Metrics in Audio Security & Privacy

Andreas Nautsch
EUROCOM

TReSPAsS-ETN TE2
2021-01-21/22 — Virtual Event

Outline

- Audio Security
  - Automatic Speaker Verification Anti-Spoofing Challenge 2019 (ASVspoof 2019)
    https://www.asvspoof.org/
  - Kinnunen et al.: “Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals,” IEEE/ACM TASLP 2020, DOI: 10.1109/TASLP.2020.3009494

- Audio Privacy
  - VoicePrivacy 2020 Challenge
    https://www.voiceprivacychallenge.org/
  - Nautsch et al.: “The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment,” Proc. Interspeech 2020

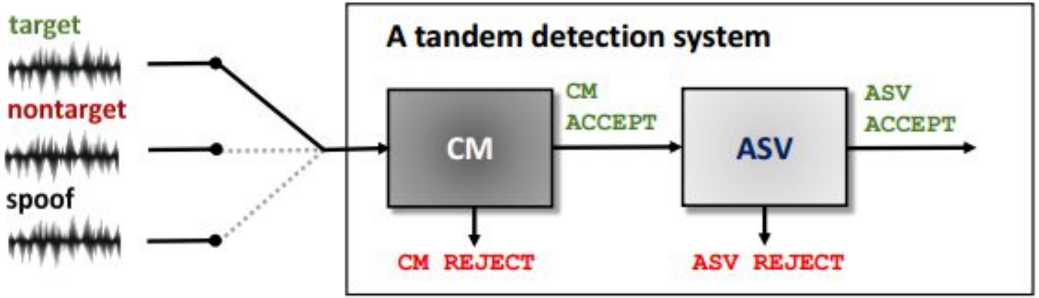
Audio Security Metric

tandem Decision Cost Function (t-DCF)

Speech	⇔	SC37 dictionary :)
tar	~	mated, bona fide, …
non	~	non-mated, non-attack, ….
spoof	~	presentation attack, logical access spoof, …
miss	~	FRR, FNMR, BPCER, …
false alarm (fa)	~	FAR, FMR, APCER, …

Audio Security: The Setting

- Anti-Spoofing
  - “Physical Access” Replay attacks (see Voice PAD)
  - “Logical Access” Voice synthesis/morphing/conversion attacks (not PAD)
- In-Scope
  - Tandem operation of countermeasure (CM) and ASV sub-systems
  - Throughout formalised assessment
- Out-Scope
  - Informal descriptors by error rates
  - Purely CM-focused performance



Audio Security: Expected Cost as Metric

- Quantification of beliefs
  - What is the impact of a decision outcome?
  - How likely is a decision outcome?
- Expected class discrimination risk
  - $\mathbb{E}$ [risk | costs, class priors, classification rates]
  - Risk minimisation: sweep thresholds, take minimum

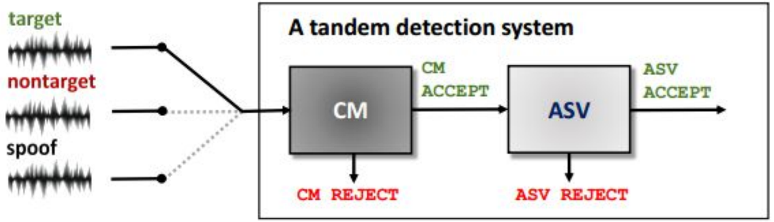
	Actual class	Tandem decision	Unit cost
a.	Target	REJECT (by ASV)	C_{miss}
b.	Nontarget	ACCEPT	C_{fa}
c.	Spoof	ACCEPT	$C_{\text{fa,spoof}}$
d.	Target	REJECT (by CM)	C_{miss}

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

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

	$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot P_{\text{a}}$	$C_{\text{fa}} \cdot \pi_{\text{non}} \cdot P_{\text{b}}$	$C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}} \cdot P_{\text{c}}$	$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot P_{\text{d}}$
Tandem action →	(CM ACCEPT, ASV REJECT)	(CM ACCEPT, ASV ACCEPT)	(CM ACCEPT, ASV ACCEPT)	(CM REJECT)
Actual class →	Target	Nontarget	Spoof	Target

Audio Security: Tandem Classification Rates



$C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}} \cdot P_{\text{c}}$

(CM ACCEPT, ASV ACCEPT)

Spoof

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

$C_{\text{fa}} \cdot \pi_{\text{non}} \cdot P_{\text{b}}$

(CM ACCEPT, ASV ACCEPT)

Nontarget

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

$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot P_{\text{a}}$

(CM ACCEPT, ASV REJECT)

Target

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

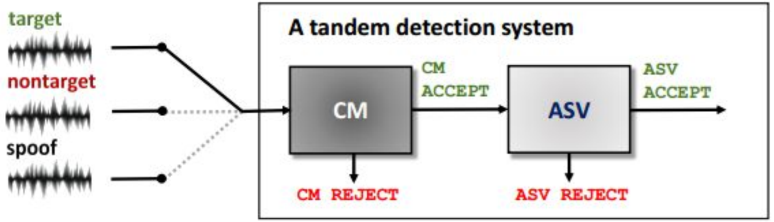
$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot P_{\text{d}}$

(CM REJECT)

Target

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

Audio Security: Metric Normalisation



- Better comparability (other costs/priors)
- What are the extreme actions?
  - 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

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

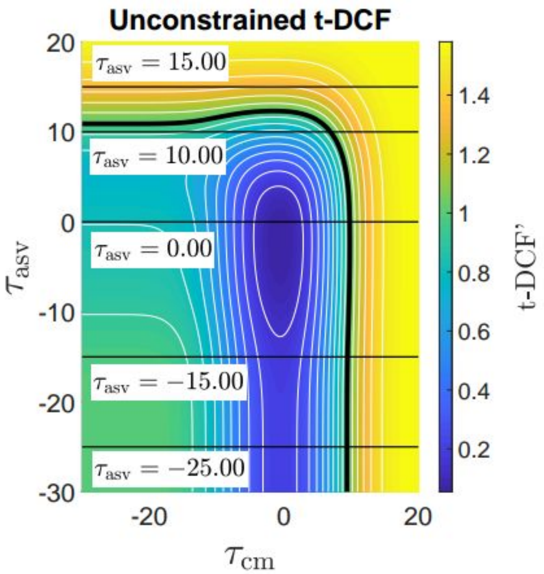
$$\text{t-DCF}'(\tau_{\text{cm}}, \tau_{\text{asv}}) = \frac{\text{t-DCF}(\tau_{\text{cm}}, \tau_{\text{asv}})}{\text{t-DCF}_{\text{default}}}$$

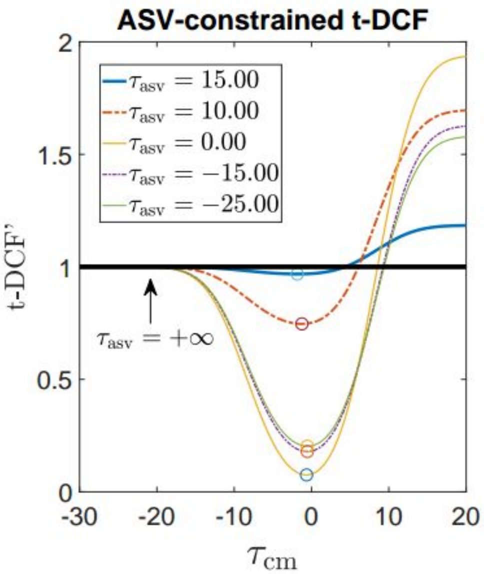
$$\text{t-DCF}'_{\text{min}} = \frac{\text{t-DCF}_{\text{min}}}{\text{t-DCF}_{\text{default}}} \leq \frac{\text{t-DCF}_{\text{min}}}{\text{t-DCF}_{\text{min}}} = 1$$

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

Audio Security: t-DCF Examples

- ASVspoof 2019 Challenge
  - Cost & prior parameters as per challenge
  - Synthetic ASV & CM scores





ASV-constrained t-DCF

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

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

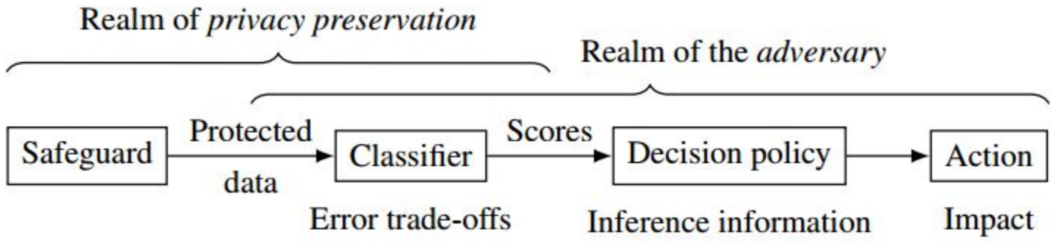
Audio Privacy Metric

Zero Evidence
Biometric Recognition Assessment
(The Privacy ZEBRA)

Picture taken in
Heidelberg Zoo, 2020

Audio Privacy: The Setting

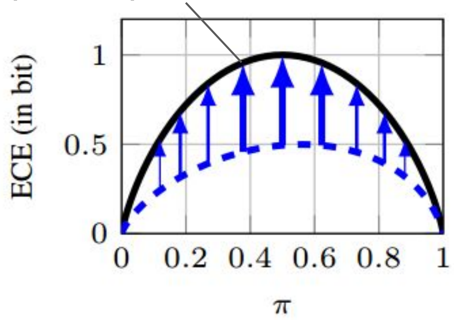
- Pseudomise audio speech data
- Decoupling layers & taking the perspective of an adversary



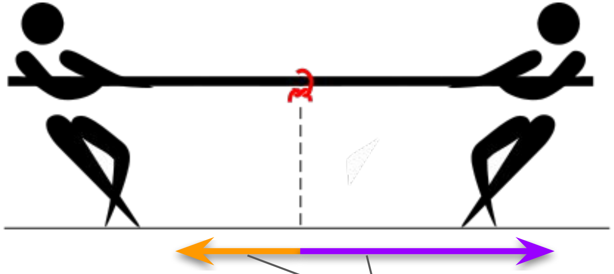
- Existing metrics do not suffice!
  - Zero-knowledge proofs are unavailable.
  - EER is the worst possible decision policy that an adversary can take for herself.
  - Unlinkability (not devised for this setting) — identity confirmation but not short-listing.
  - Any fixed error rate/cost metric prejudices privacy disclosure impacts to an individual.

Audio Privacy: Zero Evidence as Metric

Coin-tossing simulation:
all scores are equal
“prior = posterior ECE”



- Population level: Empirical Cross-Entropy (ECE)
  - Idea: prior entropy ⇒ evidence ⇒ posterior entropy
  - Cross-entropy of classification by scores from ground truth
  - Zero evidence: prior ECE = posterior ECE, regardless of prior π
- Individual level: Zero Strength of Evidence
  - Forensic sciences: likelihood ratio
  - Who is stronger: prosecutor or defendant?



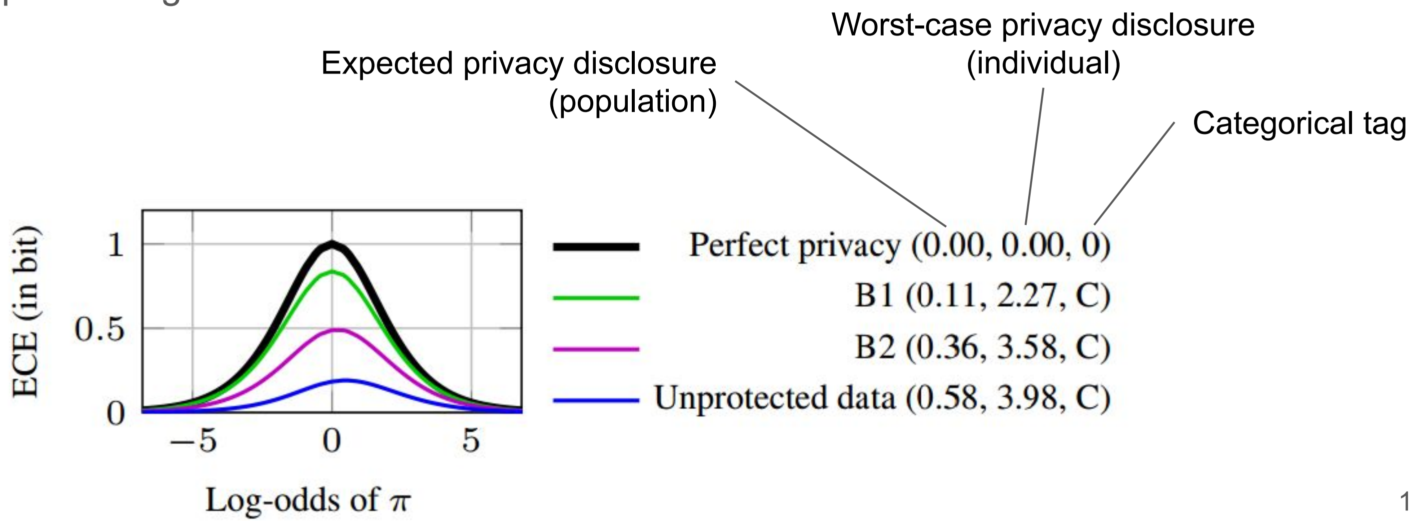
Tag	Category	Posterior odds ratio (flat prior)
0	$l = 1 = 10^0$	50 : 50 (flat posterior)
A	$10^0 < l < 10^1$	more disclosure than 50 : 50
B	$10^1 \leq l < 10^2$	one wrong in 10 to 100
C	$10^2 \leq l < 10^4$	one wrong in 100 to 10 000
D	$10^4 \leq l < 10^5$	one wrong in 10 000 to 100 000
E	$10^5 \leq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

Figure based on wikimedia.org

Ideal if equal — across individuals: worst-case privacy disclosure?

Audio Privacy: ZEBRA Examples

- VoicePrivacy 2020 Challenge
  - Task: speech recognition should work — voice biometrics not ⇒ modification of raw audio
  - ASV: pre-trained kaldi x-vector recipe
  - B1: DNN baseline
  - B2: signal processing baseline



https://gitlab.eurecom.fr/nautsch/zebra

Summary & Conclusion

- Summary
  - Audio security: cost-based approach for expected risk minimization
  - Audio privacy: relative information & strength of evidence approach
- Conclusion
  - Constrained cost as a guide for the CM optimization given a biometric system

    ⇒ taking a holistic perspective
  - Audio privacy must achieve <u>privacy for every single one</u>; *are we a marginalising society?*

    ⇒ expectation & worst-case estimates
- Security vs. Privacy: seek positive-sum solutions; not zero-sum solutions