Towards a unified assessment framework of speech pseudonymisation

Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino,
Jean-François Bonastre, Natalia Tomashenko, Driss Matrouf

Version of Record: https://www.sciencedirect.com/science/article/pii/S0885230821001005
Manuscript_fc5cb523c611faf8874813a4b9a9a39d

Towards a Unified Assessment Framework of Speech Pseudonymisation

Paul-Gauthier Noé[a,*], Andreas Nautsch[b], Nicholas Evans[b], Jose Patino[b], Jean-François Bonastre[a], Natalia Tomashenko[a], Driss Matrouf[a]

[a] *Laboratoire Informatique d'Avignon (LIA), Avignon Université, France*
[b] *Digital Security Department, EURECOM, France*

Abstract

Anonymisation and *pseudonymisation* are two similar concepts used in privacy preservation for speech data. With no established definitions for these tasks, nor standard approaches to assessment, this paper provides definitions and presents two complementary assessment frameworks. The first is based on voice similarity matrices which provide both an immediate visualisation of privacy protection performance at the speaker level and two objective measures in the form of de-identification and voice distinctiveness preservation. The approach readily highlights imbalances in system performance at the speaker level. The second, referred to as the zero evidence biometric recognition assessment (ZEBRA) framework, is based on information theory and measures the amount of private information disclosed in speech data. The paper presents also an extension to the original ZEBRA framework. It aims to reflect the robustness of the privacy safeguard when a privacy adversary adapts to the protected speech. We demonstrate the application of both frameworks to assess pseudonymisation performance on the two VoicePrivacy 2020 challenge baseline solutions plus a third one. The two frameworks were designed independently of each other. The ZEBRA framework is fully consistent with the Bayesian decision theory and the other framework focuses instead on speaker-wise visualisations of a system performance. Thus, while metrics derived from them bear similarities, they expose differences in safeguard behaviour. The assessment of pseudonymisation remains challenging and merits greater attention in the future.

Keywords: pseudonymisation, anonymisation, privacy preserving speech transformation, voice conversion, assessment methods, speaker verification.

*Corresponding author
Email address: paul-gauthier.noe@univ-avignon.fr (Paul-Gauthier Noé)

1. Introduction

The implementation of the European General Data Protection Regulation (GDPR) (European Council, 2016) and the proliferation of speech technology has given rise to the need for privacy preservation solutions (Nautsch et al., 2019a). It is well known that speech data contains a high degree of personal information such as the speaker identity, age, gender, socio-economic status, ethnicity, geographical and educational background (COMPRISE, 2019; Nautsch et al., 2019b; Schuller & Batliner, 2019). Without privacy safeguards, anyone with access to speech data is able to infer personal information that could be exploited for malicious purposes. Privacy preservation solutions aim to prevent access to such personal information contained within speech data.

Various different protections can be applied to speech data. Classical encryption techniques can be used in order to prevent man-in-the-middle attacks, exploits used to intercept data during transmission. It does not prevent unauthorised or unexpected exploitation of speech data by the receivers after decryption. The solution is to prevent access to personal information, allowing access only to information that is strictly necessary for the fulfillment of some desired service, e.g. Automatic Speech Recognition (ASR). More recent approaches support computation upon encrypted data. *Homomorphic encryption* is one example and has been applied successfully to speech recognition in the encrypted domain (Pathak et al., 2011; Zhang et al., 2019). However, the computational complexity of such techniques is prohibitive. An alternative to encryption is some form of speech transformation which refers here to any process that aims to conceal the speaker identity in an utterance where the output is still a raw speech signal. Such solutions are promoted by the VoicePrivacy initiative[1] (Tomashenko et al., 2020b) which takes the form of an internationally competitive challenge in anonymisation.

Key to driving progress, are benchmarking frameworks that allow for the meaningful comparison of competing solutions. While the VoicePrivacy initiative has established common databases and protocols, the community is yet to agree upon de facto evaluation frameworks. The strength of an anonymisation solution has traditionally been assessed using an Automatic Speaker Verification (ASV) system which produces Log-Likelihood-Ratio (LLR) scores. Different metrics can be computed from these scores: the Equal Error Rate (EER) where the decision threshold is set such that the false acceptance and false rejection rates are equal; the Log-Likelihood-Ratio Cost (C_{llr}), which is application-independent and can be decomposed into a discrimination and a calibration cost (Brümmer & du Preez, 2006). These metrics, computed using *protected* probe and *unprotected* enrolment speech segments, give insights into the ability of a protection to suppress speaker identity information. While they can be intuitive and are well known within the speech community, these metrics were not designed with the aim of assessing privacy preserving speech transformation.

[1] `https://voiceprivacychallenge.org`

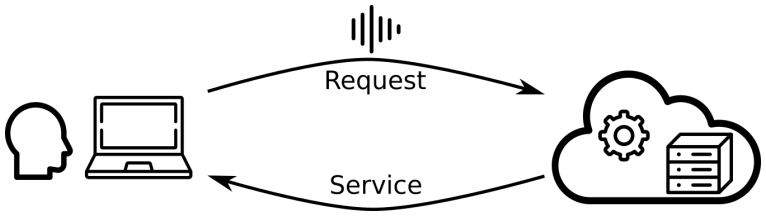
Two approaches specific to the assessment of such privacy safeguards have been proposed recently alongside the VoicePrivacy challenge. The first approach is based on *voice similarity matrices* (Noé et al., 2020). It provides intuitive visualisations and gives two metrics by comparing the diagonals of the matrices: one to measure the level of protection and one to check that the resulting protected voices remain distinguishable from each other. Visualisations of the voice similarity matrices expose any imbalance in the protection afforded to different speakers, differences that are missed by more traditional approaches that reflect protection strength in an average or pooled sense. The second approach, named the Zero Evidence Biometric Recognition Assessment (ZEBRA) framework (Nautsch et al., 2020), measures the expected and worst-case privacy disclosure of protected speech segments, i.e. a measure of the remaining information that could be used by an adversary to infer the speaker identity.

In addition to providing missing definitions for key concepts in speech transformation for privacy preservation and gathering the assessment frameworks originally presented in (Noé et al., 2020; Nautsch et al., 2020), this paper presents: an update for the computation of the voice similarity and its analogy with the C_{llr} and the Zoo plot (Dunstone & Yager., 2009); the ZEBRA framework when scores are Log-Likelihood-Ratio (LLR) instead of Likelihood-Ratio (LR) as originally presented in (Nautsch et al., 2020), as well as an extension to assess the attacker's calibration ability; and the comparison of both frameworks on the VoicePrivacy 2020 challenge's evaluation protocol. Section 2 is devoted to the definition of the main concepts, about which there is some ambiguity in the literature, e.g. *anonymisation* and *pseudonymisation*. Following prerequisites presented in Section 3, Section 4 describes and updates the voice similarity matrices based assessment framework. Section 5 describes and extends ZEBRA. The use of these frameworks is demonstrated on the VoicePrivacy 2020 challenge's baselines in Section 6. Finally, Section 7 discusses the consistency in results and findings of the two frameworks.

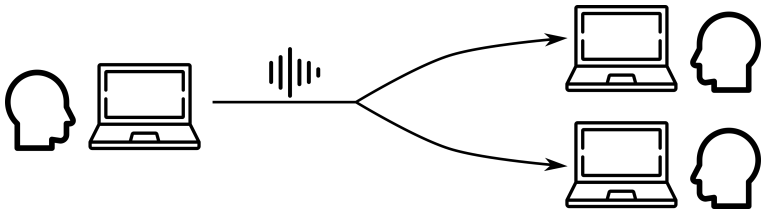
2. Anonymisation and Pseudonymisation

Speech technologies include many different applications, from smart-speakers to interactive voice response systems, touching domains such as translation, medical, educational or authentication services. Users will typically release speech data to remote service providers. This data may be stored and processed by the service provider which the user trusts not to abuse it. To better protect themselves from abuse and so as to ensure some level of privacy, users may wish to prevent their identity being revealed from the speech data entrusted to the service provider. Teleconferencing applications are also commonly used and in such a case, an user may want to hide its identity to its interlocutors. These two kinds of speech technologies are illustrated in Figure 1. Depending on the application, a user may require different kinds of privacy preservation.

Anonymisation and *pseudonymisation* are both commonly used terms, yet both are also commonly misunderstood. Both anonymisation and pseudonymisation aim to transform a speech segment in order to conceal the speaker iden-



(a) An user sending a speech based request to a service provider.



(b) An user exchanging with some people using a speech communication system.

Figure 1: Diagrams of two different kinds of speech technology applications. In 1a, the user interacts with a service provider (client-server) while in 1b, the user interacts with other users. In both cases, she or he may want to hide their identity.

tity (Tomashenko et al., 2020b; Noé et al., 2020). In both cases, other speech characteristics such as the linguistic content and the speech intelligibility should be preserved. Even if the linguistic content can be used to infer the identity of the speaker, this is beyond the scope of the work reported in this paper, where we are concerned with preventing the leakage of identity information via paralinguistic and acoustic cues that are most commonly used by automatic approaches to speaker recognition. The following definitions aim to highlight differences between anonymisation and pseudonymisation.

2.1. De-Identification

De-identification (Jin et al., 2009; Justin et al., 2015; Bahmaninezhad et al., 2018; Noé et al., 2020) also refers to the concealment of a speaker identity so that it is not possible to recover it from protected utterances. This term is also referred to as voice-disguise (Zhang, 2012; Hautamäki et al., 2018) or identity masking (Pobar & Ipšić, 2014). It is likely, however, that the full concealment of identity will be almost impossible in practice. One reason is that identity information can not be explicitly disentangled from other speech attributes that must be preserved. Accordingly, de-identification can be viewed as a process that increases the uncertainty in the linkability between an utterance and the corresponding speaker identity. A perfect de-identification would result in *perfect privacy*. The latter was introduced by Claude Shannon in (Shannon, 1949) as *perfect secrecy* and is defined as the situation where an adversary can not update their *prior knowledge* by using intercepted data (used as an evidence in a Bayesian decision framework). In other words, the adversary's *posterior knowledge* remains the same as its prior knowledge. This concept of perfect

privacy is fundamental to the development of the ZEBRA framework presented in Section 5.

2.2. Voice Distinctiveness Preservation

Depending upon the application, it can be desirable that protected voices remain consistent and distinguishable, i.e. in the protected domain, segments uttered by the same speaker are mutually linkable but distinct from protected segments uttered by another speaker. According to Noé et al. (2020), this requirement is referred to as *voice distinctiveness preservation*.

Figure 2 illustrates a scenario in which the voice distinctiveness preservation requirement is paramount. Consider three speakers who communicate using a teleconferencing system while wishing not to disclose their identities. De-identification systems can be used to conceal identity but might produce three almost identical, indistinguishable voices, as per the upper-right situation in Figure 2. Such an exchange between users with the same protected voice will result in a confusing and unnatural conversation. Voice distinctiveness preservation allows for a comfortable, natural conversation by ensuring that the three protected voices remain distinguishable as per the bottom-right situation in Figure 2. Speaker diarization, namely the who-spoke-when task, could be applied likewise to both unprotected and protected conversations. In other words, the voice mapping (original speaker's voice to its protected version) must be injective i.e. two distinct original voices must result in two distinct protected voices.

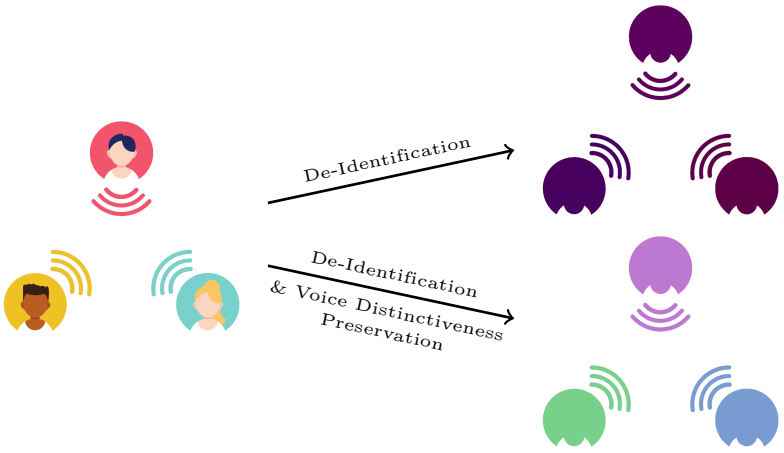


Figure 2: Application of the *de-identification* alone and together with the *voice distinctiveness preservation*. Using both results in the concealment of the speakers identity while avoiding confusion between the protected voices.

2.3. Anonymisation

Anonymisation aims to make re-identification from protected speech impossible, by any means. Thus, it ensures perfect de-identification and breaks any relation from the protected voices to their original version. Any process applied to the speech data should hence be irreversible.

Meeting anonymisation is difficult in practice, if not impossible. This is because many different attributes of the speech signal can be used to derive the speaker identity and it is generally not possible to remove all of them without also suppressing other attributes which must be preserved. An example is the linguistic content which must be preserved, but which also serves as a potential cue from which to infer the identity. The information to be preserved can not be explicitly disentangled from that relating to identity. Anonymisation may not be achievable in practice and hence the terminology is reserved to reflect a desired goal (Tomashenko et al., 2020a).

2.4. Pseudonymisation

The definition of pseudonymisation in (European Council, 2016) is softer than anonymisation, in the sense that additional information is required to recover the identity. Within the scope of privacy preservation for speech technology, we define pseudonymisation as the concealment of the speaker identity with voice distinctiveness preservation. Pseudonymisation results in distinct protected voices for each speaker. These protected voices will now be called *pseudo-voices*: like a pseudonym, the user will claim their pseudo-voice only to the people by whom she or he accepts to be identified by.

As an injective voice mapping is required in order to ensure the distinctiveness of the pseudo-voices, the existence of an inverse mapping[2] jeopardise the irreversibility requirement of anonymisation. Therefore, the anonymisation as an objective is incompatible with the pseudonymisation.

3. Prerequisites

This section reminds some prerequisites that will be necessary to tackle the presented assessment methods with better rigour and intuition. In particular, the privacy preservation task and the attack model; the strength-of-evidence score; its role in identity inference in the context of the Bayesian decision framework; the representation of posterior beliefs in binary decision making and the computation of the C_{llr}. Readers familiar with these concepts may skip this section.

[2]By reducing the codomain of the voice mapping to the image domain, the mapping is bijective.

3.1. The Privacy Preservation Task and the Attack model

The privacy preservation task is defined here as the transformation of a set O of (original) speech segments into a set P of corresponding protected segments from which an *adversary* should not be able to recover the original identity of the speakers. In order to recover the speaker who has uttered a given or intercepted protected speech segment, the adversary has to decide whether it has been uttered by the same speaker than a reference segment (*target* proposition) or by a different speaker (*impostor* proposition). As in standard Bayesian decision frameworks, the adversary has to update its *prior belief* by the strength-of-evidence produced by a biometric classifier which compares the speech segments. Human perception of the speech segments comparison is out of the scope of this paper, therefore, the adversary perceives data only through an ASV system.

In this paper, we only consider the scenario where the adversary has previously obtained unprotected speech segments (for which the identity of the speakers is known), and receives or intercepts protected speech segments whose identity she or he wishes to recover. Therefore, the unprotected reference segments are used as *enrolment* segments and the received or intercepted segments are *probes*. This corresponds to unprotected-protected comparisons (OP setting) and is referred as the *ignorant* attack model in (Tomashenko et al., 2021). This is a restricted attack scenario and many others, that we leave for future works, can be contemplated.

3.2. The Strength-of-Evidence Score, the Likelihood-Ratio

The LR is commonly used as a *score* to reflect the strength-of-evidence given by biometric classifier especially in forensic science and in particular in ASV (Drygajlo et al., 2003). It is defined as:

$$\mathrm{lr} = \frac{P(E|\theta_{\mathrm{tar}})}{P(E|\theta_{\mathrm{imp}})}, \tag{1}$$

where E is the evidence corresponding to the comparison of two speech segments for deciding between two propositions, namely, θ_{tar}: *target* (the segments have been uttered by the same unique speaker), and θ_{imp}: *impostor* (the segments have been uttered by different speakers) propositions.

As the name and Equation 1 suggest, a LR is a ratio of likelihoods. However it may not be well calibrated in the sense that it might not have a suitable probabilistic meaning for making good Bayes decision. A crucial property of perfectly calibrated likelihood-ratio is the *idempotence*. Perfectly calibrated LR let to posteriors that represent the class distribution in the evidence/feature space $P(\theta_{\mathrm{tar}}|E)$ equivalently to the class distribution in the score

space $P(\theta_{\text{tar}}|\text{lr})$ (Mandasari, 2017):

$$
\begin{aligned}
P(\theta_{\text{tar}}|E) &= P(\theta_{\text{tar}}|\text{lr}),\\
\frac{P(\theta_{\text{tar}}|E)}{1-P(\theta_{\text{tar}}|E)} = \frac{P(\theta_{\text{tar}}|E)}{P(\theta_{\text{imp}}|E)} &= \frac{P(\theta_{\text{tar}}|\text{lr})}{P(\theta_{\text{imp}}|\text{lr})} = \frac{P(\theta_{\text{tar}}|\text{lr})}{1-P(\theta_{\text{tar}}|\text{lr})},\\
\text{Bayes' rule gives}\quad \frac{P(E|\theta_{\text{tar}})}{P(E|\theta_{\text{imp}})}\frac{P(\theta_{\text{tar}})}{P(\theta_{\text{imp}})} &= \frac{P(\text{lr}|\theta_{\text{tar}})}{P(\text{lr}|\theta_{\text{imp}})}\frac{P(\theta_{\text{tar}})}{P(\theta_{\text{imp}})},\\
\text{lr} &= \frac{P(\text{lr}|\theta_{\text{tar}})}{P(\text{lr}|\theta_{\text{imp}})}.
\end{aligned}
\tag{2}
$$

On the left side, we have the LR; on the right side, we have the LR of the LR: *the LR of the LR is the LR* (Mandasari, 2017); this is the idempotence. This is a necessary property for LRs to be perfectly calibrated. Then, a non-linear monotonic function can be applied for calibration to obtain these so-called *oracle scores* (Brümmer & De Villiers, 2011; Brümmer & Preez, 2013); a simulation for the underlying ideal value of likelihood-ratios in a specific experiment. Because of this property, one can interchangeably use the evidence E (i.e. the comparison of two speech segments) and scores in equations when they are perfectly calibrated LRs.

3.3. Identity Inference and Bayesian Decision Framework

In voice biometric, to infer whether or not to reject an identity claim for lacking evidence support, the Bayesian decision framework is used. First, before creating any classification system, a risk-based decision policy is defined to quantify beliefs in (i) cost associated to correct and erroneous decision outcomes and (ii) the expected proportion of true and false identity claims to occur (the prior probabilities of θ_{tar} and θ_{imp}). Then, evidence is observed, and the prior belief is updated by the strength-of-evidence, favoring one proposition over another one; through using Bayes' theorem, the posterior belief results and informs on the likelihood of the identity claim to be true or false. Finally, an action is taken if this identity inference is convincing, i.e., whether or not the posterior belief meets the cost demand (both, posteriors and costs, are compared as trade-offs regarding θ_{tar} and θ_{imp}: the more striking trade-off wins).

However, when preserving privacy, the particular costs resulting from privacy infringements remain inaccessible to the one who creates or uses privacy safeguards. For assessing privacy preservation in the light of inference from evidence, we need to go one step back: the strength-of-evidence itself needs to be addressed and its informative capacity to decision making through inference. To make this concept more tangible, we discuss the suitability of different ways to represent posterior beliefs (probabilities, logarithms, odds, and odds ratios).

3.4. On the Representation of Posterior Beliefs in Binary Decision Making

The following discussion is solely reflective of making posterior beliefs appealing to the human eye. While the nature of probabilities is non-linear, the human mind needs linear representations to make sense out of observations. By

representing posterior beliefs in different ways, we want to briefly illustrate a different mindset that might prove useful for the topic at hand.

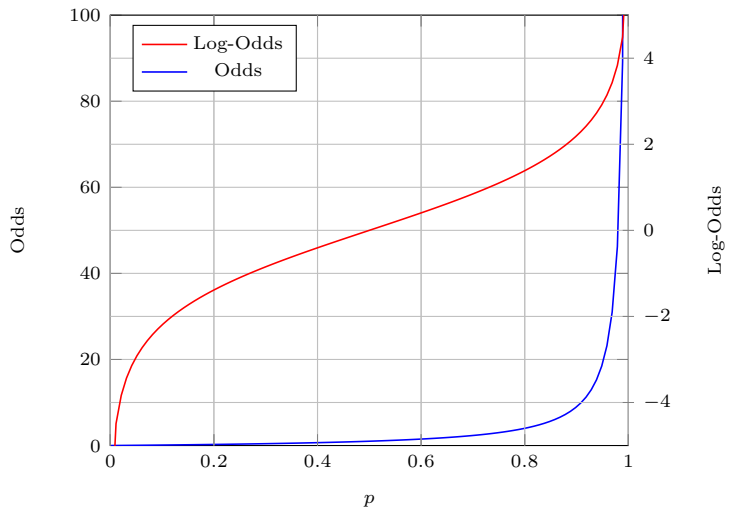


Figure 3: Odds and Log-Odds as a function of the probability p. The Log-Odds are symmetric around $p = 0.5$. Indeed, $\text{logit}(p + 0.5) = -\text{logit}(-p + 0.5)$.

Let us consider the following nine probabilities: 1%, 10%, 20%, 40%, 50%, 60%, 80%, 90%, 99%. While larger/lesser (additive) comparisons are trivial, relative (multiplicative) comparisons are less intuitive. Logarithmic representations transform multiplicative problems into additive ones. However, there are two possible decision outcomes in our model: θ_{tar} and θ_{imp} are exhaustive and mutually exclusive. Then, one probability is expressed by the other: $P(\theta_{\text{tar}}) = 1 - P(\theta_{\text{imp}})$.

Thus, the relationship between these two are of interest, rather than the value of one of them by itself, we can express it by odds: A of N trials are successful, B of N are not ($A + B = N$); the odds are $A : B$. The above nine probabilities represented as odds are: $0.010101 : 1$ $(1 : 99)$, $0.111111 : 1$ $(1 : 9)$, $0.25 : 1$ $(1 : 4)$, $0.666667 : 1$ $(1 : 1.5)$, $1 : 1$, $1.5 : 1$, $4 : 1$, $9 : 1$, $99 : 1$. For probabilities above 50%, this representation is intuitive; for probabilities below 50%, we need to invert the odds for convenience (which still blocks intuitive rigour). By keeping the $X : 1$ notation, however, we are closer to a unified scale: the X values are the odds-ratios i.e. $\frac{A}{B}$ in analogy to $A + B = N$.

The multiplicative inconvenience can be remedy by logarithms, namely, the logarithmic odds-ratios (log-odds). The log-odds representation of the above probabilities are: -4.59512, -2.19722, -1.38629, -0.405465, 0, 0.405465, 1.38629, 2.19722, 4.59512. It results in a symmetric probability representation with 50% at 0 in the context of making binary decisions (see Figure 3). Log-odds are directly computed from probabilities by the logit function: $\text{logit}(p) = \log \frac{p}{1-p}$. Its inverse function is the sigmoid function $\sigma(x) = (1 + \exp(-x))^{-1}$.

Of course, the same holds true for the posteriors: $P(\theta_{\text{tar}}|E) = 1 - P(\theta_{\text{imp}}|E)$. Posteriors $P(\theta_{\text{tar}}|E)$ are inferred from priors $P(\theta_{\text{tar}})$ and LRs as:

$$\begin{aligned} P(\theta_{\text{tar}}|E) &= \sigma\left(\log \frac{P(E|\theta_{\text{tar}})}{P(E|\theta_{\text{imp}})} + \text{logit}(P(\theta_{\text{tar}}))\right), \\ \text{logit}(P(\theta_{\text{tar}}|E)) &= \log \frac{P(E|\theta_{\text{tar}})}{P(E|\theta_{\text{imp}})} + \text{logit}(P(\theta_{\text{tar}})), \\ \text{logit}(P(\theta_{\text{tar}}|E)) - \text{logit}(P(\theta_{\text{tar}})) &= \log \frac{P(E|\theta_{\text{tar}})}{P(E|\theta_{\text{imp}})}. \end{aligned} \tag{3}$$

The term $\log \frac{P(E|\theta_{\text{tar}})}{P(E|\theta_{\text{imp}})}$ is the log-LR (LLR). In the log-odds or LLR space, binary decision making is an additive comparison problem which allows for intuitive rigour.

3.5. Goodness of LLRs

The C_{llr} (Brümmer & du Preez, 2006) quantifies the capability of a set of scores to resemble calibrated LLRs. It is derivable in three different ways: by integrating priors and costs out of detection cost functions, by measuring Empirical Cross Entropy (ECE) at the generalised uninformed prior $P(\theta_{\text{tar}}) = 0.5$, and by the logarithmic proper scoring rule. For a set $\mathcal{L} = \mathcal{L}_{\text{tar}} \sqcup \mathcal{L}_{\text{imp}}$ of LLR scores (where $\mathcal{L}_{\text{tar}}$ is the target LLRs subset and $\mathcal{L}_{\text{imp}}$ is the impostor LLRs subset) (Brümmer & De Villiers, 2011)[3]:

$$C_{\text{llr}}(\mathcal{L}) = \frac{\langle -\log \sigma(a)\rangle_{a\in\mathcal{L}_{\text{tar}}} + \langle -\log \sigma(-b)\rangle_{b\in\mathcal{L}_{\text{imp}}}}{2\ \ln(2)}. \tag{4}$$

Through monotonicity-preserving transformations (score calibration), C_{llr} can be improved; oracle score calibration (Brümmer & Preez, 2013) results in its minimum value $C_{\text{llr}}^{\text{min}}$. Then, resulting scores are perfectly calibrated LLRs (satisfying the idempotence property). C_{llr} and $C_{\text{llr}}^{\text{min}}$ are primary metrics of the 2020 VoicePrivacy challenge (Tomashenko et al., 2020b,a) for assessing biometric privacy, alongside the equal-error rate (EER). Both, EER and $C_{\text{llr}}^{\text{min}}$, relate to the convex hull of the receiver operating characteristic (Brümmer & De Villiers, 2011).

If $C_{\text{llr}} > 1$, scores are badly calibrated (better to use a coin toss instead of the particular classifier); if $C_{\text{llr}} = 1$, a classifier performs as good as a coin toss, and if $C_{\text{llr}} < 1$, scores are referred to be calibrated (scores are perfectly-calibrated if $C_{\text{llr}} = C_{\text{llr}}^{\text{min}}$). In brief, to ensure $P(\theta_{\text{tar}}|E) = P(\theta_{\text{tar}}|\text{lr})$ (LLR idempotence), oracle score calibration maps the empirical posterior $P(\theta_{\text{tar}}|E)$ to the accurate

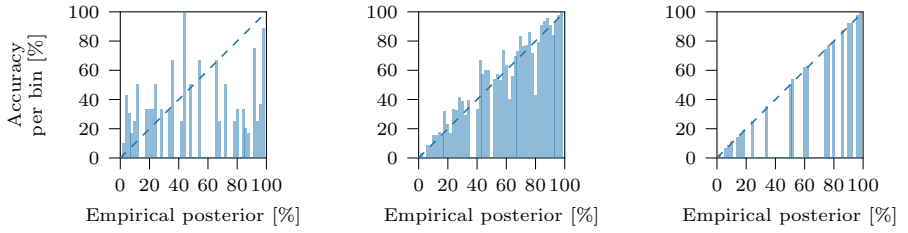
[3] $\sqcup$ refers to the disjoint union. In order lighten the equations, we use the $\langle\cdot\rangle_{\cdot}$ notation to replace summation and division: $\langle a\rangle_{a\in\mathcal{A}} = \sum_{a\in\mathcal{A}} \frac{a}{N}$. This is the average over a set $\mathcal{A}$ of N scalar values where each element occurs equally likely.

(oracle) posterior $P(\theta_{\text{tar}}|\text{lr})$. An (oracle) LLR is derived by removing the empirical database prior $\dot{\pi} = \frac{|\mathcal{L}_{\text{tar}}|}{|\mathcal{L}_{\text{tar}}|+|\mathcal{L}_{\text{non}}|}$ from the oracle posterior (Brümmer & De Villiers, 2011; Brümmer & Preez, 2013):

$$\log(\text{lr}) = \text{logit}(P(\theta_{\text{tar}}|\text{lr})) - \text{logit}(\dot{\pi}). \tag{5}$$

C_{llr} and $C_{\text{llr}}^{\text{min}}$ inform on the usefulness of a classifier to aid in making binary decisions. Most of machine learning literature considers class discrimination performance only; however, score calibration is just as relevant. Otherwise, without additional knowledge, classifier outputs mislead decision making. In preserving privacy, any means to mislead an adversary is targeted; thus, score calibration is another dimension we outline how in the following.

For visualising the calibration quality of posterior probability estimates, see Ramos & Gonzalez-Rodriguez (2013) for details. The histogram of empirical posteriors is compared to a 45° line (y=x): if the line is perfectly resembled, all scores are well-calibrated; otherwise, not. Then, the classifier is over-/underconfident in its forecast. The empirical posterior for a score s is $P(\theta_{\text{tar}}|s) = \sigma(s + \text{logit}(\dot{\pi}))$ in experiments with empirical prior $\dot{\pi}$. First, histograms are computed for each class θ_{tar}, θ_{non} (small bin widths for probabilistic assessment); second, bin by bin, the number of θ_{tar} cases is divided by the total number of cases. If this ratio of cases equals the prediction posterior, the latter is calibrated. Fig. 4 shows empirical calibration plots (unprotected speech data).



(a) Uncalibrated scores (b) Calibrated scores (c) Oracle score calibration

Figure 4: Empirical calibration plots (Ramos & Gonzalez-Rodriguez, 2013) for the kaldi x-vector recipe (Snyder et al., 2018) used on VCTK female test partition (Veaux et al., 2019): the recipe is applied out of domain; score calibration is necessary.

4. Voice Similarity Matrices based Assessment

Alongside the VoicePrivacy Challenge 2020, two assessment frameworks have been proposed in addition to the official ones: the ZEBRA framework (Nautsch et al., 2020) which will be presented in Section 5 and a framework based on voice similarity matrices (Noé et al., 2020) presented in the present section.
This framework proposes to measure the de-identification performance and the voice distinctiveness preservation of a privacy safeguard. Both measures are

based on *voice similarity matrices*. Indeed, given a set of speakers, their one-by-one comparisons can be summarised in a similarity matrix where an element reflects the resemblance between the row's speaker and the column's speaker. A similarity between two speakers is computed from scores given by the biometric comparison of their respective speech segments. Therefore, three different matrices can be computed, one for each of the following cases: the biometric classifier is fed with two original segments (OO), with one original segment and one protected segment (OP) or with two protected segments (PP). Thus, in the OO setting, a nice diagonal must appear in the similarity matrix pointing the speaker discrimination ability in the original set while in the OP setting, the diagonal must disappear if the protection is good. In the PP setting, if the resulting pseudo-voices are distinct and can be discriminated, the diagonal of the matrix should emerge. This idea of comparing the emergence of diagonals between these matrices is the founding principle of the metrics proposed in this framework and is detailed in the following.

4.1. Voice Similarity Matrix

An element of a voice similarity matrix gives a measure of similarity between two speakers. The similarity is computed with LLRs obtained from oracle calibrated (Brümmer & Preez, 2013) scores given by the one-by-one comparisons of the speech segments using a biometric classifier. More precisely, a voice similarity matrix M is defined as $M = (\mathrm{Sim}(i,j))_{1\leq i,j\leq N}$ where N is the number of speakers in the set to protect and the similarity $\mathrm{Sim}(i,j)$ is the geometric mean of the posterior, assuming an uniformed prior $P(\theta_{\mathrm{tar}}) = P(\theta_{\mathrm{imp}}) = 0.5$, and is computed[4] as:

$$\mathrm{Sim}(i,j) = \prod_{l\in\mathcal{L}_{\mathrm{i,j}}} \left(\sigma\left(l\right)^{\frac{1}{|\mathcal{L}_{\mathrm{i,j}}|}}\right), \tag{6}$$

where $\mathcal{L}_{i,j}$ is the set of LLRs from the biometric comparisons of all segments from speaker i with all segments from speaker j. When computing the similarity of a speaker with itself (self-similarity), scores from the comparison of identical segments are removed from the mean in order to avoid overestimated similarities. When computing the voice similarity matrix *within* a set (with the OO or PP setting), it might be surprising to not have all self-similarities equal to a maximum value one. However, a self-similarity has to be seen as the reflection of the ASV's behavior on a speaker rather than a measure compatible with a distance (in the mathematical sense) between a speaker and itself: if a speaker has a small self-similarity, it will suggest that she or he behaves like a goat facing the ASV system (Doddington et al., 1998). Thus, a diagonal value

[4]The similarity was initially computed as the sigmoid of the arithmetic mean of the LLRs (Noé et al., 2020). The intuition behind these two computations are the same and leads to close outcomes. However, from now on, the one proposed in Equation 6 will be preferred for its direct connection to the C_{llr} terms.

is related to the strength of the intra-speaker variability (Ajili, 2017). Actually, applying $-\log_2(\cdot)$ to the equation 6 gives:

$$-\log_2(\mathrm{Sim}(i,j)) = -\frac{1}{|\mathcal{L}_{\mathrm{i,j}}|}\sum_{l\in\mathcal{L}_{\mathrm{i,j}}}\log_2\sigma(l) = \langle -\log_2\sigma(l)\rangle_{l\in\mathcal{L}_{\mathrm{i,j}}}. \tag{7}$$

One may notice the resemblance with the target term of the C_{llr} in Equation 4 but it differs in the set of scores on which the average is done. In Equation 6, we average on the scores obtained from the comparisons of the two speakers while for the target term of the C_{llr}, the average is done on the scores from all target comparisons on the whole set. $-\log_2(\cdot)$ is not applied on the similarities in order to keep the values bounded between zero and one which facilitates the visualisation of the matrices.

The three voice similarity matrices, built from the original (O) and protected (P) sets of speech segments, used to visualise and assess the pseudonymisation performance are summarised here:

- M_{OO} which returns the speakers' similarities *within* the original set of speech segments,
- M_{OP} which returns the speakers' similarities *between* the original and protected sets,
- M_{PP} which returns the speakers' similarities *within* the protected set.

Next section presents how these matrices are used to measure the level of de-identification and voice distinctiveness preservation of a pseudonymisation system.

4.2. Assessing Pseudonymisation with the Voice Similarity Matrices

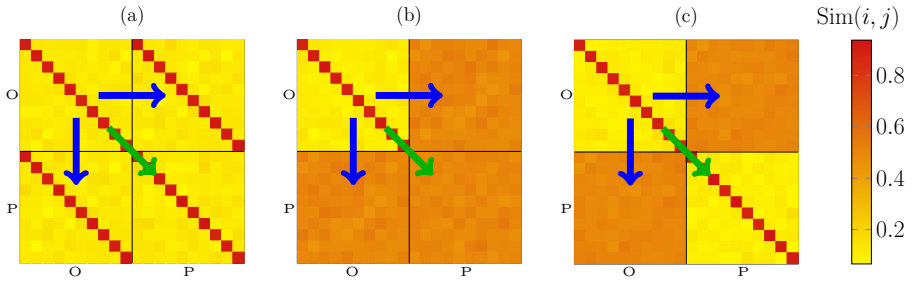


Figure 5: Three artificial examples of similarity matrices. For each case (a), (b) and (c), the upper-left matrix is M_{OO}, the upper-right and lower-left are M_{OP} (actually the lower-left matrix is $M_{\mathrm{PO}} = M_{\mathrm{OP}}^{\top}$ but as the biometric classifier is symmetric, the similarity is symmetric and $M_{\mathrm{PO}} = M_{\mathrm{OP}}$) whereas the lower-right is M_{PP}.

The emergence of the diagonal elements reflects how well an ASV system would behave when the two inputs of the biometric classifier may or may not be protected. Especially, in the case of M_{OP}, it shows if and which speaker could be

identified from its protected segments while the diagonal elements of M_{PP} show if the speakers have a low intra-speaker variability in the protected space and the off-diagonal elements show similarities between the different pseudo-voices. Therefore, the presence of a diagonal in M_{PP} suggests that each speaker has, in the protected domain, a consistent pseudo-voice that does not confuse with the others. In Figure 5 the cases (b) and (c) are examples of good de-identification (good protection) as the diagonal disappears in M_{OP} while the case (a) is an example of a poor de-identification. Cases (a) and (c) are examples where the voice distinctiveness is preserved as the diagonal remains in M_{PP}. Hence, the case (c) respects the two main objectives of the pseudonymisation where both de-identification and voice distinctiveness preservation are fulfilled. Thus, we propose to measure the pseudonymisation requirements by comparing the *dominance of the diagonal* between the matrices. The diagonal dominance of a matrix M is defined as:

$$D_{\mathrm{diag}}(M) = \left| \left(\sum_{1 \leq i \leq N} \frac{\mathrm{Sim}(i,i)}{N} \right) - \left(\sum_{\substack{1 \leq j \leq N \\ 1 \leq k \leq N \\ j \neq k}} \frac{\mathrm{Sim}(j,k)}{N(N-1)} \right) \right|, \tag{8}$$

such that it will be equal to zero for an uniform matrix and one for the identity matrix and for a matrix where all diagonal elements are zero and off-diagonal elements are one. Indeed, we consider that a scenario where a small group of segments is gathered as being from the same speaker if corresponding target trials result in high scores and impostor trials result in low scores and the scenario where a small group of segments is gathered as being from the same speaker if the target trials result in low scores and impostor trials result all in high scores are equivalent to the same risk of identification. However, if this property is appropriate for M_{OP} it might be an inconvenience for M_{PP} if a privacy safeguard results in a M_{PP} where diagonal values are low and off-diagonal values are high and close together.

The De-Identification measure is how much the diagonal disappears from M_{OO} to M_{OP} as illustrated by the blue arrows in Figure 5 and is computed, assuming that $D_{\mathrm{diag}}(M_{\mathrm{OO}}) \neq 0$, as follow:

$$\mathrm{DeID} = 1 - \frac{D_{\mathrm{diag}}(M_{\mathrm{OP}})}{D_{\mathrm{diag}}(M_{\mathrm{OO}})}. \tag{9}$$

DeID = 100% is the perfect de-identification while DeID = 0% corresponds to a system which achieves no de-identification or at least that the resulting privacy is not better than the initial one. Indeed, the initial privacy refers to the limit capacity of an ASV system to discriminate the original voices. To help illustrate this, we consider a pair of twins who have exactly the same voice such that the ASV system can not distinguish between them. In this closed case, privacy is already high, thus, only a *relative* measure of privacy

that measures the improvement in privacy, as opposed to the absolute level, will reflect meaningfully the performance of the safeguard.

The voice distinctiveness preservation is how much the diagonal remains from M_{OO} to M_{PP} as illustrated by the green arrows in Figure 5. Because a privacy safeguard can result in either a loss, or an increase of voice distinctiveness, the voice distinctiveness preservation is computed as a gain of diagonal dominance:

$$G_{\mathrm{VD}} = 10 \log_{10} \left(\frac{D_{\mathrm{diag}}(M_{\mathrm{PP}})}{D_{\mathrm{diag}}(M_{\mathrm{OO}})} \right), \tag{10}$$

such that a gain equal to zero means that the voice distinctiveness remains, in average, the same and a gain above or below zero corresponds respectively to an increase or a loss of the global voice distinctiveness.

As we assumed that a privacy safeguard will never result in less privacy, a percentage of de-identification in Equation 9 is more suitable than a gain like G_{VD}[5]. Both are global metrics and thus reflect only an averaged performance over the set of speaker. Next section motivates how the matrices can be used for an assessment speaker by speaker.

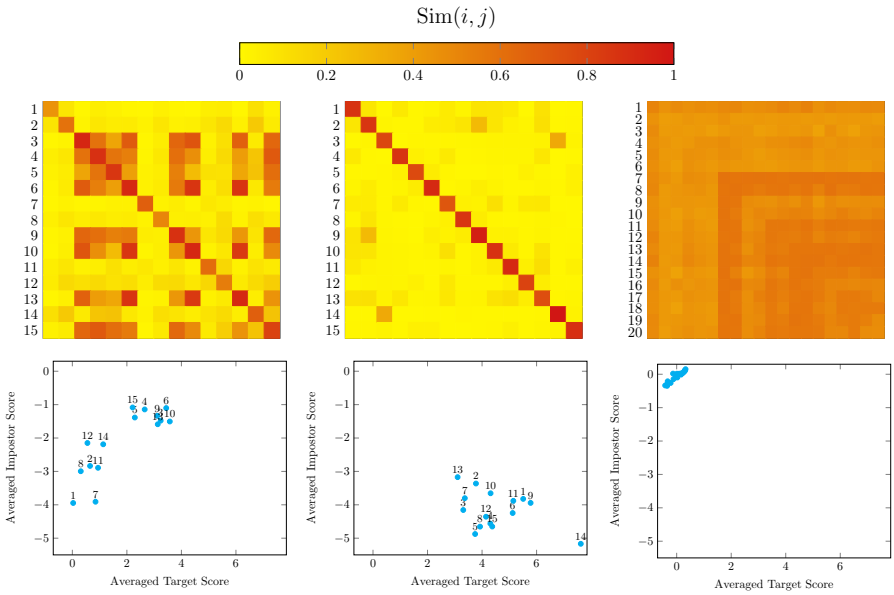
4.3. A Zoo Tour: An investigation at the Speaker level

An ASV system does not perform equally on all speakers. In order to better interpretate the ASV behavior across different speakers, George Doddington proposed to categorise speakers into four different animals (Doddington et al., 1998). This idea has been extended to other categories and the Zoo plot has been proposed in order to visualise the performance of a biometric system across different users (Dunstone & Yager., 2009). This section deals with the use of voice similarity matrix to visualise, like the Zoo plots, the performance of a system at the speaker level. Figure 6 shows few examples of voice similarity matrix with the corresponding Zoo plot below. The latter shows speakers in an averaged target/impostor score plane. In these examples, one can notice that different regions of the plane are covered. For the matrix in Figure 6a, there are mostly two kinds of resulting pseudo-voices: those which do not confuse with others but have a self-similarity not so high (1, 2, 7, 8, 11, 12 and 14) and those which confuse with some others but have a higher self-similarity (3, 4, 5, 6, 9, 10, 13 and 15). These two kinds correspond to the two clusters in the Zoo plot (bottom figure) in 6a making them close to goats and wolves respectively. Figure 6b shows an example where the ASV performs almost equally well on all speakers which correspond to high averaged target and low averaged impostor scores (sheeps). Lastly, 6c shows an example of a good privacy preservation, corresponding to poor ASV performance with an OP setting. Speakers are thus placed in the Zoo plot with zero scores making them close to worms.

These examples show how the voice similarity matrices can be used in order to visualise and analyse heterogeneous performance across the set of speakers.

[5] In electrical engineering, telecommunication and acoustics, it is common to express a gain in decibel.

Section 6 will present further results obtained on the VoicePrivacy challenge baselines.



(a) M_{PP} of vctk_test_trials_f_c with B2r and the Zoo plot.

(b) M_{OO} of vctk_test_trials_m original and the Zoo plot.

(c) M_{OP} of libri_test_trials_f with B1 and the Zoo plot.

Figure 6: Examples of Voice Similarity Matrix with the corresponding Zoo plot (Dunstone & Yager., 2009). See Section 4.3 for details on the experimental setup.

5. Zero Evidence Biometric Recognition Assessment

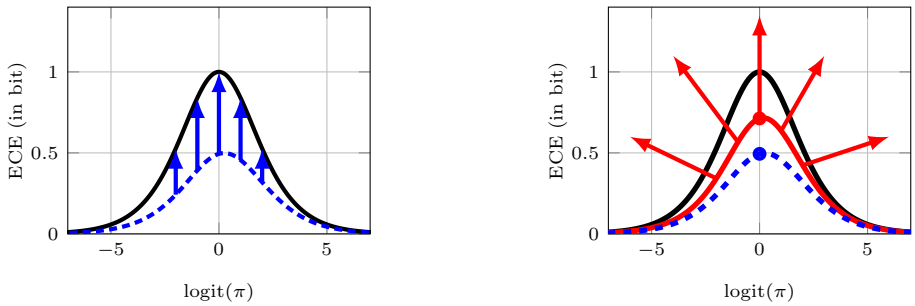
The ZEBRA framework (Nautsch et al., 2020) is based on the perfect privacy (Shannon, 1949) and the Empirical Cross Entropy (ECE) used in forensic speaker recognition (Ramos-Castro & González-Rodríguez, 2008; Ramos-Castro, 2007; Ramos et al., 2018). The narrative and derivation of ZEBRA's ECE profiles is detailed in (Nautsch et al., 2020) for LR scores. Here, we outline the relevant equations to compute the ZEBRA's metrics for LLR scores. For a set of propositions $\Theta = \{\theta_{\mathrm{tar}}, \theta_{\mathrm{imp}}\}$, the posterior ECE plot or profile is defined as a function of the prior $\pi = P(\theta_{\mathrm{tar}})$:

$$\begin{aligned} \mathrm{ECE}_{\Theta|\mathcal{L}}(\pi) = \quad & \pi \quad \langle -\log_2 \sigma(+a + \mathrm{logit}(\pi)) \rangle_{a \in \mathcal{L}_{\mathrm{tar}}} \\ & +(1-\pi) \quad \langle -\log_2 \sigma(-b + \mathrm{logit}(1-\pi)) \rangle_{b \in \mathcal{L}_{\mathrm{imp}}}. \end{aligned} \tag{11}$$

At a given prior π, the ECE provides the amount of missing information needed for making a good decision. With a prior $\pi = 0.5$, the ECE corresponds to the C_{llr} (Brümmer & du Preez, 2006). To compute the prior ECE profile, all

LLR scores are set equal to zero (score set $\mathbf{0}_{\mathcal{L}}$); the above equation simplifies because of $\sigma(\text{logit}(x)) = x$ to:

$$\text{ECE}_{\Theta|\mathbf{0}_{\mathcal{L}}}(\pi) = -\pi \log_2(\pi) - (1-\pi)\log_2(1-\pi). \tag{12}$$



(a) ZEBRA idea in ECE plot.

(b) ZEBRA idea for actual ECE.

Figure 7: ECE profiles: perfect privacy (black), expected privacy disclosure to adversary (blue), and expected privacy disclosure with adversary using imperfectly calibrated system (red). The dots at $\text{logit}(\pi) = 0$ indicate the C_{llr} (red) and the $C_{\text{llr}}^{\text{min}}$ (blue).

Figure 7 shows three ECE profiles. The prior ECE (black profile), and two evidence-informed ECEs: the blue profile shows ideally calibrated scores (discussed in the original ZEBRA framework and in Section 5.1) and the red profile shows not perfectly calibrated scores (discussed in Section 5.3). To privacy adversaries, ideally calibrated scores refer to systems where the classification model is as useful as it can be while uncalibrated scores refer to systems whose models are less useful and if the red profile exceeds the black profile (as shown by the red arrows), adversaries are better off by using random choice than the model.

5.1. The Expected Privacy Disclosure

Intuitively, for perfect privacy, the blue profile must fit the black one meaning that, for every π, the posterior knowledge remains the prior one. By comparing these two profiles, the ZEBRA approach provides the expected amount of information disclosed (by scores) from which the adversary can benefit. In addition, it provides the strongest LLR observed in an experiment as the *worst-case* scenario (for the privacy preserver while this is the *best-case* for the attacker). Alongside this score, a categorical tag is given for better human interpretation. These last two points are discussed in the next section.

The expected privacy disclosure $D_{\text{ECE}}(\Theta|\mathcal{L})$ is referred to the area between two ECE profiles: the prior ECE and the posterior ECE thus averaging the comparisons between the attacker's prior and posterior knowledge over all priors

and is computed as follow:

$$
\begin{aligned}
D_{\mathrm{ECE}}(\Theta|\mathcal{L}) &= \int \mathrm{ECE}(\Theta|\mathbf{0}_{\mathcal{L}}) - \mathrm{ECE}(\Theta|\mathcal{L})\,\mathrm{d}\pi \\
&= \frac{\langle Z(a)\rangle_{a\in\mathcal{L}_{\mathrm{tar}}} + \langle Z(-b)\rangle_{b\in\mathcal{L}_{\mathrm{imp}}}}{2\,\ln(2)} \\
\text{with } Z(l) &= \frac{1}{2} + \frac{l-(\exp(l)-1)}{(\exp(l)-1)^2}, \qquad \begin{smallmatrix}\lim\limits_{l\mapsto-\infty} Z(l)=-\infty,\\ \lim\limits_{l\mapsto 0} Z(l)=0,\\ \lim\limits_{l\mapsto\infty} Z(l)=\frac{1}{2}.\end{smallmatrix}
\end{aligned}
\tag{13}
$$

For perfect privacy, we want for every prior no information revealed to the attacker by the scores and thus a minimum D_{ECE} equal to zero. Negative input values to $Z(l)$ are possible especially for uncalibrated scores, then, depending on the prior, the posterior ECE can exceed the prior ECE like the red profile in Figure 7. In (Nautsch et al., 2020), oracle score calibration ensures the majority of target LLRs to be larger than zero and the majority of impostor LLRs to be lesser than zero. This is consistent with information theory: evidence updates knowledge and thus reduces uncertainty; badly calibrated scores indicate classifiers that are not capable of adequately resembling the feature space's class distribution in the score space (such classifiers need additional knowledge to be informative; i.e., score calibration). This discrepancy arises from the ECE being a measure of cross-entropy between the reference probability space P and a classifier's probability space $\tilde{P}$ (between nature and a model). The ECE approximates the cross-entropy $H_{P||\tilde{P}}(\Theta|\mathcal{L})$:

$$
H_{P||\tilde{P}}(\Theta|\mathcal{L}) = -\sum_{\theta\in\Theta} P(\theta) \int_l P(l|\theta)\log_2 \tilde{P}(\theta|l)\,\mathrm{d}l. \tag{14}
$$

By considering a large number of scores, the conditional probability $P(l|\theta)$ is approximated by $\frac{1}{|\mathcal{L}_\theta|}$ (Ramos-Castro & González-Rodríguez, 2008; Ramos-Castro, 2007; Ramos et al., 2018). If scores are not perfectly calibrated, we talk about the actual ECE (red profile). When the red profile indicates less uncertainty than indicated by the black profile (red profile below the black one), scores are referred to be calibrated even if not being ideally calibrated.

5.2. Worst-Case Privacy Disclosure

The worst-case privacy disclosure refers to the strongest strength-of-evidence remaining after applying a privacy safeguard on the raw audio data. It is obtained easily in the LLR domain by: $l_w = \max_{l\in\mathcal{L}'} \mathrm{abs}(l)$. $\mathcal{L}'$ are Bayesian predictions for extreme target/impostor oracle LLRs[6]. In order to facilitate its

[6]For the target and impostor propositions, both a zero and a one posterior are possible so that it is as if we had already seen such values. Therefore these extreme values are added in both target and impostor score sets when applying PAV algorithm. It is known as *Laplace's rule of succession* and is used here to avoid infinite likelihood ratios (Brümmer & De Villiers, 2011).

interpretability, it is quantised into *categorical tags*, an idea that has been already applied to forensic science (Nordgaard et al., 2011). The correspondence between the score ranges and the tags is given in Table 1 as well as the odds ratio to give more intuition.

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

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

5.3. Extension of the ZEBRA framework: Score Calibration

One strategy for the privacy preserver is to reduce the discrimination between the target and impostor score distributions. An alternative is to make score calibration impossible for the attacker. The latter calibrates scores to obtain the necessary knowledge in order to make the ideal choices regardless of its classifier's discrimination power. If scores are badly calibrated such that the ECE profile (red) fits or exceeds the black one, the attacker is better off tossing a coin. There are two commonly used calibration methods: linear and isotonic regression (Brümmer & De Villiers, 2011). Linear regression[7] is more robust but less accurate, whereas isotonic regression is sensitive to changes but results in oracle calibration.[8] Conventionally, training and testing a calibration method on the same dataset is considered cheating. However, this is useful to assess privacy preservation systems. Depending on the privacy safeguard, running it twice on a the same set can lead to different outcomes resulting in two different protected sets. In this way, training a calibration on one of these sets and testing it on the other would inform on the ability for the calibration to generalise on another instance of the privacy safeguard system. The optimisation criterion for score calibration is C_{llr} (to reach $C_{\text{llr}}^{\text{min}}$), so as a first choice one might measure the calibration loss $C_{\text{llr}}^{\text{cal}} = C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$. In ECE plots, this is representative for the prior $\pi = 0.5$ only. The ZEBRA framework is extended to quantify the extent to which score calibration fails. Here, we assess whether or not score calibration can be misused by an adversary to invert the attempt of privacy safeguards to interfere with her decision making. Let $\mathcal{E}$ be an unprotected dataset of evidence, $f_t(\mathcal{E})$ be a safeguard running that manipulates $\mathcal{E}$ at

[7] Also referred as logistic regression (Brümmer & De Villiers, 2011). It corresponds to a linear regression in the logit space.

[8] Pointers to MATLAB code: `https://sites.google.com/site/bosaristoolkit` and to Python code: `https://gitlab.eurecom.fr/nautsch/pybosaris`

a time instant t, $\mathcal{S}(\mathcal{E}')$ scores computed from a set of (protected or unprotected) evidences $\mathcal{E}'$ and we use the notation $\mathcal{S}_{\text{or}}(\mathcal{E}')$ when scores are oracle calibrated. Let also $c_\varphi(\mathcal{X}|\mathcal{Y})$ be a function to calibrate scores $\mathcal{X}$ (c_{lin} for linear and c_{iso} for isotonic regression)[9] which is trained on the score set $\mathcal{Y}$. We simulate an adversary observing two protected datasets of same origin (without knowing) who wants to yield calibrated scores $\mathcal{L}_\varphi$:

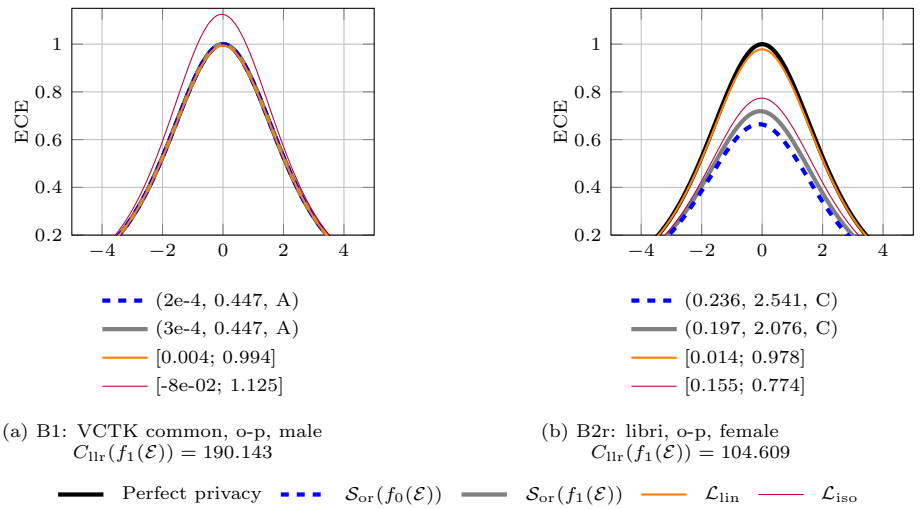
$$\mathcal{L}_\varphi = c_\varphi(\mathcal{S}(f_1(\mathcal{E}))\,|\,\mathcal{S}(f_0(\mathcal{E}))). \tag{15}$$

Then, the expected calibration distortion $C_{\text{ECE}}(\Theta|\mathcal{L}_\varphi)$ is computed simply as:

$$C_{\text{ECE}}(\Theta|\mathcal{L}_\varphi) = D_{\text{ECE}}(\Theta|\mathcal{L}_\varphi). \tag{16}$$

The difference to the above $D_{\text{ECE}}(\Theta|\mathcal{L})$ is that $\mathcal{L}_\varphi$ is not derived through oracle score calibration. The scores of $\mathcal{L}_\varphi$ are desired by an adversary to be well-calibrated but one purpose of a safeguard is to countermeasure that. This is evaluated by the expected calibration distortion which should be as low as possible: equal to zero is good, negative infinity is best. Since the expected privacy disclosure of $f_0(\mathcal{E})$ does not need to equal the expected privacy disclosure of $f_1(\mathcal{E})$, their comparison will inform on potential variations of a safeguard when ensuring zero evidence by means of class discrimination. Figure 8 shows ZEBRA plot examples to explain C_{ECE} in the context of the 2020 VoicePrivacy challenge. To simulate calibration we use two algorithms: linear regression ($\mathcal{L}_{\text{lin}}$) as in (Brümmer & De Villiers, 2011) and isotonic regression ($\mathcal{L}_{\text{iso}}$) as in (Brümmer & Preez, 2013). Here, the isotonic regression is a mapping function for all possible uncalibrated scores (it is not oracle score calibration). ECE profiles with perfectly calibrated scores are shown in blue and gray (these are the ones used to compute the D_{ECE}), actual ECE profiles are shown for (not perfectly) calibrated scores in violet and orange (used to compute the C_{ECE}) and perfect privacy profile is in black. Both figures show results where the attacker compares protected segments with unprotected ones. For (a), we can see that perfect privacy is almost reached for the two instances of the safeguard (D_{ECE} close to 0). The red profile is above the perfect privacy (negative C_{ECE}), which means that the isotonic regression learnt on the first instance of the safeguard can not be used to calibrated scores on the second instance. Even if it is better ($C_{\text{llr}}(\mathcal{L}_\phi)$ is lower than $C_{\text{llr}}(f_1(\mathcal{E}))$) for the attacker than without attempting to calibrated the scores, the resulting attacker's posterior uncertainty is still bigger than its prior one (more precisely, this is the posterior cross-entropy that is bigger than the prior entropy). In (b), the amount of privacy varies by instances. While both do not reach perfect privacy, the second instance (gray profile) fair better. Calibration with linear regression (orange profile) results almost in perfect privacy but unfortunately for the preserver, isotonic regression allows better calibration reducing the attacker's uncertainty.

[9] Actually, a calibration method could be any kind of score transformation, e.g., to relocate θ_{imp} scores that are in-between two clusters of θ_{tar} scores; see (Maouche et al., 2020).



(a) B1: VCTK common, o-p, male
$C_{\text{llr}}(f_1(\mathcal{E})) = 190.143$

(b) B2r: libri, o-p, female
$C_{\text{llr}}(f_1(\mathcal{E})) = 104.609$

Figure 8: Examples of ZEBRA plots to illustrate the calibration distortion. The black profile is perfect privacy. The blue profile is for privacy disclosure on an instance $f_0(\mathcal{E})$ of the safeguard and is thus ECE profile with oracle calibrated scores. The gray profile is the same but on another instance $f_1(\mathcal{E})$. For these two, $(D_{\text{ECE}}, l_w, \text{tag})$ is provided in the legend. Orange and violet profiles are actual ECE plots on $f_1(\mathcal{E})$ where scores have been calibrated with respectively a linear ($\mathcal{L}_{\text{lin}}$) and a isotonic ($\mathcal{L}_{\text{iso}}$) regression learnt on the other instance $f_0(\mathcal{E})$. For these two, $[C_{\text{ECE}}; C_{\text{llr}}(\mathcal{L}_\varphi)]$ is provided legend.

6. Experiment: A Case Study with both frameworks

Reported in this section are example evaluations and analyses of the VoicePrivacy 2020 Challenge baseline systems using the two assessment frameworks presented above. The analysis concerns only pseudonymisation; we report the level of protection (DeID and ZEBRA's metrics) and the voice distinctiveness preservation (G_{VD}) of a privacy safeguard only. Impacts upon other speech attributes such as quality or intelligibility are beyond the scope of this work.

Three pseudonymisation systems will be used. The first one (Fang et al., 2019) is based on x-vector pooling and neural waveform resynthesis. The second one (Patino et al., 2020) is based on vocal tract filter transformations using McAdams coefficient (McAdams, 1984). The third one is a variation of the second baseline that uses a stochastic transformation approach to improve irreversibility (Patino et al., 2020). From here on in, these three baseline systems are referred to as B1, B2 and B2r ('r' for random) respectively. In our experiments, for B2r, the McAdams coefficient is sampled, from a speaker to another, uniformly between 0.7 and 0.9 (Patino et al., 2020).

All experiments were performed using the trial part of the challenge enrolment and test sets. These are the LibriSpeech[10] test-clean dataset and a subset

[10] `https://openslr.org/12`

of the VCTK (Veaux et al., 2019) dataset. Each are split into male and female partitions. A separate partition of the VCTK dataset, denoted with the _c abbreviation, contains utterances of identical linguistic content (the same spoken sentence). Further details concerning the use of each database can be found in (Tomashenko et al., 2020a,b).

Scores are obtained by comparing segments using the x-vector speaker embedding with the probabilistic linear discriminant analysis (PLDA) (Snyder et al., 2018) trained on the LibriSpeech train-clean-360 dataset. Scores are oracle calibrated using the Pool Adjacent Violators algorithm (Brümmer & Preez, 2013).

Results on each of the baselines are presented in Table 2. It reports the metrics based on the voice similarity matrices: the de-identification (DeID) and the gain of voice distinctiveness (G_{VD}); as well as ZEBRA's outputs: the expected (D_{ECE}) and worst-case privacy disclosure (l_w) and the corresponding categorical tag (Tag).

Both frameworks agree that B1 gives good protection relatively to B2 and B2r. Indeed, it leads to a DeID close to 100% and a low D_{ECE} for all sets. However, protection is to the detriment of voice distinctiveness (low G_{VD}), making this system less suitable for applications such as teleconferencing or speaker diarization. B2 and B2r better preserve voice distinctiveness but provide a lower protection as their G_{VD} values are closer to zero ($G_{\mathrm{VD}} = 0$ means that the voice distinctiveness remains globally the same, no lose, no gain) and DeID are lower. A relatively weak protection for B2 and B2r is also shown by the high l_w values and the categorical tags. This is not surprising, since the expected privacy disclosure (D_{ECE}) for B2 and B2r are also higher. However, even if it is not shown by the presented results, one can meet a system which performs globally better than an another one but may leave some segments with a worst protection. l_w and its categorical tag are necessary to detect such a scenario. Voice similarity matrices are illustrated in Figure 9 for three of the six datasets. The better de-identification provided by B1 is evident from the lack of a distinctive diagonal and the uniformity of the M_{OP} matrices (upper-right and lower-left) which indicate good protection for each speaker. For B2 and B2r systems, the distinctive diagonal indicates inferior de-identification performance and the non-uniformity of M_{OP} matrices' diagonal values relatively to the off-diagonal values indicates that speakers do not receive the same level of protection. The lack of a distinctive diagonal for M_{PP} matrices (lower-right) confirms that B1 fails to preserve the voice distinctiveness, while B2 and B2r systems fair better. M_{PP} matrices also expose some imbalances level of voice distinctiveness. M_{PP} matrices for B2r show off-diagonal values which are, mostly, either high (red) or low (yellow) meaning that the system results in a group of voices that are well distinguishable and a group in which voices look alike as discussed in Section 4.3.

Figure 10 shows a visualisation of ZEBRA's ECE profiles for three of the six datasets. Green profiles correspond to B1, blue profiles to B2, red profiles to B2r, gray to the unprotected original data and black to perfect privacy. B1's ECE profiles overlap almost perfectly with the perfect privacy profile, thereby

Table 2: Results of the voice similarity matrices and ZEBRA based metrics on the two official challenge's baselines B1 and B2 (Tomashenko et al., 2020a) and its updated random version B2r (with $\alpha \sim \mathcal{U}(0.7, 0.9)$) (Patino et al., 2020) for each of the challenge's test sets. DeID gives the level of privacy reached in regards to the initial one and G_{VD} gives the resulting gain of voice distinctiveness. The D_{ECE} gives the expected privacy disclosure and on the right, the worst-case privacy disclosure is given in the form of l_w and categorical tag. C_{ECE} is reported for linear score calibration for B1 and B2r but not for B2 as it is deterministic such that running it twice on the same set results in two identical protected sets. Under each set name, their initial (D_{ECE}, l_w, Tag) are provided i.e. when neither the enrolment nor the test sets have been protected.

Set	System	DeID [%]	G_{VD} [dB]	D_{ECE}/ C_{ECE} [bit]	l_w	Tag
libri_test_trials_f (0.584, 3.979, C)	B1	97.85	-10.45	0.004/1e-4	0.310	A
	B2	45.02	-1.45	0.221	2.266	C
	B2r	53.51	-1.94	0.236/0.014	2.541	C
libri_test_trials_m (0.690, 3.924, C)	B1	99.99	-9.56	0.001/0.002	0.282	A
	B2	46.55	-1.40	0.354	2.614	C
	B2r	58.08	-1.66	0.224/0.267	2.528	C
vctk_test_trials_f (0.594, 3.655, C)	B1	99.87	-9.82	0.001/-4e-7	0.128	A
	B2	85.86	-4.57	0.141	2.395	C
	B2r	79.98	-1.46	0.090/0.071	2.189	C
vctk_test_trials_m (0.667, 3.921, C)	B1	99.99	-11.33	3e-05/0.003	2e-04	A
	B2	74.04	-3.42	0.196	3.012	C
	B2r	71.60	-1.69	0.162/0.128	2.425	C
vctk_test_trials_f_c (0.653, 3.557, C)	B1	99.09	-8.93	0.004/2e-4	0.668	A
	B2	69.36	-2.90	0.132	1.197	B
	B2r	64.69	-1.61	0.153/0.075	2.507	C
vctk_test_trials_m_c (0.694, 3.675, C)	B1	99.99	-10.31	2e-04/0.004	0.447	A
	B2	57.20	-2.29	0.199	2.488	C
	B2r	63.39	-1.63	0.218/0.125	2.909	C

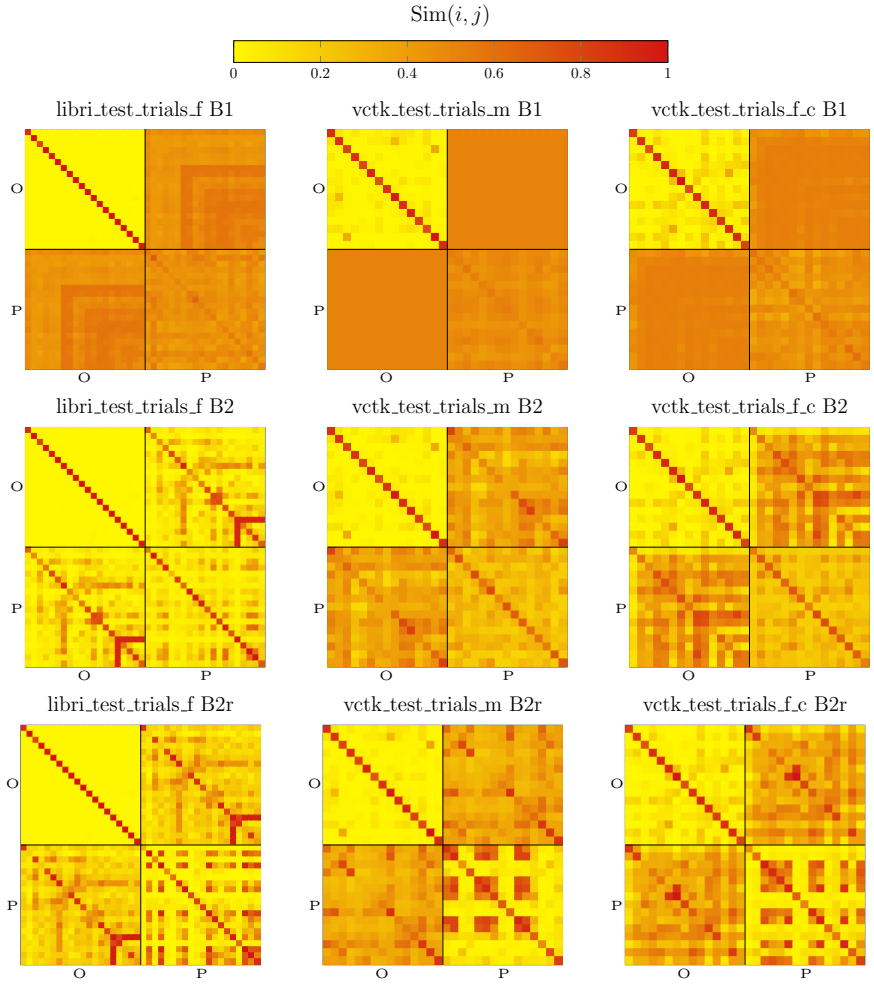
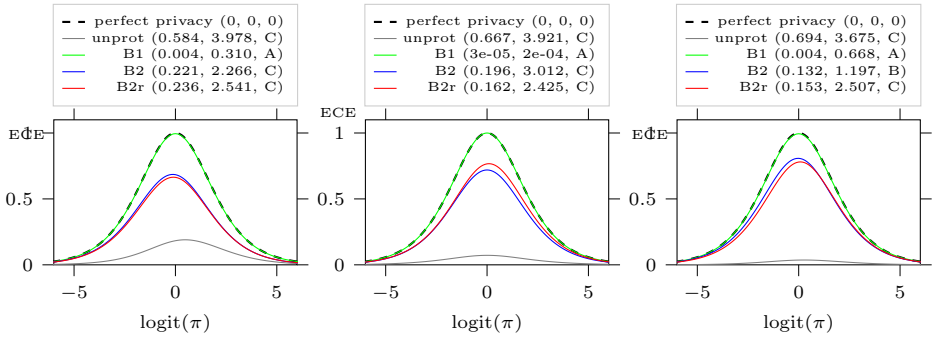


Figure 9: Voice similarity matrices obtained with the three systems applied on three of the test sets. The set name and the system used are indicated above each matrix.



(a) On libri_test_trials_f. (b) On vctk_test_trials_m. (c) On vctk_test_trials_f_c.

Figure 10: ZEBRA's ECE profiles on three of the test sets: perfect privacy (black), expected privacy disclosure to adversary without protection (unprot, gray) and expected privacy disclosure to adversary on data protected with B1 (green), B2 (blue) and B2r (red).

confirming the protection delivered by B1. There are almost no differences between B2 and B2r profiles. These observations are all consistent with the D_{ECE} values in Table 2. The ECE profiles are all symmetric around $\pi = 0.5$, however, this is the result of the present baselines and the sets used and is not an intrinsic property of the ECE; some privacy safeguards may result in asymmetric ECE profiles.

The ZEBRA's metrics reported here are used in order to measure the protection level of a privacy safeguard. However, in the next section, we will see how D_{ECE} can be used to assess the voice distinctiveness preservation and therefore compare it with the gain of voice distinctiveness G_{VD}.

7. Comparison of the two Frameworks

Before comparing the results of the two frameworks and discuss there inherent differences, this section presents reformulations, and an additional used of the expected privacy disclosure to measure the voice distinctiveness preservation, in order to better compare the two approaches.

7.1. Comparison of the frameworks' findings

By normalising $D_{\text{diag}}(M_{\text{OP}})$ (in Equation 9), DeID gives a percentage of de-identification reached by a system in relation to the initial level of privacy. Applying the same normalisation strategies to $D_{\text{ECE}}^{(\text{OP})}$ and $C_{\text{llr}}^{\text{min(OP)}}$ allows to better compare these metrics[11]:

[11] The notation $m^{(\cdot\cdot)}$ indicates in which setting the metric m is computed (OO, OP or PP).

Table 3: DeID, $D_{\text{ECE}}^{(\text{OP/OO})}$ and $C_{\text{llr}}^{\min(\text{OP/OO})}$ obtained with B1, B2 and B2r on each of the challenge's test sets. Each is a measure of the de-identification of the system. The values are reported with a two digits precision after the decimal point.

Set	System	DeID [%]	$D_{\text{ECE}}^{(\text{OP/OO})}$ [%]	$C_{\text{llr}}^{\min(\text{OP/OO})}$ [%]
libri_test_trials_f	B1	97.85	99.32	99.39
	B2	45.02	62.16	61.32
	B2r	53.51	59.59	58.87
libri_test_trials_m	B1	99.99	99.86	99.9
	B2	46.55	48.70	47.76
	B2r	58.08	67.54	66.74
vctk_test_trials_f	B1	99.87	99.83	99.78
	B2	85.86	76.26	75.21
	B2r	79.98	84.85	84.12
vctk_test_trials_m	B1	99.99	∼100	∼100
	B2	74.04	70.61	69.83
	B2r	71.60	75.71	74.78
vctk_test_trials_f_c)	B1	99.09	99.39	99.34
	B2	69.36	79.79	78.88
	B2r	64.69	76.57	75.8
vctk_test_trials_m_c	B1	99.99	99.97	∼100
	B2	57.20	71.33	70.23
	B2r	63.39	68.59	67.53

$$D_{\text{ECE}}^{(\text{OP/OO})} = 1 - \frac{D_{\text{ECE}}^{(\text{OP})}}{D_{\text{ECE}}^{(\text{OO})}}, \tag{17}$$

$$C_{\text{llr}}^{\min(\text{OP/OO})} = \frac{C_{\text{llr}}^{\min(\text{OP})} - C_{\text{llr}}^{\min(\text{OO})}}{1 - C_{\text{llr}}^{\min(\text{OO})}}. \tag{18}$$

With this normalisation, DeID, $D_{\text{ECE}}^{(\text{OP/OO})}$ and $C_{\text{llr}}^{\min(\text{OP/OO})}$ represent all a percentage of de-identification, 100% means full or perfect de-identification whereas 0% means no improvement of the average level of privacy. Table 3 reports these three measures for the three baselines. While both frameworks again show that B1 outperforms both B2 and B2r systems, there are inconsistencies in the ranking of B2 and B2r. $D_{\text{ECE}}^{(\text{OP/OO})}$ and $C_{\text{llr}}^{\min(\text{OP/OO})}$ rankings for each dataset are consistent[12] but DeID gives different findings.

It will now be shown how to measure a gain of voice distinctiveness from

[12]Even if, according to the presented results, the D_{ECE} and $C_{\text{llr}}^{\min}$ lead to the same findings, it might not be the case when studying other privacy preserving systems and other sets. Indeed, we recall that the $C_{\text{llr}}^{\min}$ consider a single prior ($\pi = 0.5$) and is thus sensitive to the adversary decision policy while D_{ECE} is independent of it.

Table 4: G_{VD}, $G_{D_{\mathrm{ECE}}^{(\mathrm{PP/OO})}}$ and $G_{C_{\mathrm{llr}}^{\mathrm{min(PP/OO)}}}$ obtained with B1, B2 and B2r on each of the challenge's test sets. Each is a measure of the gain of voice distinctiveness of the system.

Set	System	G_{VD} [dB]	$G_{D_{\mathrm{ECE}}^{(\mathrm{PP/OO})}}$ [dB]	$G_{C_{\mathrm{llr}}^{\mathrm{min(PP/OO)}}}$ [dB]
libri_test_trials_f	B1	-10.45	-7.21	-7.05
	B2	-1.45	-2.15	-2.06
	B2r	-1.94	-4.38	-4.24
libri_test_trials_m	B1	-9.56	-10.19	-9.95
	B2	-1.4	-1.20	-1.15
	B2r	-1.66	-5.19	-5.02
vctk_test_trials_f	B1	-9.82	-7.57	-7.35
	B2	-4.57	-2.75	-2.64
	B2r	-1.46	-10.26	-10.06
vctk_test_trials_m	B1	-11.33	-7.71	-7.47
	B2	-3.42	-1.97	-1.88
	B2r	-1.69	-4.51	-4.40
vctk_test_trials_f_c)	B1	-8.93	-7.47	-7.28
	B2	-2.9	-2.39	-2.29
	B2r	-1.61	-12.83	-12.60
vctk_test_trials_m_c	B1	-10.31	-7.88	-7.67
	B2	-2.29	-1.80	-1.74
	B2r	-1.63	-8.92	-8.73

D_{ECE} and also from $C_{\mathrm{llr}}^{\mathrm{min}}$ in order to compare with G_{VD}. Similarly to Equation 10, we express gains of voice distinctiveness using D_{ECE} and $C_{\mathrm{llr}}^{\mathrm{min}}$ as:

$$G_{D_{\mathrm{ECE}}^{(\mathrm{PP/OO})}} = 10\log_{10}\left(\frac{D_{\mathrm{ECE}}^{(\mathrm{PP})}}{D_{\mathrm{ECE}}^{(\mathrm{OO})}}\right), \tag{19}$$

$$G_{C_{\mathrm{llr}}^{\mathrm{min(PP/OO)}}} = 10\log_{10}\left(\frac{1-C_{\mathrm{llr}}^{\mathrm{min(PP)}}}{1-C_{\mathrm{llr}}^{\mathrm{min(OO)}}}\right). \tag{20}$$

A comparison of G_{VD}, $G_{D_{\mathrm{ECE}}^{(\mathrm{PP/OO})}}$ and $G_{C_{\mathrm{llr}}^{\mathrm{min(PP/OO)}}}$ given in Table 4 shows similar system rankings. However, G_{VD} exhibits ranking differences for B2 and B2r. For B2r, G_{VD} gives much greater values in comparison to the other gains. It suggests that G_{VD} is more lenient when a system results in a group of pseudo-voices that are distinguishable and another where they are not (see Figure 6a).

7.2. Intrinsic differences between the two frameworks

Results presented in Section 7.1 showed consistencies between D_{ECE} and $C_{\mathrm{llr}}^{\mathrm{min}}$ but differences with the metrics obtained from the voice similarity matrices. While the connection between D_{ECE} and $C_{\mathrm{llr}}^{\mathrm{min}}$ is clear (the latter is the posterior empirical cross entropy for $\pi = 0.5$), the link between D_{diag} (from which DeID and G_{VD} are computed) and the $C_{\mathrm{llr}}^{\mathrm{min}}$ is not so straightforward. The diagonal dominance measure were designed independently from the binary

decision making and the LLR goodness. It is defined in Eq. 8 as the modulus of the difference between the averaged self-similarity (target comparisons) and the averaged similarity between different speakers (impostor comparisons). In this sense, diagonal dominance is similar to the C_{llr} where there is one cost for target comparisons and another for impostor comparisons. However, the speaker-wise averaging in D_{diag} allows to consider all speakers equally independently of their number of segments in the set. Moreover, as discussed in Section 3.4, the logarithmic representation allows additive comparisons of probabilities but, unlike the C_{llr}, D_{diag} does not benefit from such a linear representation which is another distinctive characteristic of this measure.

8. Conclusion and Discussion

After providing a definition for both anonymisation and pseudonymisation in the scope of privacy preservation in speech, this paper has presented, extended and compared two frameworks for the assessment of pseudonymisation systems. The first one proposes to compute a de-identification measure DeID and a gain of voice distinctiveness G_{VD} by comparing the amount of diagonal in voice similarity matrices. While the diagonal dominance measure D_{diag} is designed independently of the goodness of the log-likelihood-ratio, its speaker-wise treatment is of interest and the intuitive visualisation of the voice similarity matrices allows to rigorously analyse the heterogeneous performance of a system at the speaker level. The other framework named ZEBRA, measures the amount of speaker identity information released by a pseudonymisation system. It can also be used to measure how score calibration by the adversary is made hard. The expected privacy disclosure D_{ECE} can be used for de-identification or for pseudo-voice distinctiveness assessment when comparisons are respectively done on unprotected-protected segments and protected-protected segments. This approach is fully compatible with the Bayesian decision framework and can be used for the assessment of any biometric systems.

In the future, comparing the diagonal and off-diagonal terms of a voice similarity matrix in the log domain in the computation of D_{diag} could benefit from the speaker-wise averaging while being consistent with the LR paradigm.

Regarding the ZEBRA framework, the worst-case privacy disclosure returns the strongest strength-of-evidence given by a segment. In future works, this framework could also be extended to the speaker level to inform on which speakers the pseudonymisation system is not or less effective.

Regarding the relevance of the voice distinctiveness preservation measures, one have to keep in mind that they depend on the ASV system and thus do not reflect the human perception of voice. Hence, voice distinctiveness in the protected domain could be improved by adding pseudo-voices in data augmentations while training the ASV system in order to compensate the nuisance that a pseudonymisation system may introduce. A low voice distinctiveness may also be the result of pseudo-voices which does not sound natural, thus, these measures should of course be interpreted along with speech quality measures.

A privacy preserver can design a protection system in many different ways. She or he might create systems that apply at different levels: a global system that is applied equally on several speakers for several sessions, or systems that are applied differently across sessions, across speakers or even across segments. The following example aims to justify the consideration of these different levels of application. Pseudonymisation may not be irreversible, thus, applying the same voice mapping over sessions would allow an adversary to accumulate knowledge in order to infer the inverse voice mapping. Having a different voice mapping for each session would prevent an adversary to use knowledge learned in a session in order to attack another one. Future works could also investigate assessment methods that take into account this multi-session aspect.

Acknowledgements

This work is supported by the JST-ANR Japanese-French project VoicePersonae. We would also like to thank Pierre-Michel Bousquet with whom we had fruitful interactions.

References

Ajili, M. (2017). *Reliability of voice comparison for forensic applications*. Ph.D. thesis Avignon Université.

Bahmaninezhad, F., Zhang, C., & Hansen, J. (2018). Convolutional neural network based speaker de-identification. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop* (pp. 255–260). doi:10.21437/Odyssey.2018-36.

Brümmer, N., & De Villiers, E. (2011). The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*, 230 – 275. URL: http://www.sciencedirect.com/science/article/pii/S0885230805000483. doi:https://doi.org/10.1016/j.csl.2005.08.001. Odyssey 2004: The speaker and Language Recognition Workshop.

Brümmer, N., & Preez, J. (2013). The PAV algorithm optimizes binary proper scoring rules.

COMPRISE (2019). Deliverable n^{o}5.1: Data protection and GDPR requirements. https://www.compriseh2020.eu/files/2019/06/D5.1.pdf.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). *Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation*. Technical Report National Inst of Standards and Technology Gaithersburg Md.

Drygajlo, A., Meuwly, D., & Alexander, A. (2003). Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recognition. In *Eighth European Conference on Speech Communication and Technology* (pp. 689–692).

Dunstone, T., & Yager., N. (2009). *Biometric System and Data Analysis: Design, Evaluation, and Data Mining*. (1st ed.). New York, NY: Springer-Verlag.

European Council (2016). 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).

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., & Bonastre, J.-F. (2019). Speaker Anonymization Using X-vector and Neural Waveform Models. In *Proc. 10th ISCA Speech Synthesis Workshop* (pp. 155–160). URL: http://dx.doi.org/10.21437/SSW.2019-28. doi:10.21437/SSW.2019-28.

Hautamäki, R. G., Kanervisto, A., Hautamaki, V., & Kinnunen, T. (2018). Perceptual evaluation of the effectiveness of voice disguise by age modification. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop* (pp. 320–326). URL: http://dx.doi.org/10.21437/Odyssey.2018-45. doi:10.21437/Odyssey.2018-45.

Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation. In *2009 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 529–533).

Justin, T., Štruc, V., Dobrišek, S., Vesnicer, B., Ipšić, I., & Mihelič, F. (2015). Speaker de-identification using diphone recognition and speech synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1–7). volume 04.

Mandasari, M. I. (2017). *Speaker Recognition Systems in Forensic Conditions: The Calibration and Evaluation of the Likelihood Ratio*. Ph.D. thesis Radboud University Nijmegen.

Maouche, M., Srivastava, B. M. L., Vauquier, N., Bellet, A., Tommasi, M., & Vincent, E. (2020). A comparative study of speech anonymization metrics. In *Proc. Interspeech* (pp. 1708–1712).

McAdams, S. (1984). *Spectral Fusion, Spectral Parsing and the Formation of Auditory Images*. Master's thesis Stanford University Stanford, California. URL: https://ccrma.stanford.edu/files/papers/stanm22.pdf.

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. (2019a). The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding. In *Proc. Interspeech 2019* (pp. 3695–3699). doi:10.21437/Interspeech.2019-2647.

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.-F., Raj, B., Trancoso, I., & Busch, C. (2019b). Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, *58*, 441 – 480. doi:10.1016/j.csl.2019.06.001.

Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.-G., Bonastre, J.-F., Todisco, M., & Evans, N. (2020). The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment. In *Proc. Interspeech 2020* (pp. 1698–1702). doi:10.21437/Interspeech.2020-1815.

Nordgaard, A., Ansell, R., Drotz, W., & Jaeger, L. (2011). Scale of conclusions for the value of evidence. *Law, Probability and Risk*, *11*, 1–24. URL: https://doi.org/10.1093/lpr/mgr020. doi:10.1093/lpr/mgr020. arXiv:https://academic.oup.com/lpr/article-pdf/11/1/1/3012102/mgr020.pdf.

Noé, P.-G., Bonastre, J.-F., Matrouf, D., Tomashenko, N., Nautsch, A., & Evans, N. (2020). Speech Pseudonymisation Assessment Using Voice Similarity Matrices. In *Proc. Interspeech 2020* (pp. 1718–1722). URL: http://dx.doi.org/10.21437/Interspeech.2020-2720. doi:10.21437/Interspeech.2020-2720.

Pathak, M., Rane, S., Sun, W., & Raj, B. (2011). Privacy preserving probabilistic inference with hidden markov models. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5868–5871). doi:10.1109/ICASSP.2011.5947696.

Patino, J., Todisco, M., Nautsch, A., & Evans, N. (2020). *Speaker anonymisation using the McAdams coefficient*. Technical Report EURECOM+6190 Eurecom. URL: http://www.eurecom.fr/publication/6190.

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., & Evans, N. (2020). Speaker anonymisation using the McAdams coefficient. arXiv:2011.01130.

Pobar, M., & Ipšić, I. (2014). Online speaker de-identification using voice transformation. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1264–1267). doi:10.1109/MIPRO.2014.6859761.

Ramos, D., Franco-Pedroso, J., Lozano-Diez, A., & Gonzalez-Rodriguez, J. (2018). Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, *20*, 208.

Ramos, D., & Gonzalez-Rodriguez, J. (2013). Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, *230*, 156–169. doi:10.1016/j.forsciint.2013.04.014.

Ramos-Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. Ph.D. thesis Universidad Autónoma de Madrid.

Ramos-Castro, D., & González-Rodríguez, J. (2008). Cross-entropy analysis of the information in forensic speaker recognition. In *Odyssey*.

Schuller, B., & Batliner, A. (2019). Interspeech computational paralinguistics challenge (compare). http://www.compare.openaudio.eu/.

Shannon, C. E. (1949). Communication theory of secrecy systems. *The Bell System Technical Journal*, *28*, 656–715. doi:10.1002/j.1538-7305.1949.tb00928.x.

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

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., & Todisco, M. (2020a). The VoicePrivacy 2020 Challenge evaluation plan. URL: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf.

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., & Todisco, M. (2020b). Introducing the VoicePrivacy Initiative. In *Proc. Interspeech 2020* (pp. 1693–1697). URL: http://dx.doi.org/10.21437/Interspeech.2020-1333. doi:10.21437/Interspeech.2020-1333.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., Maouche, M., & Meunier, C. (2021). The VoicePrivacy 2020 Challenge: Results and findings. Submitted to Special Issue on Voice Privacy in Computer Speech and Language.

Veaux, C., Yamagishi, J., & Macdonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). URL: https://datashare.ed.ac.uk/handle/10283/3443.

Zhang, C. (2012). Acoustic analysis of disguised voices with raised and lowered pitch. In *2012 8th International Symposium on Chinese Spoken Language Processing* (pp. 353–357).

Zhang, S., Gong, Y., & Yu, D. (2019). Encrypted speech recognition using deep polynomial networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5691–5695). doi:10.1109/ICASSP.2019.8683721.