The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment

Andreas Nautsch[1], Jose Patino[1], Natalia Tomashenko[2], Junichi Yamagishi[3],
Paul-Gauthier Noé[2], Jean-François Bonastre[2], Massimiliano Todisco[1] and Nicholas Evans[1]

[1] Digital Security Department, EURECOM, France
[2] Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
[3] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

nautsch@eurecom.fr

Abstract

Mounting privacy legislation calls for the preservation of privacy in speech technology, though solutions are gravely lacking. While evaluation campaigns are long-proven tools to drive progress, the need to consider a privacy *adversary* implies that traditional approaches to evaluation must be adapted to the assessment of privacy and privacy preservation solutions. This paper presents the first step in this direction: *metrics*.

We introduce the zero evidence biometric recognition assessment (ZEBRA) framework and propose two new privacy metrics. They measure the *average* level of privacy preservation afforded by a given safeguard for a population and the *worst-case* privacy disclosure for an individual. The paper demonstrates their application to privacy preservation assessment within the scope of the VoicePrivacy challenge. While the ZEBRA framework is designed with speech applications in mind, it is a candidate for incorporation into biometric information protection standards and is readily extendable to the study of privacy in applications even beyond speech and biometrics.

1. Introduction

Spoken language contains a wealth of personal information including the biometric identity [1, 2]. Such sensitive information is clearly susceptible to being exploited for unscrupulous and ethically reprehensible purposes. Unsurprisingly, speech data falls within the scope of recent privacy legislation, e.g., the California Consumer Privacy Act (CCPA) [3] in the US (effective January 1, 2020), the European General Data Protection Regulation (GDPR) [4] (implemented May 25, 2018) and EU directive 2016/680 [5] (the Police Directive, also implemented May 25, 2018). Privacy is a fundamental human right [6] and the failure to respect privacy legislation can attract significant fines. Speech technology providers and operators are thus obliged to ensure adequate provisions for privacy preservation.

There are two general approaches to guard against privacy intrusions in the case of speech data. The first is to protect access to speech data, usually via some form of encryption and secure computation. The second, and the focus in this article, is to strip the speech signal of personally identifiable information such that it cannot be linked (with some level of certainty) to a specific individual. Pseudonymisation, de-identification and anonymisation are all examples of such approaches, but are all relatively embryonic research topics within the speech community; only few solutions have been proposed thus far.

One reason for why progress in privacy preservation has not kept pace with legislation is due to the lack of frameworks for assessment. While privacy preservation solutions may take many different forms, their goals are common: they should prevent the use of speech being used to infer identity. Accordingly, the same or similar metrics can be used for the assessment of any form of privacy preservation solution. The design of such metrics is then a priority and stands to boost progress in privacy preservation whatever the particular approach.

Upon first consideration, metrics for the assessment of privacy preservation solutions may seem straightforward. This is not the case, however, since many obvious metrics do not reflect the *decision policy of a privacy adversary*. Consequently, they will give a misleading measure of privacy. Inspired by metrics used in forensics research, this paper reports our proposals for two different privacy preservation metrics that disentangle the considerations of the privacy safeguard and the privacy adversary. The paper shows how the *empirical cross entropy* and the *strength of evidence* can be harnessed within a so-called *zero evidence biometric recognition assessment* (ZEBRA) framework to measure the expected and worst-case privacy disclosure for a given privacy preservation solution.

The motivation for this work is described in Section 2. Section 3 describes background work and the empirical cross-entropy. Section 4 describes its use in the ZEBRA framework and demonstrates its application to the assessment of privacy preservation solutions within the context of the VoicePrivacy 2020 challenge [7, 8]. A discussion of the work and directions for the future are presented in Section 5.

2. Towards empirical privacy metrics

Privacy metrics are needed in order to gauge and to compare the level of privacy preservation offered by different solutions. Such a metric should also reflect the *gain in privacy* delivered by a given safeguard as well as the *remaining potential for privacy disclosure*. We also seek metrics which reflect not just the *average* level of privacy preservation provided by a particular solution, but one that can also be used to understand the *variation* in privacy preservation provided to a population; there may be differences in the level of protection provided to different users. Finally, metrics should be based not upon the prior beliefs and costs of a privacy preservation system designer or evaluator but should, instead, reflect those of a privacy adversary. Only then, can we gain meaningful insights to privacy and the gap to *perfect privacy*.

Perfect privacy was introduced as *perfect secrecy* by Shannon [9]: the posterior probabilities of intercepted data are identical to the prior probabilities of an adversary. This led to *theoretically* founded assessment of privacy safeguards in modern cryptography [10] (*zero knowledge proofs*); here, input data has a *mathematical definition*. Speech data is different (we use *models*, not *definitions*), hence we seek *empirical* approaches to assessment. Unfortunately, despite some obvious candidates, existing empirical metrics do not meet the above requirements.

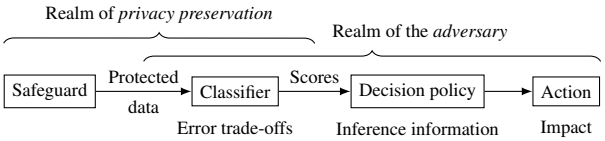


Figure 1: *Decoupling the classifier, decision policy and action to estimate privacy from the perspective of an adversary.*

Obvious candidates are to measure the, e.g., equal error rate (EER), *unlinkability* of protected biometric datasets [11], cost risks and application-independent risks by detection cost functions (DCFs) [12] and the goodness of so-called log-likelihood ratio (LLR) scores, C_{llr} [13][1]; all are ill suited to the privacy scenario.[2] The envisaged scenario is illustrated in Fig. 1. Speech data is processed according to some form of privacy preservation algorithm (safeguard) to suppress speaker-discriminative information. The resulting data is used by an adversary who still seeks to infer the speaker's identity. The adversary does this using some biometric classifier which assesses data and reports on the *strength-of-evidence* in the form of scores. In this scenario, the classifier can be in the realm of either the privacy preserver, or the privacy adversary. If the privacy adversary were to use the best technology available to them, then a classifier can be used within the realm of the privacy preserver so long as it is representative of the state of the art. This is where the realm of the privacy preserver ends.

From here on, everything is within the realm of the privacy adversary, not the privacy preserver. A privacy metric must hence reflect the adversary's decision policy (what action to take) parameterised by their prior and cost beliefs. It is then necessary to assume minimum knowledge, or maximum uncertainty. Whereas priors can be simulated, costs of privacy infringement cannot. In this case, the *prior of the prior* is assumed to be uniformly distributed. Accordingly, classifier outputs are assumed to be likelihood ratios, obtained by the *pool adjacent violators to LLR algorithm* [15, 16, 17], a non-linear transform of observed scores resulting in *oracle score calibration*.

The scenario described above may be familiar to some readers for it resembles that of a (counter) forensics scenario. The counterparts here are the forensic practitioner (the privacy preserver) and the judge/jury (the privacy adversary). The forensic practitioner must present evidence only; s/he must not make decisions. The decision policy is that of the judge/jury; it is unknown to the forensic practitioner who must assume maximum uncertainty of prior/cost beliefs so as not to encroach upon the province of the court. While they provide the basis for the work presented in this paper, even the metrics used by the forensics community are not sufficient for the assessment of privacy. This is because the forensic practitioner simulates the empirical cross-entropy over different prior values only, without quantifying either the expected or highest strength-of-evidence (worst-case disclosure), which we need for assessing privacy.

[1]The C_{llr} metric also relates to empirical cross-entropy (see Section 3) for one specific prior; *a computational co-incidence* [14].

[2](1) By its very definition, the EER reflects a privacy adversary's worst possible decision policy [15]; the EER will hence reflect an unduly optimistic estimate of privacy protections. (2) Unlinkability considers the potential of evidence to *confirm* an identity, it overlooks the relevance to privacy of evidence that might *exclude* an identity. The exclusion of identities is still evidence, however, and so the unlinkability metric also gives a potentially distorted measure of privacy protections. (3) For cost based metrics, the impact of privacy disclosure depends on subcultures, such that we cannot possibly treat cost impacts.

3. Empirical cross entropy

This section presents a brief overview of empirical cross entropy (ECE), a metric developed for the assessment of forensic speaker recognition systems and originally reported in [14, 18, 19]. It ends with a discussion of how the ECE can be adapted to assess the performance of privacy preserving systems.

A speaker recognition system is furnished with two utterances, one for training with known speaker identity and one for testing an identity claim. There are two propositions $\mathcal{A}, \mathcal{B}$, namely that either the identity of the speakers in each utterance is the same $\mathcal{A}$, or that the speakers are different $\mathcal{B}$. We denote the set of propositions $\Theta = \{\mathcal{A}, \mathcal{B}\}$ and refer to a single proposition as $\theta \in \Theta$. Before using the speaker recognition system, we have some prior belief in the truth of each proposition, which we quantify by the prior probability $P(\theta)$. We denote two probability spaces: the ground truth or *reference* probability $P(\cdot)$, and the *classifier* probability $\tilde{P}(\cdot)$; the latter is the forecast whose desired value is the reference. The prior entropy in making a binary decision $H_P(\Theta)$ is:

$$H_P(\Theta) = -\sum_{\theta\in\Theta} P(\theta)\log_2 P(\theta). \tag{1}$$

The recognition system compares the training utterance to the test utterance in order to compute (ideally calibrated) recognition scores S. The goal of an attacker is to reduce *prior posterior entropy* about the propositions Θ by updating the prior with observed scores S which results in the *posterior entropy*, see [20]. The reference *posterior entropy* $H_P(\Theta \mid S)$ is:

$$H_P(\Theta \mid S) = -\sum_{\theta\in\Theta} P(\theta)\int_s P(s \mid \theta)\log_2 P(\theta \mid s)\,\mathrm{d}s. \tag{2}$$

Because we are in an empirical setting, it is not possible to derive reference likelihoods $P(s \mid \theta)$ from a theoretical footing; in general, they remain unknown [14, 19]. We can, however, quantify the *cross-entropy* $H_{P||\tilde{P}}(\Theta \mid S)$ between the posterior distributions of a system $\tilde{P}$ and the reference P:

$$H_{P||\tilde{P}}(\Theta \mid S) = -\sum_{\theta\in\Theta} P(\theta)\int_s P(s \mid \theta)\log_2 \tilde{P}(\theta \mid s)\,\mathrm{d}s. \tag{3}$$

We can approximate $P(s \mid \theta) \approx |S_\theta|^{-1}$ for a *large number* of classifier posteriors $\tilde{P}(\theta \mid s)$, where S_θ denotes the set of scores for class θ and $|S_\theta|$ is its size. In the forensic setting, systems do not compute $\tilde{P}(\theta \mid s)$ directly, since priors $\tilde{P}(\theta)$ are decoupled. The choice of reference and classifier prior values $P(\theta), \tilde{P}(\theta)$ is external to the classifier and considered a parameter: $\pi = P(\mathcal{A}) = \tilde{P}(\mathcal{A})$ and $1-\pi = P(\mathcal{B}) = \tilde{P}(\mathcal{B})$. Consequently, systems estimate likelihood ratio (LR) scores $S = \frac{\tilde{P}(X \mid \mathcal{A})}{\tilde{P}(X \mid \mathcal{B})}$ from features X. The ECE is computed as [14]:

$$\begin{aligned}\mathrm{ECE}(\Theta \mid \mathcal{S}) := &\frac{\pi}{|\mathcal{S}_\mathcal{A}|}\sum_{a\in\mathcal{S}_\mathcal{A}}\log_2\left(1+\frac{1-\pi}{a\,\pi}\right)\\ &+\frac{1-\pi}{|\mathcal{S}_\mathcal{B}|}\sum_{b\in\mathcal{S}_\mathcal{B}}\log_2\left(1+\frac{b\,\pi}{1-\pi}\right).\end{aligned} \tag{4}$$

It can be shown that the ECE reflects the expected amount of additional information that is needed in order to know the true proposition θ. If the classifier is unreliable (it performs poorly, requiring more information), then the ECE will be higher than if the classifier is more reliable.

4. Zero evidence framework

The goal of privacy preservation is to strip a speech utterance of personally identifiable information such that an adversary cannot recognise the identity of a speaker from an protected recording of their voice. In terms of speaker recognition, it should not be possible to match with certainty a training utterance to an anonymised test utterance. The adversary has some prior belief and seeks to use evidence provided by a speaker recognition system to update their prior belief for identity inference. Since use of a speaker recognition system should not result in an information gain, privacy preservation should leave the adversary in a position where they are left making decisions based only upon their prior belief (whatever it is). This is *perfect secrecy* [9], a concept re-coined here as *perfect privacy*.

This section sets out our ideas to make use of the ECE as a means of assessing the level of privacy provided by a privacy preserving solution. We propose two metrics that can be used for optimisation, assessment and ranking according to a so-called *zero evidence biometric recognition assessment* approach.[3] The two metrics aim to measure the extent to which a safeguard preserves privacy, or rather what degree of speaker discriminative information remains in an utterance. The first metric reflects the gain in information that an adversary can obtain by using a speaker recognition system. This is equivalent to the expected privacy disclosure regardless of the adversary's prior belief. Realising that the ECE reflects no more than an expected value (*population* level), the second metric reflects the worst-case scenario (*individual* level), i.e. the maximum level of privacy that may be disclosed despite privacy preservation.

4.1. Expected privacy disclosure

The expected privacy disclosure is the relative information that can be gained from use of a speaker recognition system. This difference is illustrated in terms of the ECE profiles in Fig. 2 where the black line represents the *perfect privacy* ECE (*zero evidence* scores $\mathbf{0}_{\mathcal{S}}$; all LRs have the value 1) and the blue, dashed line represents the *adversary* ECE, i.e. for a system that would yield oracle calibrated scores $\mathcal{S}$ (oracle LRs).[4]

By removing as much as possible any biometric information in speech data, the gap between these two profiles would reduce and result in an increase in the adversary ECE, as indicated by the blue arrows in Fig. 2. In the case that the safeguard is successful in removing *all* the speaker specific information, then the perfect privacy and adversary ECEs would be identical, i.e. we have *perfect secrecy*; there remains *zero evidence*.[5]

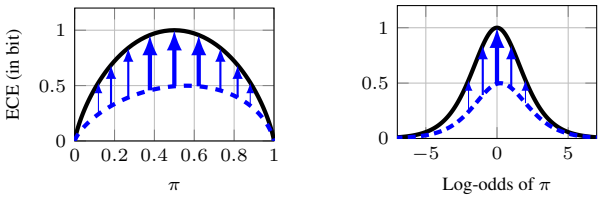
In practice, a safeguard is unlikely to remove *all* evidence; it will likely still result in the disclosure of some privacy and different solutions will disclose different levels of privacy. Hence, we need some means to compare solutions. The answer is a metric which measures the difference between the *perfect privacy* and *adversary* ECE. Since the adversary's prior is unknown, the evaluator must assume maximum uncertainty (different π values have the same probability of occurrence). Accordingly, the metric must reflect the difference between both ECE profiles in Fig. 2 for the full range of priors π. This gives the *expected privacy disclosure* $D_{\text{ECE}}(\Theta \,|\, \mathcal{S})$:

$$D_{\text{ECE}}(\Theta \,|\, \mathcal{S}) = \int_0^1 \text{ECE}(\Theta \,|\, \mathbf{0}_{\mathcal{S}}) - \text{ECE}(\Theta \,|\, \mathcal{S})\, \mathrm{d}\pi. \qquad (5)$$

[3] Code: `https://gitlab.eurecom.fr/nautsch/zebra`

[4] In [14, 18, 19], the *default* and the *minimum* ECE, respectively.

[5] $\text{ECE}(\Theta \,|\, \mathbf{0}_{\mathcal{S}}) = \pi \log_2(1 + \frac{1-\pi}{\pi}) + (1-\pi)\log_2(1 + \frac{\pi}{1-\pi})$.



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

Figure 2: *(a) Adversary ECE (blue profile) for some oracle calibrated scores. The idea is to make the blue equal the black profile (perfect privacy ECE). (b) Conventionally, ECE profiles are plotted against* $\log \frac{\pi}{1-\pi}$*, the log-odds of* π*.*

It can be shown that the integral is given by:

$$D_{\text{ECE}}(\Theta \,|\, \mathcal{S}) = \frac{\langle Z(a)\rangle_{a\in\mathcal{S}_{\mathcal{A}}} + \langle Z(\frac{1}{b})\rangle_{b\in\mathcal{S}_{\mathcal{B}}}}{\log(2)}$$
$$\text{with} \quad Z(x) = \frac{(x-3)(x-1) + 2\log(x)}{4\,(x-1)^2}, \qquad (6)$$
$$x > 0, \qquad \lim_{x\mapsto 1} Z(x) = 0, \quad \text{and} \quad \lim_{x\mapsto\infty} Z(x) = \frac{1}{4},$$

where $\langle\cdot\rangle_{s\in\mathcal{S}_\theta}$ computes the average over scores s in a set $\mathcal{S}_\theta$.

$D_{\text{ECE}}(\Theta \,|\, \mathcal{S})$ signifies the expected disclosure to an adversary in bits and is independent of the adversary's prior belief (the prior is marginalised and integrated out in Eq. 5). It has an intuitive interpretation: if $D_{\text{ECE}}(\Theta \,|\, \mathcal{S}) = \frac{1}{2\,\log(2)} \approx 0.721$, then classes are perfectly separated for all priors (*no privacy*); if $D_{\text{ECE}}(\Theta \,|\, \mathcal{S}) = 0$, then we have *perfect privacy* (*zero evidence*).

4.2. Worst-case privacy disclosure

The worst-case privacy disclosure is reflected by the highest strength of evidence observed during testing. While we can use posterior probabilities of a non-informative prior π as in [21], we propose the use of an additional approach which improves human interpretability, especially for the non-expert. This entails *categorical tags*, see Table 1, for LR values l that are adapted from methodological guidelines for forensic practitioners published by the European Network of Forensic Science Institutes [22] and by forensics work in [23]. Additionally, we provide the odds ratio for recognising a biometric identity by the lowest LR value of a category (assuming no knowledge of the adversary prior: π is flat); lower odds mean less precision for an adversary, thus privacy is more preserved (best is $50:50$).

To determine l, the LR value of the worst-case privacy disclosure, we must consider both positive strength of evidence

Table 1: *Categorical tags of worst-case privacy disclosure.*

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

(proposition $\mathcal{A}$, it *is* the speaker) and negative strength of evidence (proposition $\mathcal{B}$, it is *not* the speaker). The former corresponds to LR values $\mathcal{A}: 1 < l < \infty$, whereas the latter corresponds to LR values $\mathcal{B}: 0 < l < 1$. In LR space, however, the extent to which one proposition is favoured over another is non-linear; LRs of $l = 0.9$ or $l = 1.1$ do not reflect the same strength of evidence in support of each proposition. A linear space is obtained easily by operating in the log-LR (or LLR) space $\mathcal{S}'_{\log}$ of ideally calibrated[6] scores $\mathcal{S}'$:

$$\mathcal{S}'_{\log} = \left(\log(s) \,|\, s \in \mathcal{S}'\right). \tag{7}$$

In log-LR space, a value of *zero* implies neither $\mathcal{A}$ nor $\mathcal{B}$ is favoured (zero evidence; full privacy). In addition, scores in log-LR space are symmetric; scores of -0.1 and +0.1 reflect the same strength of evidence for each proposition. As a result, the worst-case privacy disclosure is obtained as $\log(l)$ by taking the maximum of the absolute value:[7]

$$\log(l) = \max_{s \in \mathcal{S}'_{\log}} (\text{abs}(s)). \tag{8}$$

As for log-LRs, the log-odds of posteriors/priors are also symmetric; Table 1 defines categories by magnitudes—for base 10 log-LRs as $\log_{10}(l)$ values and posterior log-odds.

4.3. Assessing privacy disclosure, an example

Reported here is an example case study performed using the ZEBRA framework in the context of the *VoicePrivacy 2020* challenge [7, 8], which involves the design of anonymisation solutions in privacy preservation. Experiments were performed with the female subset of the LibriSpeech test data using: the two challenge baselines, B1 [24] (a pre-trained x-vector approach) and B2 [25] (a formant shifting technique; no training)—the *safeguard* component in Fig. 1; a state-of-the-art x-vector [26] speaker recognition system—the *classifier* component in Fig. 1. We report results for all subsets online.[3]

Results are illustrated in Fig. 3. The legend also shows ZEBRA results in the form of a (D_{ECE}, $\log_{10}(l)$, tag) tuple, where computations are according to (6) for D_{ECE} and to (8) for $\log_{10}(l)$, with categorical tags referenced in Tab. 1. The lower blue profile shows the ECE without protection (D_{ECE}: 0.58, $\log_{10}(l)$: 3.98, tag: C). The magenta and green profiles show the ECE of each baseline. The black profile corresponds to perfect privacy or *zero evidence*.

The baselines offer varying degrees of privacy preservation. B1 (D_{ECE}: 0.11) appears to perform considerably better than B2 (D_{ECE}: 0.36); the green and black profiles are relatively close together whereas the gap between black and magenta profiles is substantial. $\log_{10}(l)$ results show a somewhat different picture: 3.98 for the unprotected system, 3.58 for B2 and 2.27 for B1.

[6] We refer to $\mathcal{S}'$ instead of $\mathcal{S}$ on purpose. To avoid infinite LR values, see the code of [15], we extend the calibration used in Section 4.1, and apply *Laplace's rule of succession* (also known as *the sunrise problem*). Two dummy scores are added to the extremities of all observed scores—one for class $\mathcal{B}$ and one for class $\mathcal{A}$. The former serves as a Bayesian predictor to the LR value for the highest class $\mathcal{A}$ scores (that are larger than the highest class $\mathcal{B}$ score), and the latter captures the infinity.

[7] The metric $\log(l)$ corresponds to the L^∞ length-norm of a vector $\boldsymbol{x}$ with n LLRs: $||\boldsymbol{x}||_\infty = \max\{|x_1|, \ldots, |x_n|\}$. If this length is zero, there is *zero evidence* disclosure (the vector points from its origin to itself). A privacy disclosure, a non-zero dimension, increases the vector length. The worst-case metric is *optimistic*, since any other L^p norm with $p < \infty$ is larger (despite the Bayesian prediction of LLRs).

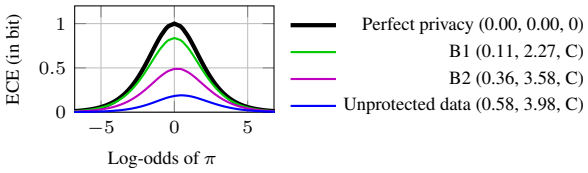


Figure 3: *ZEBRA assessment with ECE profiles.*

These results *all* correspond to tag C in Tab. 1.[8] Hence, while B1 is substantially better than B2 in terms of *average* privacy disclosure, there is less to choose between them in terms of the categorical *worst-case* privacy disclosure.

These observations highlight the value of the proposed ZEBRA framework for privacy preservation. Whereas the EER and other metrics may give an unduly optimistic and distorted sense of protection, the D_{ECE} provides a reliable estimate of protection in terms of the *average* protection afforded to a *population* whereas $\log(l)$ provides additional insights into the level of protection afforded in a *worst-case* to an *individual*.

5. Discussion and future work

With mounting legislation demanding protections for personal information, provisions for privacy are today of paramount consideration. This paper presents the ZEBRA framework for the evaluation of privacy safeguards. It overcomes many of the weaknesses of existing metrics such as the equal error rate. Crucially, the framework formulates the problem in the realm of a privacy adversary and provides the means to assess the average and worst-case privacy disclosure of a given safeguard.

While the work makes inroads towards an adversary-centric metric, the arguments presented in this paper suggest that we need to go one step further; metrics are only one part of an assessment strategy. The case study presented in this paper uses a protocol designed by an *evaluator*. Just like the decision policy, the protocol is also in the realm of the *adversary*. Future work should hence extend the current study to consider the privacy that can be disclosed when the adversary chooses both the decision policy *and* the protocol. Similarly, the current work assumes oracle score calibration, whereas it too is in the realm of the adversary. Future work could hence study the impact of calibration within the ZEBRA framework, such that they cannot be used by adversaries.

Finally, the ZEBRA framework proposed in this paper is relevant to the *biometric information protection* standard, currently in revision. It is also readily extendable to the study of privacy concerning sensitive speech data, e.g. health and emotional status, or particularly sensitive spoken/transcribed content, e.g. political and religious beliefs. Finally, speech serves as only one example application of the ZEBRA framework; since it operates in the score domain, it can be applied with minimal effort to the study of non-speech problems such as privacy preservation in video surveillance.

6. Acknowledgements

This work is partly funded by the projects: ANR-JST VoicePersonae, ANR Harpocrates and ANR-DFG RESPECT.

[8] Categorical tags for LRs and their tables evolved over time since 1961 starting with Jeffrey, whose table ended with $2 \leq \log_{10}(l)$ [27]; category C resulted after the introduction of DNA evidence in media.

7. References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Proc. Interspeech*, 2019, pp. 3695–3699.

[2] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, "Privacy implications of voice and speech analysis – information disclosure by inference," in *Proc. Privacy and Identity Management. Data for Better Living: AI and Privacy: IFIP Int'l Summer School 2019, Revised Selected Papers*. Springer, 2020, pp. 242–258.

[3] California State Legislature, "Assembly bill no. 375, chau. privacy: personal information: businesses (California Consumer Privacy Act)," 6 2018.

[4] European Council, "Regulation 2016/679 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," 4 2016.

[5] ——, "Directive 2016/680 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA," 4 2016.

[6] Council of Europe, "European convention on human rights," 6 2010.

[7] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. M. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 challenge evaluation plan," 2020.

[8] ——, "Introducing the VoicePrivacy initiative," in *Proc. Interspeech*, 2020.

[9] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 10 1949.

[10] J. Katz and Y. Lindell, *Introduction to modern cryptography*. Chapman and Hall/CRC, 2014.

[11] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Trans. on Information Forensics and Security (TIFS)*, vol. 3, no. 6, pp. 1406–1420, Jun. 2018.

[12] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Elsevier Science Speech Communication*, vol. 31, pp. 225–254, 6 2000.

[13] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Elsevier Computer Speech and Language (CSL)*, vol. 20, no. 2, pp. 230–275, 7 2006.

[14] D. Ramos and J. Gonzalez-Rodrigues, "Cross-entropy analysis of the information in forensic speaker recognition," in *Proc. IEEE Odyssey*, 2008.

[15] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," [Online] https://sites.google.com/site/bosaristoolkit, accessed 2020-01-15, AGNITIO Research, South Africa, Tech. Rep., 12 2011.

[16] N. Brümmer and J. du Preez, "The PAV algorithm optimizes binary proper scoring rules," 2009.

[17] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, University of Stellenbosch, 2010.

[18] D. Ramos-Castro, "Forensic evaluation of the evidence using automatic speaker recognition systems," Ph.D. dissertation, Universidad Poliécnica de Madrid, 2007.

[19] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," *Entropy*, vol. 20, no. 3, p. 208, 3 2018.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[21] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delcrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language, Special issue on Speaker and language characterization and recognition: voice modeling, conversion, synthesis and ethical aspects*, vol. 58, pp. 441–480, 11 2019.

[22] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, "Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition including guidance on the conduct of proficiency testing and collaborative exercises," Europen Network of Forensic Science Institutes, Tech. Rep., 2015.

[23] A. Nordgaard, R. Ansell, W. Drotz, and L. Jaeger, "Scale of conclusions for the value of evidence," *Law, Probability and Risk*, vol. 11, no. 1, pp. 1–24, 2012.

[24] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using X-vector and neural waveform models," in *Proc. Speech Synthesis Workshop (SSW)*, 2019, pp. 155–160.

[25] J. Patino, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," Eurecom, Tech. Rep. Research Report RR-20-343, 02 2020. [Online]. Available: http://www.eurecom.fr/publication/6190

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[27] D. H. Kaye, "The weight of evidence in law, statistics, and forensic science," in *Proc. NIST Technical Colloquium on Quantifying the Weight of Forensic Evidence*, 2016, [Online] https://www.nist.gov/system/files/documents/2020/01/22/03_kaye_16-nist-woe-linear.pdf, 2020-05-06.