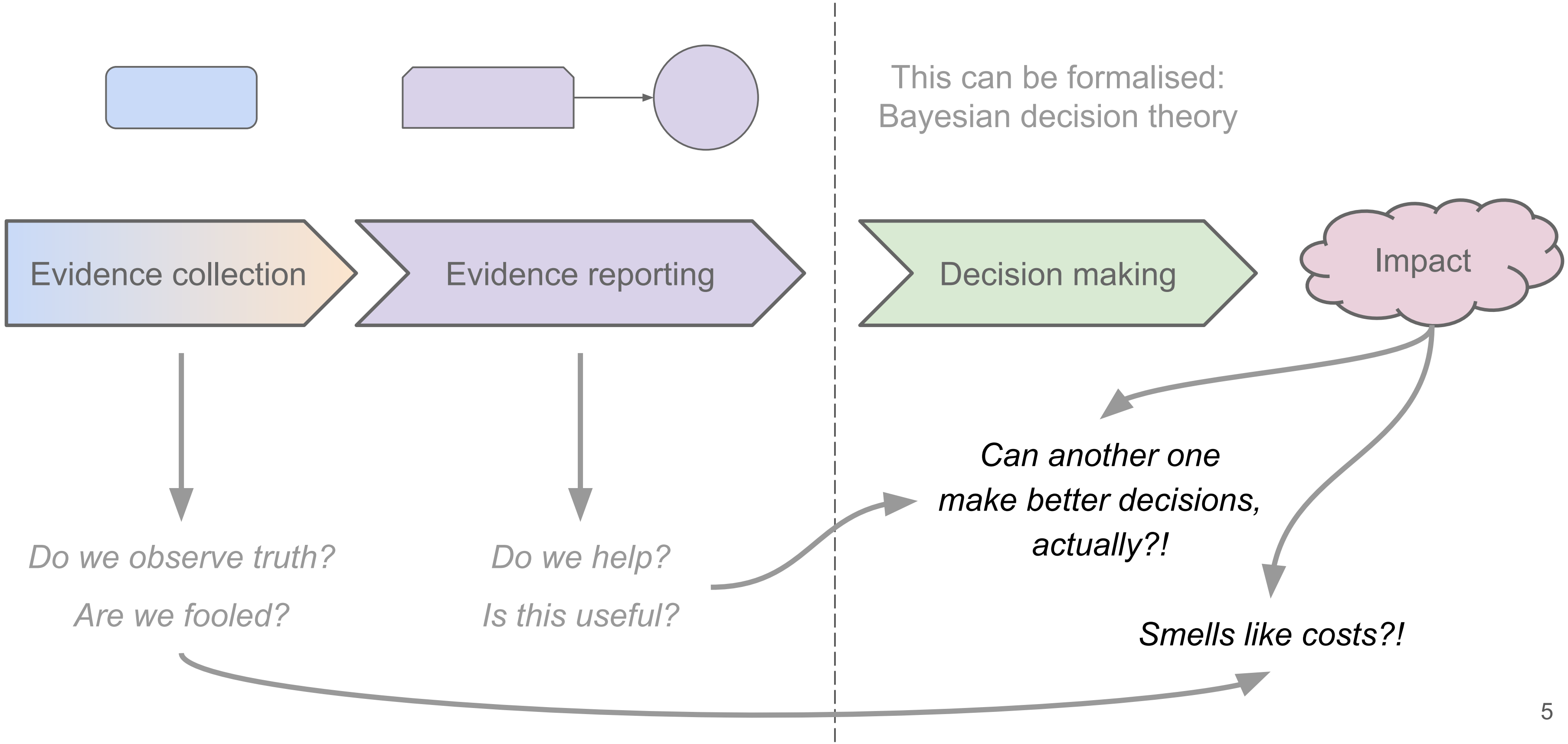
Biometrics with voice: WHO is speaking?

Same person?

Biometrics with voice: WHO is speaking?



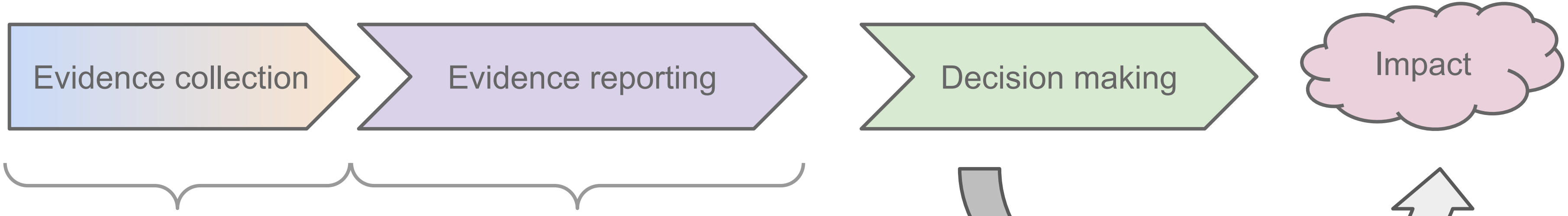
What’s actually going on here? *… think forensic sciences*



Security & privacy — voice biometrics two-ways

Focus: common evaluation methodology to the assessment of …

a) Security
b) Privacy



Spoofing detection

ASVspoof challenge

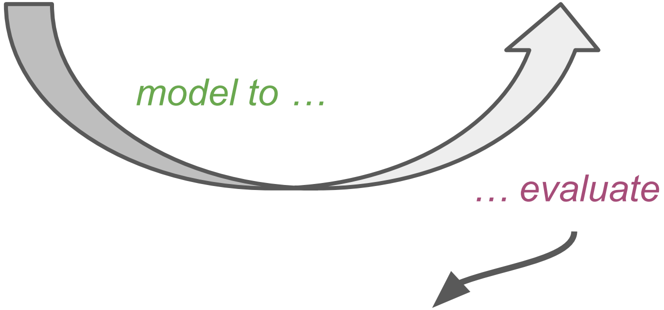
- Physical access
- Logical access
- Speech deepfakes

Zero evidence

VoicePrivacy challenge

- What was said? ☑️
- Who spoke? ☒

⇒ Modify raw audio



Focus here:

HOW to <u>evaluate</u> :)

Why security? — Why privacy?

Why security?

www.about.hsbc.co.uk/news-and-media/hsbc-uks-voice-id-prevents-gbp249-million-of-attempted-fraud

E.g. Fraud detection in online banking

- HSBC refers to £249 mio saved through voice biometrics
- Attacking voice biometrics is possible
- Needs to be prevented

Why privacy?

www.stasimuseum.de

E.g. surveillance through speech data

- Enabling human rights for individuals
- Bad example: GDR Ministry for State Security German (Stasi)
- Needs to be prevented

Automatic Speaker Verification Anti-Spoofing (ASVspoof)

Kinnunen et al.: "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals," IEEE/ACM TASLP 2020

DOI: 10.1109/TASLP.2020.3009494

https://arxiv.org/abs/2007.05979

ASVspoof metric: tandem detection cost function (t-DCF)

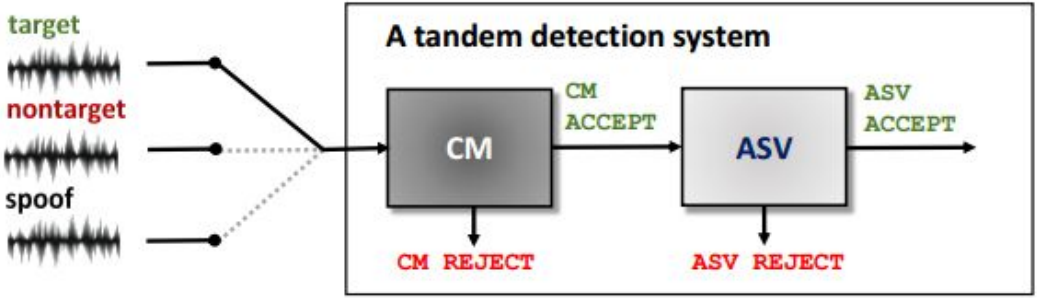
- Cascaded system design
  - ASV is given
  - Countermeasure (CM) ⇒ add-on security

- ASV classification task

  target vs. nontarget

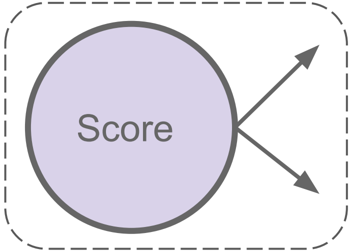
- CM classification task

  non-/target vs. spoof



- Evaluation: overall expected operational cost from employing ASV & CM

Looking glass: Bayesian decision theory

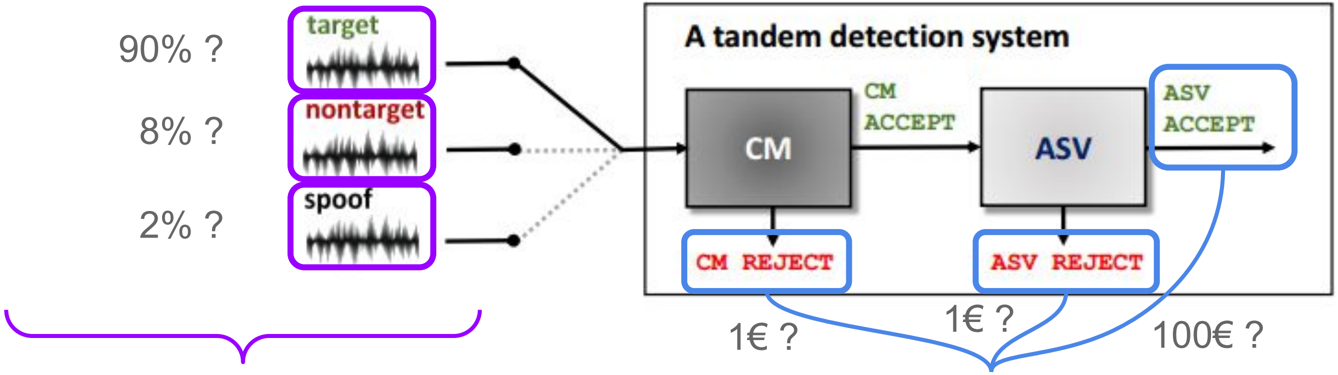


	Actual class	Tandem decision	Unit cost
a.	Target	REJECT (by ASV)	C_{miss}
b.	Nontarget	ACCEPT	C_{fa}
c.	Spoof	ACCEPT	$C_{\mathrm{fa,spoof}}$
d.	Target	REJECT (by CM)	C_{miss}

Actual class	Asserted prior
Target	π_{tar}
Nontarget	π_{non}
Spoof	π_{spoof}
	$\Sigma = 1$

How often do these happen?

// fa: false alarm



Prior probabilities

⇒ *subjective quantification*

// "policy" is trading-off beliefs

Risk costs

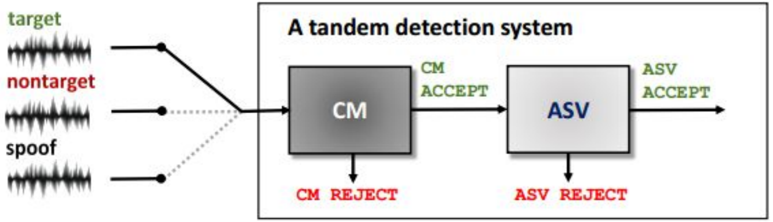
⇒ *subjective quantification*

// "policy" is trading-off beliefs

This can be formalised:
Bayesian decision theory

Decision making

t-DCF: at a glance



$$\boxed{\text{t-DCF} = \underbrace{C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot P_{\text{a}}} + \underbrace{C_{\text{fa}} \cdot \pi_{\text{non}} \cdot P_{\text{b}}} + \underbrace{C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}} \cdot P_{\text{c}}} + \underbrace{C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot P_{\text{d}}}}$$

Tandem action →	(CM ACCEPT, ASV REJECT)	(CM ACCEPT, ASV ACCEPT)	(CM ACCEPT, ASV ACCEPT)	(CM REJECT)
Actual class →	Target	Nontarget	Spoof	Target

$$P_{\text{c}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) \times P_{\text{fa,spoof}}^{\text{asv}}(\tau_{\text{asv}})$$

// fa: score ≥ threshold → P(...)

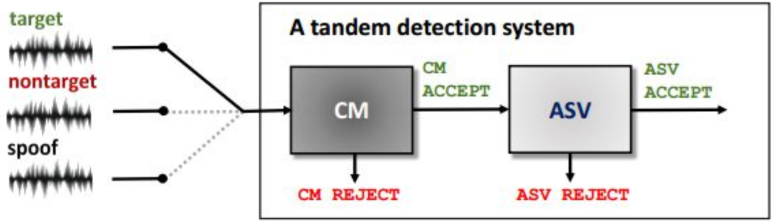
$$P_{\text{d}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$$

// miss: score < threshold → P(...)

$$P_{\text{b}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = (1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) \times P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}})$$

$$P_{\text{a}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = (1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) \times P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}})$$

How to compare t-DCFs of different priors/costs?



- Default: simulate coin tossing performance!
- Playing through the extrema…
  - CM & ASV: all-pass

$$C_{\text{fa}} \cdot \pi_{\text{non}} \cdot \boxed{1} + C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}} \cdot \boxed{1}$$

  - CM: no-pass

$$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot \boxed{1}$$

  - CM: all-pass & ASV: no-pass

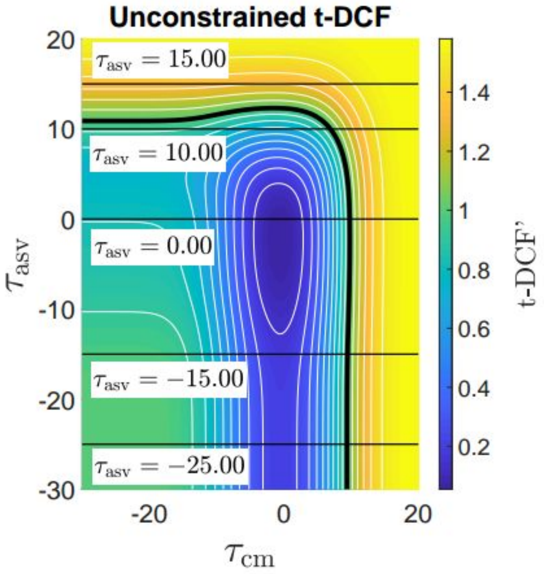
$$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot \boxed{1}$$

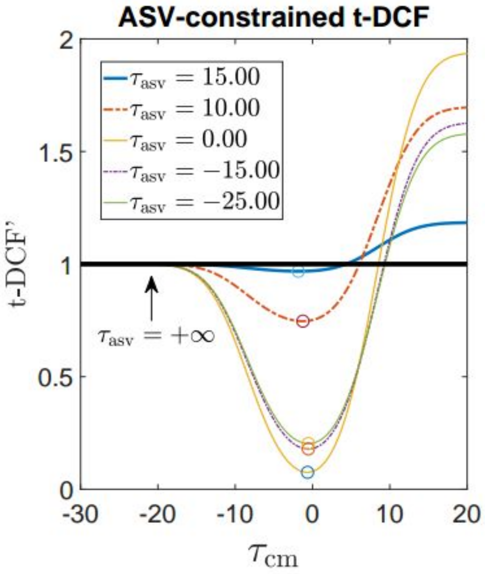
$$\text{t-DCF}'(\tau_{\text{cm}}, \tau_{\text{asv}}) = \frac{\text{t-DCF}(\tau_{\text{cm}}, \tau_{\text{asv}})}{\text{t-DCF}_{\text{default}}}$$

$$\text{t-DCF}'_{\text{min}} = \frac{\text{t-DCF}_{\text{min}}}{\text{t-DCF}_{\text{default}}} \leq \frac{\text{t-DCF}_{\text{min}}}{\text{t-DCF}_{\text{min}}} = 1$$

$$\text{t-DCF}_{\text{default}} = \min\{C_{\text{fa}} \cdot \pi_{\text{non}} + C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}}, C_{\text{miss}} \cdot \pi_{\text{tar}}\}$$

Synthetic scores; parameters as of ASVspoof 2019/21





ASV-constrained t-DCF

$$\text{t-DCF}(\tau_{\text{cm}}) = C_0 + C_1 P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) + C_2 P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})$$

$$C_0 = \pi_{\text{tar}} C_{\text{miss}} P_{\text{miss}}^{\text{asv}} + \pi_{\text{non}} C_{\text{fa}} P_{\text{fa}}^{\text{asv}}$$
$$C_1 = \pi_{\text{tar}} C_{\text{miss}} - (\pi_{\text{tar}} C_{\text{miss}} P_{\text{miss}}^{\text{asv}} + \pi_{\text{non}} C_{\text{fa}} P_{\text{fa}}^{\text{asv}})$$
$$C_2 = \pi_{\text{spoof}} C_{\text{fa,spoof}} P_{\text{fa,spoof}}^{\text{asv}}$$

VoicePrivacy

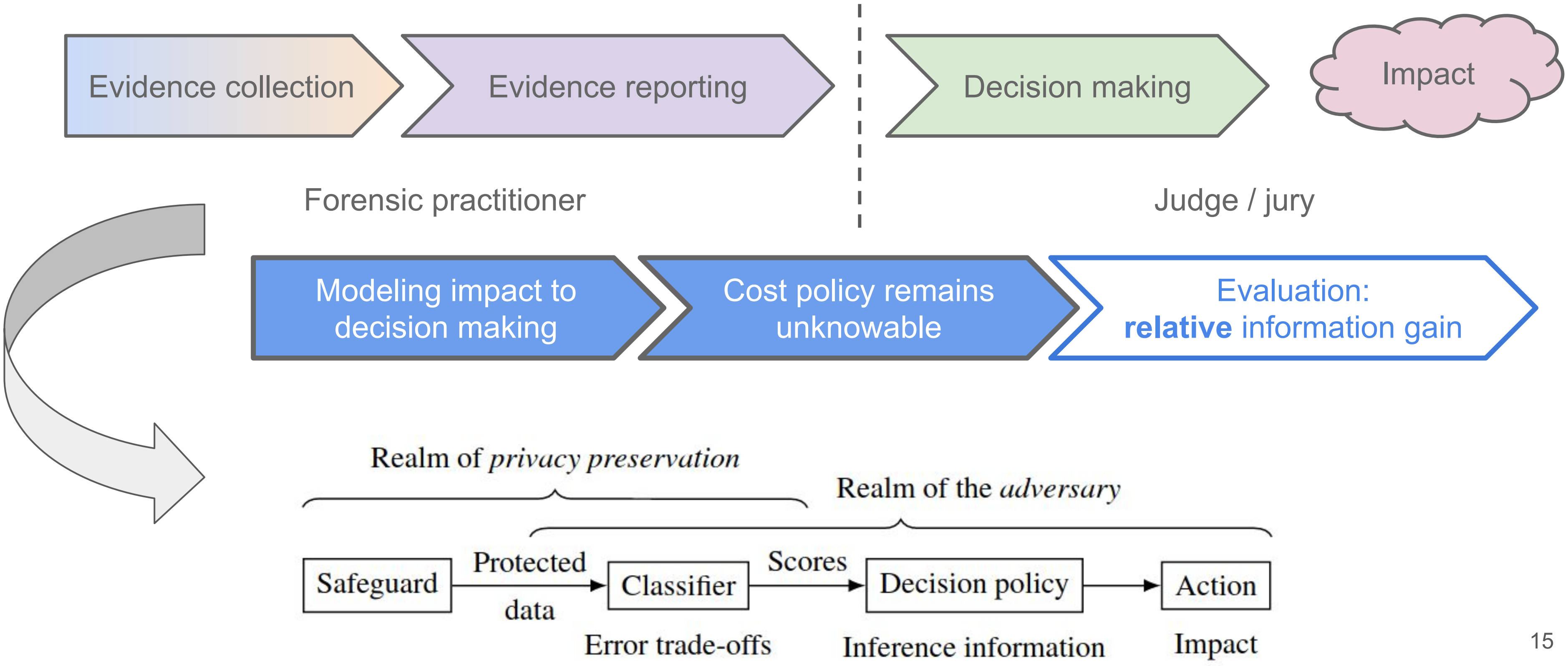
Nautsch et al.: “The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment,” Proc. Interspeech 2020

DOI: 10.21437/Interspeech.2020-1815

https://arxiv.org/abs/2005.09413

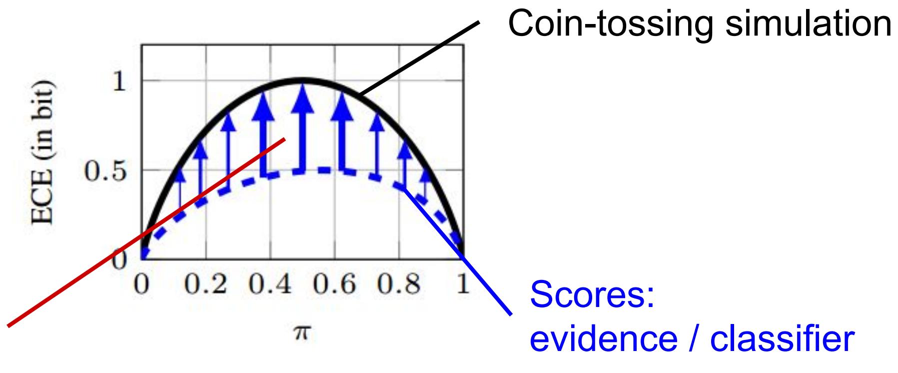
Picture taken in
Heidelberg Zoo, 2020

Motivation: evidence in court & decoupled provinces



Zero evidence “ZEBRA” framework: two metrics

- Expected privacy disclosure
  - Population level
  - Minimise empirical cross-entropy (ECE); regardless of prior probability



- Worst-case privacy disclosure
  - Individual level
  - Minimise strength-of-evidence; across data records

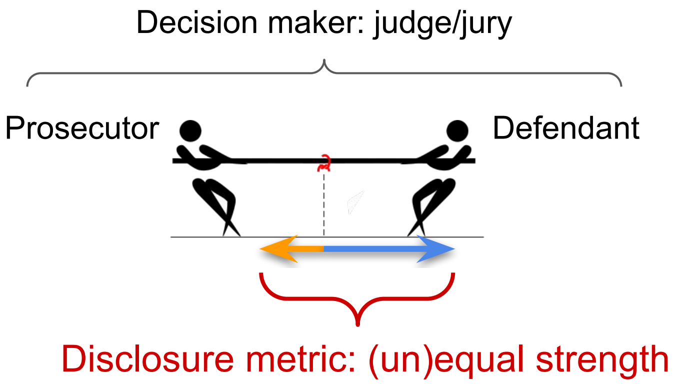


Figure based on wikimedia.org

Textbook: empirical cross-entropy (ECE) — step by step

Prior uncertainty

Evidence

Posterior uncertainty

// P(s | θ) is a research topic

RELATIVE information gain

Expectation across scores

Θ = { A: "same person" , B: "different person" }

P: reference probability space

$$H_P(\Theta) = -\sum_{\theta\in\Theta} P(\theta)\log_2 P(\theta)$$

$$H_P(\Theta\,|\,S) = -\sum_{\theta\in\Theta} P(\theta)\int_s P(s\,|\,\theta)\log_2 P(\theta\,|\,s)\,\mathrm{d}s$$

Set of scores

$$H_{P||\tilde{P}}(\Theta\,|\,S) = -\sum_{\theta\in\Theta} P(\theta)\int_s P(s\,|\,\theta)\log_2 \tilde{P}(\theta\,|\,s)\,\mathrm{d}s$$

$$P(s\,|\,\theta) \approx |S_\theta|^{-1}$$

reference and classifier prior values $P(\theta), \tilde{P}(\theta)$

$\pi = P(\mathcal{A}) = \tilde{P}(\mathcal{A})$ and $1-\pi = P(\mathcal{B}) = \tilde{P}(\mathcal{B})$

Well-calibrated scores (strength-of-evidence)

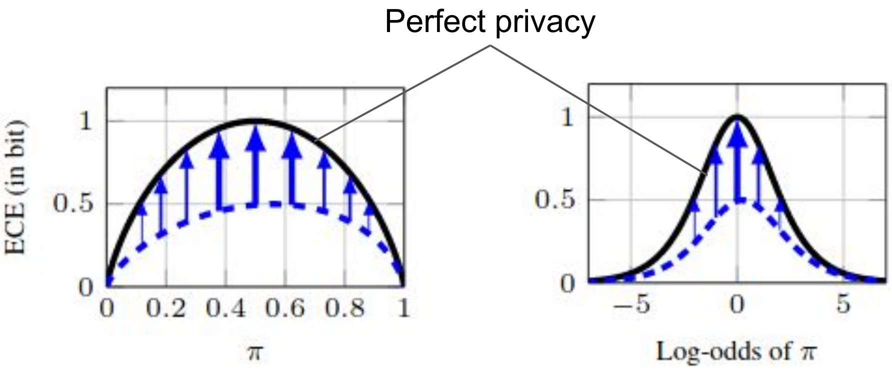
"the classifier"

$$\mathrm{ECE}(\Theta\,|\,\mathcal{S}) := \frac{\pi}{|\mathcal{S}_\mathcal{A}|}\sum_{a\in\mathcal{S}_\mathcal{A}}\log_2\left(1+\frac{1-\pi}{a\pi}\right) + \frac{1-\pi}{|\mathcal{S}_\mathcal{B}|}\sum_{b\in\mathcal{S}_\mathcal{B}}\log_2\left(1+\frac{b\pi}{1-\pi}\right)$$

Integrate π out

Ramos & Gonzalez-Rodriguez: Cross-entropy Analysis of the Information in Forensic Speaker Recognition, in Proc. Odyssey, 2008
Ramos, Franco Pedroso, Lozano-Diez, Gonzalez-Rodriguez: Deconstructing Cross-Entropy for Probabilistic Binary Classifiers, Entropy 20(3), 2018

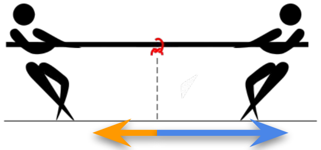
Shannon’s perfect secrecy to strength-of-evidence



(a) *ZEBRA idea* (b) *ECE plot*

On the highest strength-of-evidence

- Basic idea
  - Sustain probabilistic interpretation of scores
  - Account for binary decision setting // P(“yes”) = 1 - P(“no”)
  - Take the highest strength-of-evidence
  - Keep in mind the world is larger than one dataset
    ⇒ apply Laplace’s rule of succession &
    return a prediction of the worst case disclosure



- Make reporting digestible, lessons from forensic sciences
  - Everyone interprets numbers & ratios differently
  - Thus: categorical tags & scale

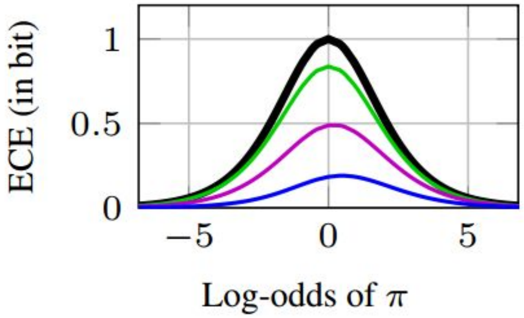
Tag	Category	Posterior odds ratio (flat prior)
0	$l = 1 = 10^0$	50 : 50 (flat posterior)
A	$10^0 < l < 10^1$	more disclosure than 50 : 50
B	$10^1 \leq l < 10^2$	one wrong in 10 to 100
C	$10^2 \leq l < 10^4$	one wrong in 100 to 10 000
D	$10^4 \leq l < 10^5$	one wrong in 10 000 to 100 000
E	$10^5 \leq l < 10^6$	one wrong in 100 000 to 1 000 000
F	$10^6 \leq l$	one wrong in at least 1 000 000

VoicePrivacy 2020 challenge; an example

Expected privacy disclosure (population)

Worst-case privacy disclosure (individual)

Categorical tag



Perfect privacy (0.00, 0.00, 0) … coin-tossing

B1 (0.11, 2.27, C) … a DNN baseline

B2 (0.36, 3.58, C) … signal processing (on-the-fly)

Unprotected data (0.58, 3.98, C) … the x-vector DNN (SOTA)

Wrapping up …

Summary

- One framework, two application spaces
  ⇒ Bayesian decision theory

- Security focus: ASVspoof challenge
  ⇒ tandem detection cost function (t-DCF)

- Privacy focus: VoicePrivacy challenge
  ⇒ zero evidence “ZEBRA” with expectation & worst-case metrics

Take-home message(s)

- Towards a holistic approach
  - Think interdisciplinary for solutions
  - Develop multidisciplinary skills

- Expectation is not the sole metric
  - One might not know all parameters all the time — theory & models are indispensable
  - Consider the worst-case — avoid running into marginalising societies

- “Privacy as anti-biometrics”

  ⇒ we need more conversations across fields :)

  // there’s so much more in speech than biometrics alone