

Preserving privacy in speaker and speech characterisation

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, et al.

▶ **To cite this version:**

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, et al.. Preserving privacy in speaker and speech characterisation. Computer Speech and Language, 2019, 58, pp.441-480. 10.1016/j.csl.2019.06.001 . hal-02307615

HAL Id: hal-02307615
https://hal.science/hal-02307615

Submitted on 7 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 58 (2019) 441–480

www.elsevier.com/locate/csl

Preserving privacy in speaker and speech characterisation ☆

Andreas Nautsch*,[a,f], Abelino Jiménez[b], Amos Treiber[c], Jascha Kolberg[a], Catherine Jasserand[d], Els Kindt[e], Héctor Delgado[f], Massimiliano Todisco[f], Mohamed Amine Hmani[g], Aymen Mtibaa[g], Mohammed Ahmed Abdelraheem[h], Alberto Abad[i], Francisco Teixeira[i], Driss Matrouf[j], Marta Gomez-Barrero[a], Dijana Petrovska-Delacrétaz[g], Gérard Chollet[h,g], Nicholas Evans[f], Thomas Schneider[c], Jean-François Bonastre[j], Bhiksha Raj[k], Isabel Trancoso[i], Christoph Busch[a]

[a] *da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany*
[b] *Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA*
[c] *Cryptography and Privacy Engineering Group (ENCRYPTO), Technische Universität Darmstadt, Germany*
[d] *Security, Technology & e-Privacy Research Group, Transboundary Legal Studies Department, University of Groningen, the Netherlands*
[e] *Centre for IT & IP Law (CITIP), KU Leuven, Belgium*
[f] *Digital Security Department, EURECOM, France*
[g] *Samovar CNRS UMR 5157, Télécom SudParis Université Paris-Saclay, France*
[h] *Intelligent Voice Ltd., London, UK*
[i] *INESC-ID/IST, University of Lisbon, Portugal*
[j] *Laboratoire Informatique d'Avignon (LIA), Université d'Avignon, France*
[k] *Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA*

Received 30 November 2018; received in revised form 31 March 2019; accepted 2 June 2019
Available online 8 June 2019

Abstract

Speech recordings are a rich source of personal, sensitive data that can be used to support a plethora of diverse applications, from health profiling to biometric recognition. It is therefore essential that speech recordings are adequately protected so that they cannot be misused. Such protection, in the form of privacy-preserving technologies, is required to ensure that: (i) the biometric profiles of a given individual (e.g., across different biometric service operators) are *unlinkable*; (ii) leaked, encrypted biometric information is *irreversible*, and that (iii) biometric references are *renewable*. Whereas many privacy-preserving technologies have been developed for other biometric characteristics, very few solutions have been proposed to protect privacy in the case of speech signals. Despite privacy preservation this is now being mandated by recent European and international data protection regulations. With the aim of fostering progress and collaboration between researchers in the speech, biometrics and applied cryptography communities, this survey article provides an introduction to the field, starting with a legal perspective on privacy preservation in the case of *speech data*. It then establishes the requirements for effective privacy preservation, reviews generic cryptography-based solutions, followed by specific techniques that are applicable to *speaker* characterisation (biometric applications) and *speech* characterisation (non-biometric applications). Glancing at non-biometrics, methods are presented to avoid *function creep*, preventing the exploitation of biometric information, e.g., to *single out* an identity in speech-assisted health care via

☆ Recent advances in speaker and language recognition and characterisation.

* Corresponding author at: Digital Security Department, EURECOM, France.
E-mail address: andreas.nautsch@eurecom.fr (A. Nautsch).

https://doi.org/10.1016/j.csl.2019.06.001
0885-2308/ 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license.
(http://creativecommons.org/licenses/by/4.0/)

speaker characterisation. In promoting harmonised research, the article also outlines common, empirical evaluation metrics for the assessment of privacy-preserving technologies for speech data.
© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license.
(http://creativecommons.org/licenses/by/4.0/).

Keywords: Data privacy; Voice biometrics; Standardisation; Cryptography; Legislation

1. Introduction

Many of today's smart devices and services support voice-driven interaction. Speech recordings are increasingly being collected by televisions, smart phones, loudspeakers, watches, intelligent digital assistants, and even smart cars. In all of these cases there is potential for speech recordings to be transmitted over untrusted public networks and then stored and processed on untrusted, third-party cloud-based infrastructures. Stemming from the thrust towards seamless user interaction, rather than being based upon traditional push-to-talk functionality, voice-driven interactive technologies are increasingly *always-listening*. The potential for speech data to be exposed and intercepted by fraudsters or hackers raises grave privacy concerns.

Perhaps the most privacy-sensitive information contained in speech signals is that relating to personal identity, information revealed through automatic speaker and speech characterisation technology. While this article has a focus on both speaker and speech characterisation, the main focus is on the former (voice biometrics). Over the last decade, biometrics technology has revolutionised the traditional approach to authentication of individuals using tokens or passwords. Whereas the latter can be lost, stolen, or easily forgotten, biometrics technology automatically infers the identity from biological and behavioural traits which cannot be lost or forgotten, and are more difficult to steal or pass on. In contrast to passwords, *biometric characteristics* are not renewable: once compromised, e.g., intercepted by fraudsters following a data breach, a *biometric trait* is rendered useless in terms of security (a person cannot replace their fingerprint with a new one), even if cryptographic methods can be used to achieve some level of *renewability*.

The human voice is among the most natural, non-intrusive and convenient of all such characteristics. Progress in automatic speaker characterisation (Kinnunen and Li, 2010; Hansen and Hasan, 2015) has advanced tremendously over the last 20 years. Covering not only biometric recognition in terms of verification (one-to-one comparisons) and identification (one-to-many comparisons), speaker characterisation also concerns the biometric annotation of speech sequences (i.e., the field of *speaker diarization*).[1] Today, speaker characterisation technology is increasingly ubiquitous, being used for authentication of individuals and access control across a broad range of different services and devices, e.g., telephone banking services and smart devices that either contain or provide access to personal or sensitive data. Despite the clear advantages and proliferation of biometrics technology, persisting concerns regarding intrusions into privacy (Prabhakar et al., 2003) have dented public confidence. But not only the *identity of a speaker* is sensitive information, as the uttered content might be of sensitive nature as well. Here, we refer to any other (non-biometric) form of speech (or sound) based processing as *speech characterisation*, since privacy can only be preserved if human dignity and freedom of expression are protected in any form of automatic processing. Speech recognition is one field, but other tasks operating on sounds can be deliberately used to characterise speech, such as in emotion classification, continuous sleepiness detection, diagnosis of diseases, correction of speech production problems in children, and processing of babies' sounds.

Privacy concerns relate to the potential interception and misuse of biometric and non-biometric speech data. All biometric systems involve the storage of biometric references (*biometric enrolment*) and comparisons with new probe samples (*biometric verification*). The interception of such biometric data constitutes an undeniable intrusion to personal privacy, as the information used in speaker characterisation can also be misused by fraudsters for other purposes (Prabhakar et al., 2003). Intrusions into personal privacy are clearly unacceptable and the responsibility to preserve privacy is now enshrined in the recent EU General Data Protection Regulation (European Parliament and Council, 2016a, GDPR). Adequate privacy preservation is therefore essential to ensure that sensitive biometric data, including voice recordings or speech data, are properly protected from misuse.

[1] We refer to *biometric recognition* as defined by ISO/IEC JTC1 SC37 Biometrics (2017b), whereas we define *biometric characterisation* amply: the biometric processing of *characteristics* for any purpose including *biometric recognition, many-to-many comparisons*, and *speaker diarisation*.

While various forms of encryption have been developed to preserve privacy in, e.g., face recognition (Erkin et al., 2009; Sadeghi et al., 2009; Osadchy et al., 2010; Bringer et al., 2014), iris recognition (Blanton and Gasti, 2011; Bringer et al., 2014; Gomez-Barrero et al., 2017a) and fingerprint recognition (Barni et al., 2010; Bianchi et al., 2010; Blanton and Gasti, 2011; Evans et al., 2011), there are very few solutions for speaker characterisation. Furthermore, the existing solutions for other biometric characteristics are not readily transferable to speaker characterisation. Their adaptation to preserve privacy in speech signals is also far from trivial. As argued later in this article, the continued success of speaker characterisation technology, and indeed speech technologies in general, appears to hinge upon the development of reliable and efficient privacy preservation capabilities specifically designed for the automatic processing of speech signals.

This article aims to stimulate research interest in the development of privacy-preserving technologies for the automatic processing of speech signals (in speaker and speech characterisation). Privacy preservation refers to and includes in most cases data protection preservation as well, without explicitly mentioning data protection. In targeting researchers and practitioners in speaker characterisation, biometrics and applied cryptography, it specifically aims to catalyse the cross-fertilisation that will be needed in the coming years to deliver effective, reliable and efficient solutions. As such, this article serves as a general introduction to the field. It is aimed both at speaker characterisation researchers, who may be new to the field of privacy preservation; and also to cryptography researchers, who may be new to the field of speaker characterisation. This article should also be of interest to researchers in the fields of privacy-in-biometrics, though perhaps concerning other biometric characteristics.

In order to motivate the study of privacy-preserving speaker characterisation, Section 2 of the article provides a brief survey of European and international data protection regulations that govern the collection and automatic processing of personal, sensitive biometric information such as voice data. Section 3 then outlines the requirements to preserve privacy in a generic biometric system and presents a review of different privacy-preserving technologies. Section 4 describes the established encryption primitives that have been explored previously as a means of preserving privacy in different biometric systems. The material in these first three sections is agnostic to a particular biometric characteristic.

One of the primary focus of this article is *speaker characterisation* on which Section 5 will expand. It is intended to give a comprehensive but brief overview of the different techniques used in today's state-of-the-art speaker characterisation systems and is aimed at non-expert readers. Further reading on state-of-the-art speaker characterisation technology can be found in other papers within this CSL special issue. It is our intend to highlight the the challenges involved in the preservation of privacy in speech signals and their treatment by automatic speaker characterisation systems to an audience from a wider background. Specific approaches to privacy-preserving speaker characterisation are the focus of Section 6. The goal is to show how the general techniques described in Sections 3 and 4 can be applied to the speaker characterisation systems described in Section 5, rather than to compare different techniques by quantitative means; an undertaking of the latter is far beyond the scope of this article.

Section 7 extends the focus of this article to *speech* characterisation, encompassing other speech-driven applications beyond those with a direct emphasis upon speaker characterisation. Many such applications, e.g., speech recognition and paralinguistics, may not be designed to recognise specific speakers, but may nonetheless present threats to privacy through the potential for so-called *function creep* which could also happen in smart home applications. This threat is credible, for many speech and speaker characterisation technologies operate with similar feature representations; there is little to stop speech characterisation systems performing speaker characterisation as well. Potential function creeps are avoidable at different levels, e.g., by computing upon encrypted speech data from which biometric information is suppressed. The material presented in Section 7 therefore extends that of Section 6 to show how similar privacy preserving technologies may be applied to speech characterisation. Section 8 covers evaluation metrics that can be used to compare the strength or level of privacy preservation offered by competing solutions. A summary of the article is given in Section 9 alongside a discussion of critical directions for future work.

2. Speech data: a legal perspective

Speech recordings are an especially rich source of personal information (Shafran et al., 2003). While speaker characterisation technologies (Kinnunen and Li, 2010; Hansen and Hasan, 2015) can reveal an individual's identity, their use has wider implications, for the same technology could conceivably be used to track a person's whereabouts, their personal habits, and their interactions with personal acquaintances. The degree of personal information captured in speech signals also extends well beyond the notion of identity. The literature shows that speech recordings

can reveal a person's gender with almost certainty (Harb and Chen, 2005) or be used to estimate their age with reasonable accuracy (García et al., 2015; Haderlein et al., 2015; Sadjadi et al., 2016). Moreover, speech recordings may also be used to determine the native language of a person, even if s/he is speaking a foreign language (Abad et al., 2016). Speech recordings can also be used to profile a person's health condition, general well-being (Vilda et al., 2009) or emotional state (Mencattini et al., 2014). This information can be classified as either *paralinguistic* or *extralinguistic. Linguistic* information is related to communication, and to what the speaker intends to relay onto the listener through language; *paralinguistic* information, on the other hand, is transmitted in the form of non-linguistic and non-verbal communication, through which the speaker may convey to the listener an attitude, feeling or emotional state. Finally, non-communicative information such as the speaker's age, gender, personality and health status is categorised as *extralinguistic* (however, this information is often referred to as *paralinguistic*) (Schuller and Batliner, 2013). All of this information is personal, private information that most of us would not willingly entrust to others. It is hardly surprising, then, that the collection, storage, and processing of speech data is now subject to regulatory control. This section presents a legal perspective of the use of speech data for biometric applications, in the European Union (EU) and beyond.

2.1. Concept and status of biometric data

Regulations covering the use of biometric data have been introduced by various countries around the world. The following examines some of the specific regulations in the EU and the US.

2.1.1. At the EU level

Since May 2018, new data protection rules apply in the EU. They are split between a General Data Protection Regulation (GDPR or Regulation 2016/679) (European Parliament and Council, 2016a) and a specific Directive regulating the processing of personal data in the law enforcement context (the *Police* Directive or Directive 2016/680) (European Parliament and Council, 2016b). Both instruments define biometric data as:

> "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data."[2]

Following this definition, biometric data is a type of personal data, representing biometric characteristics and processed to *allow or confirm the unique identification* of an individual. The definition does not refer to speech data (whether audio recordings, speech patterns or speech representations, such as templates and models), as it only mentions facial images and fingerprints as examples of this type of data. But it can be assumed that those forms are covered by the regulatory definition of biometric data, as long as they result from specific technical processing (that allow or confirm the individual's unique identification).

The new regulatory framework is recent, and the legal literature on the topic of the processing of biometric data is still scarce. However, it focuses in part on the meaning of *unique identification* and its consequences on the classification of biometric data as sensitive data. Scholars have different interpretations: some consider that *unique identification* could be a reference to the use of the biometric identification modality (Kindt, 2018), whereas others argue that *unique identification* might be understood as a threshold of identification from a data protection perspective (Jasserand, 2016). Following the latter interpretation, biometric data processed for either verification purposes or identification purposes are classified as sensitive data, whereas following the first interpretation only biometric data processed for identification purposes may have to be regarded as sensitive data. These discussions are mainly linked to a discrepancy of wording between, on the one hand, Recital 51 GDPR, which specifies the conditions under which photographs fall within the category of biometric data, i.e., *only when processed through a specific technical means allowing the unique identification or authentication of an individual*, and, on the other hand, Article 4(14) GDPR that defines the notion of biometric data making a reference to *allow[ing] or confirm[ing] the unique identification*.[3]

[2] As defined in Art. 4(14) GDPR and Article 3(13) Directive 2016/680.

[3] In the recital 51 *unique identification* further seems to be used as a synonym of (biometric) identification as it is opposed to *authentication*. But the wording and this interpretation is not consistent with Article 4(14) GDPR where the functions are described as *allow[ing] or confirm[ing] the unique identification*.

Besides, Recital 51 GDPR raises other questions, in particular, whether only photographs of faces *should not systematically be considered* as biometric data if not processed through specific technical biometric processes, or also other *images*, including for example fingerprint images. This is important as such data collections would not be considered biometric data collections. If it is argued that *unique identification* should not be in the context of a technical biometric comparison but approached as a threshold of identification from a data protection perspective, it is clear that all categories of biometric data processing, whether for identification or verification, would fall under the general prohibition of Article 9(1) GDPR.

According to Article 9(1) GDPR,[4] only biometric data *processed to uniquely identify* an individual are sensitive data. As a consequence, data classified as sensitive cannot be processed, unless an exception applies.[5] Article 9(2) GDPR provides a list of ten exceptions, which are restrictive in their application.[6] Should *uniquely identifying* be understood as only referring to the process of biometric identification, biometric data processed for authentication or verification purposes would then be considered as *ordinary* personal data, for which the general legal grounds for processing could be invoked. Following Article 6 GDPR, these legal grounds include the individuals' consent but also the legitimate interests of the controller. The ambiguity of the phrase *for purposes of uniquely identifying* and hence the precise scope of biometric voice data as a special category is also present in the police Directive 2016/680. This is not a mere theoretical discussion, as in this case Law Enforcement Authorities (LEAs) shall restrict such processing unless it is *strictly necessary*, subject to appropriate safeguards and authorised by law (safe two other grounds).[7]

As the GDPR has replaced the Data Protection Directive (Directive 95/46/EC), which is the basis of many other European legislations, the implications of the GDPR are far reaching. In this paragraph, an example is elaborated which concerns the use of biometric voice characterisation to *uniquely identify* a person for the purposes of securing financial transactions (e.g., online banking and mobile payment solutions). The unclarity surrounding the meaning of *uniquely identifying* and consequently the legal basis applicable to the processing of voice data is also important in view of the second Payment Services Directive (EU) 2015/2366 (PSDII) which has introduced the factor of *something the user is* for strong customer authentication (Kindt, 2019). Voice data as authentication element is on the rise, while the legal framework proposed by the GDPR does not offer clarity to the financial service sector on the legal basis required for such processing. Due to the varying interpretations (see above) further clarification by the legislator is highly relevant. This discussion about the exact scope of biometric data processing falling in the special category and the legal grounds on which such processing can be based is somewhat less pressing for scientific research. Such research could be based not only on (explicit) consent, but also on the *research exception* that allows the processing of special categories of personal data for research purposes.[8]

Biometric data could still fall in the category of *sensitive data* independently of the purpose for which they are processed. In particular, if the data reveal *sensitive information*, defined as *racial or ethnic origin, political opinions, religious or philosophical beliefs* as well as *data concerning health*, their processing should be regarded as sensitive.[9] Speech data revealing this type of sensitive information could thus be considered sensitive data. In addition, they could also be classified as sensitive data due to the national regime applicable to their processing. Following Article 9(4) GDPR, member states are in addition allowed to introduce further conditions for the processing of biometric data, whether or not they are processed to uniquely identify an individual.

Crucial for voice data processing is the obligation of the controllers and processors to implement technical and organisational measures to ensure a level of security appropriate to the risks, including as appropriate, by encryption, by ensuring the confidentiality, integrity and availability and resilience, and by assessing and evaluating the effectiveness of the security measures.[10] Finally, it is important to retain that the GDPR requires, when voice data are processed *on a large scale* (for example when large financial institutions deploy user verification by call centres, or when any other personal data processing activity is *likely to result in high risks*) that an assessment of the risks and

[4] And Article 10 of Directive 2016/680.

[5] By contrast, in a law enforcement context, the processing of sensitive data is not prohibited but subject to strict conditions, see Article 10 of Directive 2016/680.

[6] Such as explicit consent, substantial public interests or research purposes.

[7] Article 10 Directive 2016/680, to be implemented in national laws.

[8] Article 9.2(a) and Article 9.2(j) GDPR. The *research exception* requires, however, that Article 89 GDPR is complied with and safeguards are in place, especially for reaching data minimisation.

[9] Article 9(1) GDPR and Article 10 of Directive 2016/680.

[10] Article 32 GDPR.

safeguards (Data Protection Impact Assessment, DPIA) is made.[11] The privacy and data protection preservation techniques and methods described in this contribution are therefore also highly relevant for such DPIA, whether under the GDPR or under the Police Directive 2016/680. The latter also requires the implementation of security, privacy, and data protection by design and by default measures.

2.1.2. In the US

Other countries have adopted specific laws either regulating data privacy (and including biometric data) or privacy rules for the processing of biometric data alone. In the United States, for instance, three States (Illinois, Texas, and Washington) have adopted specific acts or provisions to regulate the collection and use of biometric information or identifiers for commercial purposes. The notion of *biometric data* is not defined. Instead, the different acts and provisions refer to *biometric identifier* and *biometric information*.

In the laws of Illinois and Texas, a biometric identifier is restrictively defined as being limited to *retina or iris scan, fingerprint, voiceprint, or scan of hand or face geometry.*[12] By contrast, in the Washington law, the definition of *biometric identifier* provides examples of those identifiers and include *voiceprint*.[13]

As for biometric information, it is defined in the Illinois law as *information [...] based on an individual's biometric identifier used to identify an individual.* This notion is encapsulated in the definition of *biometric identifier* in the Washington law, which is defined as *data generated by automatic measurements of an individual's biological characteristics, such as a fingerprint, voiceprint, eye retinas, irises, or other unique biological patterns or characteristics that is used to identify a specific individual.*

Besides these three sectorial state laws, California has adopted a general data privacy law designed to protect consumers (California Consumer Privacy Act). In the definition section of the act, *biometric information* is defined as a very comprehensive notion that includes not only physical characteristics but also behavioural ones. As for the examples, they expressly refer to voice recordings. The definition also explains that an *identifier* can be extracted from *biometric information* and provides examples of these identifiers (including a voiceprint).[14] Biometric identifier is classified as an example of personal information.[15] Thus, a voiceprint is considered as personal information.

2.2. Privacy by Design

Privacy by Design is a policy concept that was developed in the 1990s. In 2012, it appeared under the form of the obligation of *Data Protection by Design and by Default* in the EU legislative proposals for the GDPR and the *Police Directive*.

2.2.1. Origin of the concept

It is difficult to pinpoint the exact origin of the concept. It has been linked to PETs (Privacy-Enhancing Technologies), to the concept of *Value Sensitive Design* in ethics, as well as to Laurence Lessig's concept of *code is law* (Klitou, 2014; Lessig, 2006). In the 1990s, the concept was popularised by Ann Cavoukian, the former Information and Privacy Commissioner of Ontario. She developed a policy concept around seven foundational principles, including *Privacy by Default*.[16] The idea behind *Privacy by Design* is to incorporate privacy into the design and development of products, services, and systems. Endorsed by the International Conference of Data Protection and Privacy Commissioners in 2010,[17] her approach has been harshly criticised for its lack of implementability (Rubinstein and

[11] Article 35 GDPR. Note that Member States generally listed biometric data processing as requiring such DPIA.

[12] Section 10 Illinois Biometric Information Act, 740 ILCS 14, with a very similar definition in the Texas Business and Code of Commerce – BUS&COM 503.001. Capture or Use of Identifier (a).

[13] Washington, Chapter on Biometric Identifier, RCW 19.375.010 (1).

[14] California Consumer Privacy Act, 1798.140 (b).

[15] California Consumer Privacy Act, 1798.140 (o)(1).

[16] The seven foundational principles are (1) Proactive not Reactive, Preventative not Remedial, (2) Privacy as the Default Setting, (3) Privacy Embedded into Design, (4) Full Functionality-Positive-Sum, not Zero-Sum, (5) End-to-End Security—Full Lifecycle Protection, (6) Visibility and Transparency-Keep it Open, and (7) Respect for User Privacy: Keep it User-Centric.

[17] 32nd International Conference of Data Protection and Privacy Commissioners: *Resolution on Privacy by Design*, Jerusalem, Israel, 27–29 Oct. 2010. https://edps.europa.eu/sites/edp/files/publication/10-10-27_jerusalem_resolutionon_privacybydesign_en.pdf.

Good, 2013). Several computer scientists proposed their own engineered approach to the concept (Spiekermann and Crannor, 2009; Gürses et al., 2011; Hoepman, 2013).

2.2.2. Data protection by design and by default

Before the adoption of the new data protection framework, the Data Protection Directive 95/46 already contained the premise of the concept in Article 17 on the security of processing. That provision imposed the implementation of *technical and organisation measures* for security purposes (such as to prevent accidental and unlawful destruction, loss, alteration, transmission or access to the data). In the GDPR, the obligation which is imposed on data controllers to adopt *technical and organisational measures* goes beyond security measures. As a general rule, controllers shall protect individuals' rights and freedoms in relation to the processing of their personal data.[18] Article 25 GDPR furthermore describes the obligations of Data Protection *by Design* and *by Default*.

Following Article 25 (1) GDPR, data controllers must adopt *technical and organisational measures* to implement data protection principles, such as data minimisation. The only example of measures that can be found in the provision is the *pseudonymisation* of data. Recital 78 GDPR refers to data transparency or data security as examples of these measures. The provision does not specify how to implement the obligation, but lists factors that should be taken into account: the state of the art, the cost of implementation, as well as the nature, scope, context, purposes of the processing, and the risks to individuals' rights. Article 25(2) GDPR lays down the obligation of data protection by default. Contrary to the policy concept of Privacy by Design, data protection by default is conceived as a separate obligation from data protection by design. The requirement of data protection by default focuses on the implementation of the principle of data minimisation, limiting the collection of personal data to what is *strictly necessary*.

Some authors suggest that the obligations of *data protection by design and by default* find their roots, or at least are inspired, by the *Privacy by Design* concept, developed by Cavoukian (Costa and Poullet, 2012). However, there are major differences between the two. *Privacy by Design* is a policy principle applicable to all the actors involved in the processing of personal data, whereas data protection by design and by default is a legal obligation limited to data controllers. According to Recital 78 GDPR, producers of products, services, and applications are only *encouraged* to take into account the principle of data protection when they design or develop their products, services, and applications. This encouragement does not constitute an obligation. However, to sell their products, producers and manufacturers might be obliged by contract to comply with the obligations of data protection by design and by default.

2.3. Discussion

One of the issues that emerges from this Section is the reference to *voiceprints* as an example of biometric identifiers in legal texts. The reference to the term is unfortunate as scientists challenge its accuracy and meaning. In 2000, Boë argued that voiceprint is an *erroneous metaphoric terminology [that] leads many people (not only the general public) to believe that a graphical representation of the voice (the sonogram, as it happens) is just as reliable as the structure of the papillary ridges of the fingertips, or genetic fingerprints, and that it allows reliable identification of the original speaker* (Boë, 2000) (referring to Bimbot and Chollet, 1997). The reference to such a term wrongly gives the impression that voice data can be graphically represented, which is not the case. Further research on the discrepancy of terms used by both the legal and technical communities is necessary.

Building on the legal provisions described in this section, the paper suggests technical solutions in Sections 6 and 7 to preserve the individuals' right to privacy and data protection when their speech data are processed for biometric and non-biometric characterisation applications. Before describing solutions, we provide a general overview on preserving privacy in biometrics, covering other biometric modalities than voice as well as standardisation on biometric information protection.

3. Biometric privacy preservation at a glance

Conventional biometric systems involve separate *enrolment* and *verification* phases. During enrolment, biometric references B_R are captured from a sensor before features and their representations (as templates or models) T_R are

[18] For instance: to protect the right to data protection, the right to privacy, and the freedom of information.

extracted and then stored in a database (DB). During verification, a biometric probe B_P is captured from the same sensor, giving features T_P. Subsequently, the comparison subsystem compares probe T_P with the reference T_R that corresponds to the claimed identity ID_R, thereby giving a dis/similarity score S. The score is then compared to a threshold η, in order to determine an accept or reject decision. In contrast to the one-to-one comparison performed in verification scenarios, *identification* scenarios involve comparisons between the probe and the full set of references stored in the database.

The storing of biometric references (as templates or models) without protection is undesirable since this practice raises grave privacy concerns. Research has shown that it is possible to recover original biometric data from templates and models. Examples of such work include that in fingerprint (Cappelli et al., 2007), handshape (Gomez-Barrero et al., 2014), face (Adler, 2003), and iris (Galbally et al., 2013) recognition (using templates). While speaker models cannot be used to recover estimates of the exact speech data used for their training (as with templates), they can be used to generate speech that is representative in some sense of the speaker, e.g., through model-based voice conversion (Toda et al., 2007). The potential for privacy intrusions stems from the ability to use such information for purposes other than those originally intended, or to create presentation attack instruments that can be used to impersonate bona fide or enrolled subjects. Database controllers can also use such information to link the biometric information collected from the same subjects across different databases. This would enable the profiling and tracking of individuals.

Privacy-preserving techniques have been devised in order to mitigate privacy concerns. These typically involve some form of encryption as part of appropriate security measures and safeguards, usually applied to protect the information stored in biometric databases. The quest to preserve privacy, among other factors, has led to the development of three general requirements (ISO/IEC JTC1 SC27 Security Techniques, 2011), namely:

- *Unlinkability:* Given only protected biometric information, it is not possible to say whether two protected biometric sample representations belong to the same subject. This prevents cross-comparisons for databases of different applications and ensures the privacy of the subject.
- *Renewability:* If a protected biometric reference is leaked or lost, the reference data can be revoked and renewed from the same biometric trait without the need to re-enroll.
- *Irreversibility:* Recovering biometric data from leaked protected biometric information is impossible without knowing the secret used to protect the biometric information. Restoring of valid biometric features or samples is prevented.

The techniques to meet these requirements generally refer to some form of biometric information protection. Since other elements of the biometric system may also require protection, the relevant standard concerning privacy preservation refers instead to *Biometric Information Protection* (BIP). The ISO/IEC 24745 (ISO/IEC JTC1 SC27 Security Techniques, 2011) standard advocates the adoption of the BIP scheme depicted in Fig. 1 in order to fulfill the requirements of unlinkability, renewability, and irreversibility. In contrast to conventional systems, the feature extraction stage, used during enrolment, is followed by a Pseudonymous Identifier Encoder (PIE). The PIE generates Auxiliary Data (AD) and a Pseudonymous Identifier (PI) for the Reference (R). The corresponding data AD_R and PI_R are then stored in separate databases. PI_R, in the form of a protected template or model, contains no information relating to the subject's ID. Furthermore, as it is protected, it cannot be used to restore the original biometric data. The subject's ID-related data is instead contained within AD_R and is stored in a separate database.

During verification, features extracted from the probe (P) serve as input to a corresponding Pseudonymous Identifier Recorder (PIR). The PIR may use AD_R in order to generate the Pseudonymous Identifier PI_P. In the next step, the Pseudonymous Identifier Comparator (PIC), a component of the comparison subsystem, compares the reference and probe identifiers, PI_R and PI_P, respectively, in order to produce the score S. The remaining threshold comparison and decision steps are then identical to those of the conventional, unprotected system; the dis/similarity score S is compared to threshold η and, depending on this result, the subject's claim is either accepted or rejected.

The leakage of protected biometric information does not provide any information about the ID of the subject itself. Templates and models cannot be used to reconstruct or estimate original biometric features or samples; protected information is *irreversible*. Protection also means that it is possible to create a new PI_R without requiring a subject to re-enroll; protected templates and models are *renewable*. Accordingly, since successive enrolments using the same biometric information will produce different PI_R, they are also *unlinkable* across different databases.

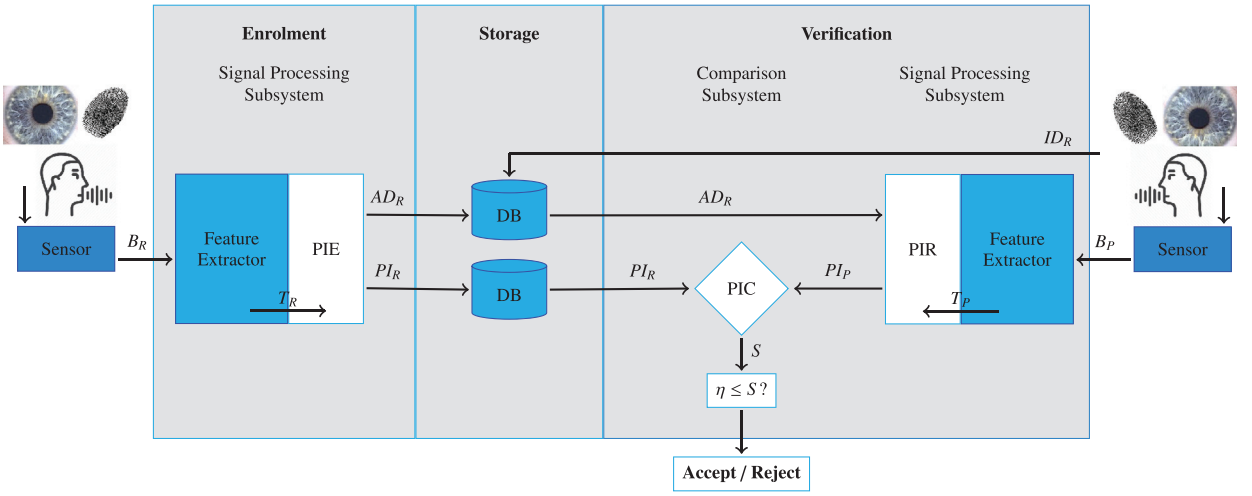


Fig. 1. Enrolment (left) and verification (right) processes in a biometric information protection system based on ISO/IEC 24745 (ISO/IEC JTC1 SC27 Security Techniques, 2011).

While not always fully achievable, BIP should incur no degradation to recognition accuracy. In order to support real time verification, BIP should not significantly increase the computation time, or at least keep it manageable. With many such BIP schemes having been proposed in recent years, comparing approaches in a meaningful way has become critical to progress. The work in Simoens et al. (2012) proposes suitable techniques to benchmark different BIP solutions.

The following identifies a number of general cryptographic techniques that may be applied to preserve privacy in biometric information. The emphasis here is high level. A more detailed discussion of specific techniques is presented in Section 4. A treatment of their application in speaker and speech characterisation is the subject of Sections 6 and 7. Current approaches to BIP can be broadly classified into three categories (Rathgeb and Uhl, 2011), namely: (i) cancellable biometrics (Patel et al., 2015), where irreversible transformations are applied on the sample or template level; (ii) cryptobiometric systems (Campisi, 2013), where a key is either bound or extracted from the biometric data; and (iii) biometrics in the encrypted domain (Aguilar-Melchor et al., 2013), where techniques based on homomorphic encryption (HE) and secure two-party computation (STPC) are used to protect biometric data. Whereas cancellable biometrics and cryptobiometric systems usually incur some accuracy degradation (Rathgeb and Uhl, 2011), the use of biometrics in the encrypted domain prevents such loss, since operations performed in the encrypted domain are equivalent to those performed with plaintext data. Additionally, biometrics in the encrypted domain can be provably secure for biometric template protection (Cavoukian and Stoianov, 2011).

A number of homomorphic encryption solutions for biometric information protection are based on *ideal lattices* (Stehlé et al., 2009), which are assumed to be *post-quantum secure* (Bernstein et al., 2009). The work of Yasuda et al. (2013, 2015) proposes an approach that uses binary feature vectors with a constant size of 2048 bits for any biometric characteristic to compute the Hamming distance between reference and probe samples in the encrypted domain. An alternative approach proposed in Patsakis et al. (2015) encrypts 2048-bit iris codes with NTRU encryption (Hoffstein et al., 1998) and computes the distance in the encrypted domain while never disclosing biometric data.

STPC has been applied successfully in a number of privacy preservation studies involving various biometric characteristics. Most of this literature relates to biometric identification, where client feature vectors are compared to a database of biometric feature vectors stored on a remote server. In a similar way to the classical STPC setting, the database is stored on the server in plaintext format. Critically, though, the computation does not leak any information about the client feature vectors to the server, whereas the server database is hidden from the client. These solutions have been explored in face (Erkin et al., 2009; Sadeghi et al., 2009; Osadchy et al., 2010; Bringer et al., 2014), iris (Blanton and Gasti, 2011; Bringer et al., 2014), and fingerprint (Barni et al., 2010; Bianchi et al., 2010; Blanton and Gasti, 2011; Evans et al., 2011) identification scenarios. In contrast to the voice characteristic, the application of STPC for these biometric characteristics has attracted considerable attention, since it relies upon computations that

can be efficiently computed by STPC, e.g., iris identification consists mostly of XOR gates that are inexpensive in STPC.

However, in order to achieve biometric information protection, the server is required to perform its computation without knowledge of the protocol input, i.e., the reference data. This is only achieved in some STPC-based works, which in Osadchy et al. (2010) and Chun et al. (2014) involves hybrid approaches which combine HE with STPC, e.g., by using HE for encryption while performing computations with HE and/or STPC. The approach proposed in Blanton and Aliasgari (2012) uses STPC to protect iris templates in a biometric identification architecture. A pair of servers that are assumed to not collude securely compute the identification by storing the templates using a cryptographic primitive called secret sharing, thereby achieving unlinkability. A different line of research (Wang et al., 2015; Hu et al., 2018) uses special-purpose encryption schemes to allow specific computations over the encrypted templates. In more general terms, the secure and efficient training of neural networks (Mohassel and Zhang, 2017) and classification using neural networks (Sadeghi and Schneider, 2008; Barni et al., 2011; Liu et al., 2017; Riazi et al., 2018; Juvekar et al., 2018; Riazi et al., 2019) have also been in the focus of the STPC community, enabling the application of neural network techniques without revealing client input data and without revealing server templates or models.

For the sake of improving efficiency in identification, Adjedj et al. (2009) applied searchable symmetric encryption (Curtmola et al., 2006) to iris features, which are transformed via Locality Sensitive Hashing (LSH). Functional encryption schemes with function-hiding property that compute the inner product of two encrypted vectors can provide secure authentication for biometric data as outlined by Kim et al. (2018), where an inner product between two bipolar vectors is used to compute the Hamming distance between their corresponding binary vectors.

The most efficient BIP schemes operate on binarised data, e.g., in iris, fingerprint, and face recognition systems where sample quality/precision is typically high. In these cases, binary representations can be reasonably consistent across multiple acquisitions. Operation upon low-quality samples can be avoided by issuing probe re-captures so that high-quality samples are acquired. The same approach is not applicable in the case of speech, however, which typically exhibits considerable intersession nuisance variation, e.g., phonetic content, vocal effort, and both additive and convolutive noise. As a result, features extracted from speech signals are rarely as consistent as they are for some other biometrics. Conventionally, speaker characterisation systems employ some form of *model*, rather than *templates*; models accommodate data uncertainty in the manner of varying signal contents and qualities. Thus, in the case of speaker characterisation, instead of protecting *biometric templates* (i.e., point estimates), privacy-preserving techniques must be applied to *biometric models* and be robust to inherent sample variation. One approach is to combine model binarisation techniques with variability compensation methods to provide the required robustness; other approaches may preserve principled uncertainty propagation for the purpose of inferring identities robustly. In speaker characterisation, BIP extends the perspective of operating on *templates* as feature representations already comprising highly *precise information* for a characterisation on observed data to *models*, which analytically characterise the rather *uncertain information* (not on observed but) on inferred data. Eventually, privacy-preserving safeguards need to be able to help the acceptance of voice in biometrics applications by the citizen, if the underlined applications are reliable. For example, the requirements of ISO/IEC 17025 (ISO/CASCO Committee on Conformity Assessment, 2017), a framework of implementing reliability assessment in biometric applications (motivated by Meuwly et al., 2017), need to be satisfied. In summary, the application of BIP to speech data can be particularly challenging, when implementing security measures for biometric information in terms of templates or models. Existing techniques and strategies are outlined in the following Section.

4. Overview on cryptographic approaches

Since the groundbreaking introduction of *public-key* cryptography in the 1970s, modern cryptography has grown to encompass more than just the classical goal of ensuring confidentiality and authenticity of messages and has become highly relevant to everyday users. Sensitive data is increasingly outsourced to service providers and thus encryption has to be used in applications. However, since encryption hides any information concerning the underlying data, the service provider can no longer operate on the data without knowing the secret encryption key. As a result, additional primitives of advanced cryptography that still allow the secure computation of such services have moved into the focus of the cryptographic research community and received increasing interest by industrial players.

In this section, we give an overview of various secure computation primitives that have been developed by the cryptographic community and have served as the basis of the privacy-preserving architectures surveyed in this article. We refer to Katz and Lindell (2014) for an introduction of basic cryptographic notions. Provable security, achieved by rigorous security proofs as summarised in Lindell (2017), are the de facto standard in modern cryptography and should also be used for privacy-preserving solutions. Most of the protocols surveyed in the following have such proofs.

4.1. Homomorphic encryption (HE)

Homomorphic encryption (HE) has become a popular tool for allowing computations on encrypted outsourced data. In general, the structure of the plaintext space is preserved in the ciphertext space for additions and/or multiplications of plaintext data under encryption. Therefore, HE enables operations on the encrypted data without requiring any decryptions. For instance, in the case of an *additively homomorphic encryption* scheme, with encryption function *enc*, the addition operation + can be preserved in the cipherspace under the *public key pk* and another operation, e.g., the multiplication: $enc_{pk}(x) \cdot enc_{pk}(y) = enc_{pk}(x + y)$. It follows that the multiplication with a *constant a* can also be computed under encryption using exponentiation: $enc_{pk}(x)^a = enc_{pk}(ax)$. However, multiplication between two ciphertexts is not preserved.

We differentiate between three types of HE: *(i) partially homomorphic encryption (PHE), (ii) somewhat homomorphic encryption (SHE)*, and *(iii) fully homomorphic encryption (FHE)*. The latter, FHE, allows unlimited additions and multiplications at the cost of an increased computational load (Gentry, 2009), while SHE schemes have a fixed limit of multiplications to speed up their execution. PHE schemes support either additions or multiplications, hence, they are only partially homomorphic.

The problem for SHE schemes is that the resulting ciphertext cannot be decrypted when the limit of multiplications is exceeded. Furthermore, there are SHE schemes that cannot correctly decrypt when different operations are combined, i.e., only additions or multiplications are possible but combining both operations in the encrypted domain cannot be handled. Popular HE schemes include the *Paillier* (Paillier, 1999; Paillier and Pointcheval, 1999; Bellare et al., 1998; Damgård and Jurik, 2001), *ElGamal* (ElGamal, 1984), and *NTRU* (Hoffstein et al., 1998) cryptosystems. Efficient implementations of various HE schemes like Python-Paillier,[19] HElib,[20] the NTRU Open Source Project,[21] or SEAL[22] have been widely used to perform computation on encrypted data. The drawback of these schemes, usually, is a relatively high computational overhead and the relatively large ciphertexts and keys. HE schemes can be used to preserve privacy in speaker and speech characterisation, see Sections 6.1 and 7.1.

4.2. Secure two-party computation (STPC)

Apart from HE, there are other methods that allow secure computation without revealing the plaintext input or any intermediate information. Generally, the notion of STPC allows two parties to compute any computable function $f(x, y)$ on input x provided by any party and input y provided by the other party, without revealing anything else about one party's input to either. A comparative overview of frameworks that implement STPC protocols can be found in Hastings et al. (2019). Additionally, the MATRIX framework[23] allows to deploy and compare the performance of different STPC frameworks (Barak et al., 2018). STPC protocols can be used to preserve privacy in speaker characterisation, see Section 6.2.

In STPC, the to-be-computed function f is usually represented as a circuit, i.e., a directed acyclic graph where the edges represent intermediate *wires*, and the vertices indicate input and output wires, as well as *gates* that compute a basic function. For instance, any function can be represented as a Boolean circuit consisting of only XOR and AND gates. For the secure evaluation of f, parties rely on securely computing every gate in the function's circuit. The various STPC protocols supply a way to securely compute those basic gate functionalities and ensure confidentiality of

[19] https://github.com/n1analytics/python-paillier
[20] https://github.com/shaih/HElib
[21] https://github.com/NTRUOpenSourceProject/ntru-crypto
[22] https://www.microsoft.com/en-us/research/project/simple-encrypted-arithmetic-library/
[23] https://github.com/cryptobiu/MATRIX

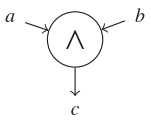
inputs and intermediate values for a composition of gates (i.e., a circuit). For protocols operating on Boolean inputs, any computation is possible on values of any input space with a binary representation, e.g., even floating point operations are possible by using appropriate circuit building blocks, though this incurs a higher overhead.

Compared to HE, STPC (mostly) relies on relatively cheap symmetric cryptographic operations (e.g., AES) and, as such, has the potential to be a more practical privacy-preserving solution. It has become apparent that for STPC communication is the bottleneck: rounds of interaction or communication volume mostly determine the execution time, whereas HE relies on computationally rather expensive operations. The main difference to HE is that STPC requires *interaction* per gate between the parties, while HE does this only for the inputs and outputs. In an effort to overcome the limitations of PHE and SHE schemes, STPC has received great attention from the cryptographic community in recent years, which resulted in efficient and practical schemes (Malkhi et al., 2004; Kolesnikov and Schneider, 2008; Bellare et al., 2013; Zahur et al., 2015). Further efficiency improvements stem from the ability to split the circuit evaluation into an offline and an online phase, where the offline phase is used for input-independent pre-computation that is required for the input-dependent online phase. It follows that for the execution time itself, only the online phase is relevant. Efficiency can also highly depend on the specific STPC protocol and the desired functionality (Demmler et al., 2015b). Protocols exist with security against *semi-honest* and *malicious* participants. Semi-honest parties follow the protocol specification honestly, whereas malicious parties can actively deviate from the protocol. Transformations of semi-honest protocols into maliciously secure versions (e.g., Lindell and Pinkas, 2012) usually incur some overhead.

Yao's garbled circuits (GCs) protocol (Yao, 1982), introduced by Andrew Yao in the 1980s as a first solution for STPC, has a constant number of rounds. The protocol has been subject to many optimisations and has been used in various applications. GCs operate on binary inputs and compute a functionality by evaluating its *garbled* Boolean circuit representation.

In the protocol, the first party, also called the garbler, creates random labels $k_0^w \in_R \{0,1\}^\kappa$ and $k_1^w \in_R \{0,1\}^\kappa$ for each wire w in the circuit, where κ is the security parameter (e.g., $\kappa = 128$ in practice). For every gate g in the circuit, the party generates a *garbled* gate $\widetilde{g}$ with the property that, given $\widetilde{g}$ and the labels corresponding to the values of the input wires of g, the other party (the evaluator) can compute the label corresponding to the correct value of the output wire of g. The construction makes use of a symmetric encryption scheme with an encryption function *enc* and a decryption function *dec*. Fig. 2 outlines how the plaintext AND gate on the left is garbled: for every possible input bit combination, the label corresponding to the output value is encrypted using the labels corresponding to the respective input bits. From this example, it is easy to see how to garble any two-input gate g: for all possible input bit values a and b, the garbled entry is $enc_{k_a^0}\left(enc_{k_b^1}\left(k^2_{g(a,b)}\right)\right)$. Given $\widetilde{g}$, k_a^0, and k_b^1, one can evaluate the garbled gate by trying to decrypt each entry. Assuming a symmetric encryption scheme with an elusive and efficiently verifiable range (Lindell and Pinkas, 2009), only the entry corresponding to a and b can be decrypted to obtain the result $k^2_{g(a,b)}$.

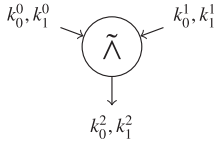
After garbling each gate in the circuit, the garbler randomly permutes the entries of each garbled gate and sends the resulting garbled circuit to the evaluator. A final task is that, in order to enable the evaluation of the garbled circuit, the garbler has to transfer the labels corresponding to the input bits of both parties without revealing one party's input to the other. Since the labels are chosen uniformly at random, just sending the labels corresponding to the garbler's inputs to the evaluator reveals nothing about the actual input. However, sending the labels for the input of the evaluator is problematic, as the garbler must not know which labels to send but also cannot send both labels for each bit because then the evaluator could learn information about the garbler's input by evaluating the circuit on other inputs than its own. The solution to this problem is a cryptographic primitive called *Oblivious Transfer* (OT). Such an OT allows a sending party with the input (x_0, x_1) to send the value x_b to a receiving party with input



(a) AND gate.

$\wedge$	$b=0$	$b=1$
$a=0$	0	0
$a=1$	0	1

(b) Truth table of AND gate.



(c) Garbled AND gate.

$\tilde{\wedge}$	$b=0$	$b=1$
$a=0$	$enc_{k_{\mathbf{0}}^0}\left(enc_{k_{\mathbf{0}}^1}\left(k^2_{\mathbf{0}}\right)\right)$	$enc_{k_{\mathbf{0}}^0}\left(enc_{k_{\mathbf{1}}^1}\left(k^2_{\mathbf{0}}\right)\right)$
$a=1$	$enc_{k_{\mathbf{1}}^0}\left(enc_{k_{\mathbf{0}}^1}\left(k^2_{\mathbf{0}}\right)\right)$	$enc_{k_{\mathbf{1}}^0}\left(enc_{k_{\mathbf{1}}^1}\left(k^2_{\mathbf{1}}\right)\right)$

(d) Truth table of garbled AND gate.

Fig. 2. Computation of a garbled AND gate $\widetilde{\wedge}$.

bit $b \in \{0, 1\}$ without learning anything about b and without revealing anything about x_{1-b} to the receiver. In order to send the evaluator its corresponding labels, both parties engage in an OT protocol for every input bit y_j of the evaluator's input y, with the garbler acting as sender with input (k_0^j, k_1^j) and the evaluator acting as receiver with input y_j. OT can be efficiently instantiated based on mostly symmetric cryptography using an optimisation called OT extension (Ishai et al., 2003; Asharov et al., 2013).

To summarise, Yao's GC protocol consists of the following steps:

1. The garbler transforms f into a Boolean circuit consisting of XOR and AND gates. It generates labels for all wires and garbled gates for every gate in the circuit.
2. The garbler sends the permuted garbled gates and the labels corresponding to its own input bits to the evaluator. Then, both parties engage in OTs, where the evaluator obliviously receives the labels corresponding to its input bits.
3. The evaluator evaluates each gate in the garbled circuit using the labels obtained in step 2.
4. To reveal the output, the garbler can reveal the plaintext bits corresponding to the circuit output labels obtained in step 3.

There have been numerous improvements to Yao's GC. The "point-and-permute" optimisation (Malkhi et al., 2004) just requires one decryption per gate, the "free-XOR" technique (Kolesnikov and Schneider, 2008) allows to the computation of XOR gates without communication and negligible computation; and the "half-gates" construction (Zahur et al., 2015) yields AND gates that require only 2κ bits of communication. Yao's GC has been used to preserve privacy in speaker characterisation, see Section 6.2.

4.3. Search on encrypted data

Apart from the privacy and security concerns, outsourcing data to the cloud has some advantages to both individuals and companies such as huge storage capacity and robust backup services. The privacy and security concerns can be addressed by encrypting all the user files using symmetric-key encryption algorithms where only the client (data owner) knows the encryption keys. However, whenever there is a need to search for a specific uploaded file, the client has no choice other than downloading all the uploaded files, since the cloud does not know the client's key and thus, cannot decrypt and search on the encrypted files. Methods used for searching on encrypted data often involve hashing techniques, which have been used—often relying on some form of *binarisation* of already discrete (biometric) data—for symmetric and asymmetric protocols in speaker and speech characterisation, see Sections 6.3, 6.4, and 7.2.

There are several cryptographic methods to search on encrypted data such as fully homomorphic encryption (Gentry, 2009) and oblivious RAMs (Pinkas and Reinman, 2010) but they are not yet practical. One of the practical methods to search on encrypted data is Searchable Symmetric Encryption (SSE). It was firstly proposed by Song et al. (2000) and later improved by Curtmola et al. (2006). Based on the improved security model, many SSE schemes were proposed such as Kamara et al. (2012) and Cash et al. (2013, 2014).

An SSE scheme allows only a single user to encrypt documents and make them searchable without decrypting them. It enables an untrusted server, which is semi-honest, to search on a database consisting of encrypted documents without learning the client's secret keys. This is done via the use of trapdoors, which are generated by applying a deterministic encryption algorithm or a keyed-hash function on the set of keywords contained in the documents. However, the efficiency of SSE schemes comes at the cost of leaking some information about the plaintext, namely, the *access pattern*, i.e., the result of the query or the document IDs corresponding to the queried keyword, and the *search pattern*, i.e., searchable encryption schemes use deterministic encryption for the search keywords (two searches are deterministically the same or not). This leakage can make SSE schemes vulnerable to inference attacks (Islam et al., 2012; Cash et al., 2015) if the attacker has enough background knowledge about the encrypted documents.

Most SSE schemes are efficient with search complexity that is sublinear in the number of documents. However, SSE schemes support only equality queries and allow only the owner of the secret key to write and read data. On the other hand, public key searchable encryption schemes allow multi-users to write using the public key and only the private key holder to read the data. They also support more queries such as range queries. All this functionality of

public key searchable encryption schemes comes at the cost of reduced efficiency, as most schemes make use of pairing-based cryptosystems such as the scheme proposed in Boneh and Franklin (2001) which is not as efficient as standard public key cryptosystems.

The first public key searchable encryption scheme was proposed by Boneh et al. (2004). It is based on the anonymous identity-based encryption scheme by Boneh and Franklin (2001) and supports equality search. Later, Boneh and Waters (2007) proposed another public key searchable encryption scheme supporting conjunctive, subset, and range queries. Their scheme is a form of predicate encryption (Katz et al., 2008) which is a generalisation of attribute-based encryption where certain conditions are examined in the generated ciphertext. The conditions that are tested are represented by predicates (Boolean functions). A predicate encryption scheme uses a public key pk and a secret key sk. The public key pk is used to encrypt a pair (i, m), where i represents the attribute (searchable field) and m represents the data or payload message.

In a predicate encryption scheme, ciphertexts are associated with attributes and search tokens are related to predicates. For each description of a predicate f, the owner of the master secret key sk can generate a search token sk_f to perform search over a given ciphertext by evaluating the predicate f. Supposing that the ciphertext is $c = enc_{pk}(i, m)$ and sk_f is the search token for a predicate f, then $dec_{sk_f}(c) = m$, if $f(i) = 1$. Otherwise the decryption result will be empty. One major drawback of public key searchable encryption schemes (including predicate encryption schemes based on public key encryption) is that they cannot provide keyword privacy (or predicate privacy). For public key encryption schemes supporting equality or range queries after receiving a token sk_f, the server can reveal information about the predicate f by guessing the attribute i and, using the public key pk, generating the corresponding ciphertext c_j. In case the guess is correct, c_j should be decryptable by sk_f. Protecting the search tokens from external adversaries is a requirement when using public key searchable encryption schemes. However, the existing techniques still rely on a trusted server. Therefore, symmetric predicate encryption schemes (Shen et al., 2009) and functional encryption schemes (Bishop et al., 2015; Kim et al., 2018)[24] with function-hiding property have been proposed to prevent adversaries and malicious servers from learning any information about the ciphertext.

4.4. Functional encryption

Functional encryption (FE) is a generalisation of predicate encryption (Boneh et al., 2011). It is a cryptographic mechanism (Agrawal et al., 2013), which lets an authorised entity evaluate the value of an encrypted function on encrypted data and obtain the result in clear text. This is of particular interest for biometric verification because a biometric template or model is hard coded in the function and biometric data is kept secret as well. Take the example of a bank willing to use a cloud infrastructure to verify the identity of a caller. The bank has stored a private key on the SIM cards of the smartphones of every enrolled client. When calling the bank, a client claims an identity. The cloud has an encrypted function associated with that identity. This function is able to compute a verification score on the encrypted speech data of the presumed client (the biometric reference data is encrypted with the private key stored in the SIM card of the smartphone). If the score is high enough, the identity verification is successful. For example, in Kim et al. (2018), Hamming distances $d(x, y)$ between n-bit binary vectors x, y (as binary form: x', y') are computed by $< x', y' > = n - 2\, d(x, y)$, based on which an inner product encryption scheme is outlined. Thus far, functional encryption has not attracted significant attention in terms of application to privacy preservation for speaker and speech characterisation.

4.5. Differential privacy

Dwork (2006) proposed the *differential privacy* model, which has been used as a standard for data privacy. A data set D is a collection of elements and a randomised *query mechanism* M produces a response when performed on a given data set. Two data sets D and D' are said to be *adjacent* if they differ by at most one element. There are two proposed definitions for adjacent data sets. The stronger one is based on deletion: D' contains one entry less than D. The weaker one is based on substitution: one entry of D' differs in value from that in D. Differential privacy is used in speech characterisation, see Section 7.3.

[24] https://github.com/kevinlewi/fhipe

The query mechanism M is said to satisfy differential privacy if the probability of M resulting in a solution S when performed on a data set D is very close to the probability of M resulting in the same solution S when executed on an adjacent data set D'. Formally, we say that a randomised function M satisfies ε-differential privacy if for all adjacent data sets D and D' and for any $S \in range(M)$, denoted by:

$$\left| \log \frac{p(M(D) = S)}{p\Big(M(D') = S\Big)} \right| \leq \varepsilon. \tag{1}$$

The value of the ε parameter is referred to as *leakage* and determines the degree of privacy. A smaller ε represents a stronger privacy. In practice, ε is set less than 1 (e.g., 0.1 or $\ln(2)$).

As a consequence of this model, we may consider that if an individual chooses to contribute to the data set, there is a little or no increase in privacy risk for the individual as compared to not choosing to contribute to the data set. With the aim of designing mechanisms that satisfy differential privacy, two composition theorems are used:

Theorem 1 (Sequential composition). If a set of mechanisms $\{M_1, M_2, \ldots, M_n\}$ is sequentially performed on a data set, and each M_i provides ε_i differential privacy, then the entire mechanism will provide $\left(\sum_{i=1}^{n} \varepsilon_i\right)$-differential privacy.
Theorem 2 (Parallel composition). If $\{D_1, \ldots, D_n\}$ is a partition of the data set and M_i is a ε-differential privacy mechanism applied on D_i, then the mechanism that applies the M_1, M_2, ..., M_n in the corresponding sets $D_1, \ldots, D_n$ is a $\left(\max\{\varepsilon_1, \ldots, \varepsilon_n\}\right)$-differential privacy mechanism.

These two theorems help to control the degradation of privacy when we need to compose several differentially private mechanisms. To achieve differential privacy, Dwork (2006) proposed the *exponential mechanism* for releasing continues-valued functions satisfying ε-differential privacy. Given a function f to be evaluated over the data set D, we need to add a perturbation to the value of $f(D)$ to prevent leakage. To do this, the mechanism M adds the appropriate perturbation η, such that $f(D) + \eta$ satisfies differential privacy. The distribution of η is determined by the *sensitivity* of f of the data set D. This is the maximum difference between $f(D)$ and $f(D')$ where D' is an adjacent data set. Formally, the sensitivity S of f is given by:

$$S = \max_{D,D'} \| f(D) - f(D') \|_1. \tag{2}$$

Hence, the sensitivity S of the function f indicates how much the function is likely to change after changing one instance from the data set. Dwork (2006) showed that if η is sampled from Laplace(S/ε), then the mechanism $f(D) + \eta$ satisfies ε-differential privacy. Note that the perturbation η introduces an error with respect to the true value $f(D)$ which is inversely proportional to ε. This implies a trade-off between privacy and utility. Differential privacy has become an important research field; for a general overview and further details see Dwork et al. (2014) and Cormode et al. (2018).

4.6. Hardware-assisted security

Secure computation between mutually non-trusting parties can also be achieved via hardware-assisted security by executing the code in a trusted execution environment (TEE). For instance, the Intel Software Guard Extension (Intel SGX) (McKeen et al., 2013) is a widely available TEE implementation deployed by Intel in their recent CPUs that can also be used to implement secure two-party computation (Koeberl et al., 2015; Gupta et al., 2016; Bahmani et al., 2017). A user can upload their data and code into the TEE, called an *enclave* in SGX, attest to a remote party that the right code has been loaded, and the code can then be executed on the trusted SGX hardware of the untrusted party. This provides confidentiality and integrity to both the user's data and the computation. Security guarantees rely on the CPU architecture itself isolating the enclave code and data from other environments, thereby making it impossible for any other part of the machine to tamper with the computation. Additionally, enclaves are encrypted on the machine to ensure confidentiality. However, SGX security guarantees can be violated by side-channel attacks (Xu et al., 2015; Costan and Devadas, 2016). Hardware-assisted security has been used for speech characterisation, see Section 7.4.

On a more technical level, initial code and data are copied from unprotected memory into an enclave. Afterwards, the code is run inside the enclave. To ensure that the user is communicating with the right enclave and that the

enclave code and data has not been altered, SGX incorporates a primitive called remote attestation (RA). To perform RA, a hash of the enclave's contents is signed using a specific key accessible only to the trusted hardware, and verified using a certificate issued by the manufacturer. This signature serves as proof that the user is communicating with an enclave running on SGX hardware loaded with code and data associated with a specific hash. Since code and data are copied from an unprotected location, they have to be sent over via a secure channel, which can be set up by the enclave code itself via public-key cryptography.

4.7. Floating point operations

The cryptographic techniques outlined so far are able to securely compute any or some specific functionalities and can be practical especially for applications that rely on integer or binary operations: the vast majority of encryption approaches are based on the modulo operation, more accommodating for integers than for floats; and STPC of arithmetic circuits operating in the integer domain has been an efficient solution for privacy-preserving arithmetic operations. This becomes an issue for speaker and speech characterisation as probabilistic predictions need to be carried out to compute (log-likelihood ratio) scores, requiring scalar, vector, and matrix floating point operations.

For these operations, either a conversion into the integer domain is necessary to enable secure computation on floating point inputs, or special protocols need to be employed. A trivial solution is to just scale the floats to integers, which can often be a trade-off between accuracy and efficiency. Of course, one can securely evaluate a Boolean circuit that implements floating point operations on binary float inputs (Demmler et al., 2015a), or use dedicated protocols (Aliasgari et al., 2013), but both approaches come with a heavy overhead. For some applications, it can also be more efficient to convert between the integer and floating point spaces for different parts of the computation (Aliasgari et al., 2013; Tkachenko et al., 2018).

In order to preserve accuracy, the floating point space needs to be addressed (floats are compositions of integers). The IEEE 754 floating point standard (IEEE Standards Association, 2008) encodes floats f with a sign S and a mantissa M times a base B raised to an exponent E: $f = (-1)^S \cdot M \cdot B^E$. The trivial scaling and truncating solution translates to approximating the base-exponent term with one (i.e., $B^E \approx 1$). The remaining float term is, then, represented by a signed integer. However, to preserve accuracy, an auxilliary compound representation for float values based on *S, M, B, E* is required that encrypts one or multiple integers that encode a float. In Thorne (2017), accuracy is preserved using the Paillier cryptosystem: solely the mantissa is encrypted, whereas other terms remain in plaintext. Due to the probabilistic nature of the employed cryptosystem, two encrypted representations of the same plaintext value are not the same value in the encrypted domain. Using unsigned integers, negative float values are accounted for by an alternative float representation: the plaintext integer domain is divided into four intervals: $\left[0, \frac{n}{3}\right)$ for positive float representations, $\left[\frac{2n}{3}, n\right)$ for negative float representations, and $\left[\frac{n}{3}, \frac{2}{3}n\right)$ as well as $[n, \infty)$ for the purpose of detecting overflows resulting from previous HE operations. This auxiliary floating point representation is used in HE schemes for speaker characterisation, see Section 6.1.

4.8. Summary

The advanced cryptographic primitives presented in this section allow the secure computation of many functionalities relevant to speaker and speech characterisation. Although efficiency is continually optimised, the deployment of these solutions outside of the cryptographic community still faces some hurdles. On the one hand, the need for cryptographic solutions for outsourced computation problems recently gained wide interest because of legislative efforts to preserve privacy. On the other hand, the solutions require expert knowledge for effective deployment, as the different solutions achieve varying degrees of efficiency, accuracy, and security.

5. Automatic speaker characterisation

This section provides a brief overview of automatic speaker characterisation, that is, fundamental approaches to biometric verification, biometric identification, and speaker diarization (the annotation of the speech sequence with speaker labels that indicate *who spoke when*). The presentation is aimed at the non-expert reader and so the terminology used below is adapted to that used in other fields of biometrics (ISO/IEC JTC1 SC37 Biometrics, 2017b) and computer science (diarization is currently not covered in biometrics standardisation, compare ISO/IEC JTC1 SC37

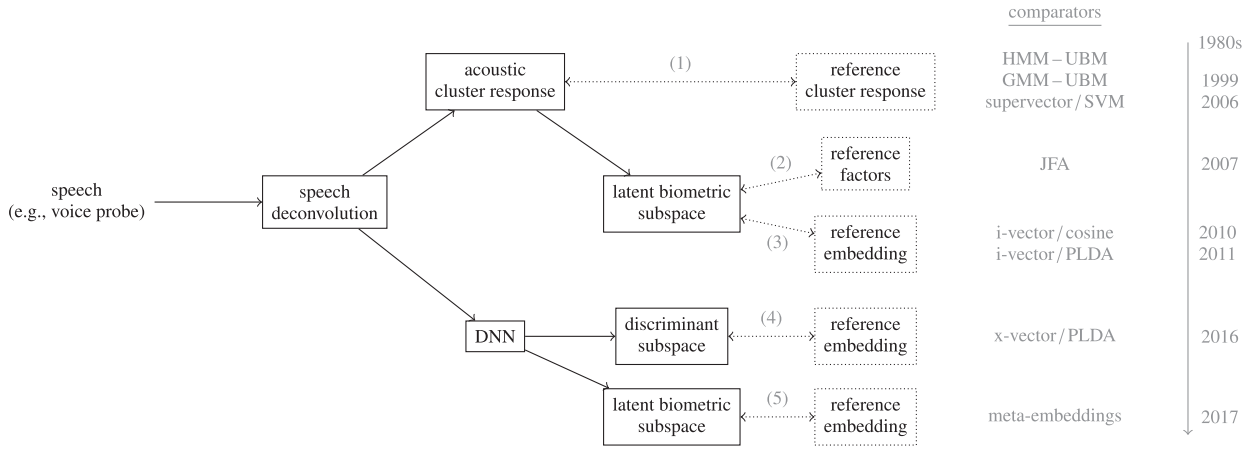


Fig. 3. Overview on speaker recognition with processing (solid), groups of comparators (dotted), and timeline.

Biometrics, 2017b). It also covers the full processing pipeline, illustrated in Figs. 3 and 4, which cover both feature extraction and biometric comparison. While the treatment focuses specifically upon automatic speaker recognition, many of the techniques described in this section are also applicable to diarization and other speaker characterisation applications.

Fig. 3 shows the most dominant automatic speaker recognition technologies. They are all based upon the deconvolution of speech signals. This pre-processing step is necessary to separate so-called source and filter components. This is achieved by a process known as homomorphic analysis (Oppenheim and Schafer, 1968) which is typically applied to short-term intervals of the speech signal. The source component comprises *pitch* and *glottal pulse* information, whereas the filter component represents *vocal tract* information. Traditionally, acoustic features used for automatic speaker recognition encompass only the latter. In contrast to other fields of biometrics, acoustic features used for automatic speaker recognition are an *inferred* representation of the vocal tract, rather than being a directly *observed* biometric characteristic. With speech being a dynamic signal, feature vectors are extracted periodically, e.g., from *sliding windows* typically representing in the order of 25 ms of consecutive speech and with a 10 ms window offset. While there are a host of alternatives, the most popular acoustic features are mel-frequency cepstrum coefficients (MFCCs) (Stevens et al., 1937; Bridle and Brown, 1974; Davis and Mermelstein, 1980).

Approaches to biometric comparison, also shown in Fig. 3 by groups of comparators, encompass probabilistic identity inference by subspace identity models and DNN based feature extraction. Group (1) of Fig. 3 corresponds to comparisons using probabilistic cluster responses between an acoustic cluster, i.e., a universal background model (UBM) and a reference model, i.e., a hidden Markov model (HMM) (Rabiner, 1989) or a Gaussian mixture model (GMM) (Reynolds et al., 2000). In this case, the biometric information is the reference-specific cluster centroids (i.e., the mean values of the probabilistic clusters). The concatenation of these mean values is referred to as a *supervector* (Kinnunen and Li, 2010). Groups (2,3) in Fig. 3 correspond to techniques that decompose supervectors into factors that represent both biometric and non-biometric subspaces. These factors are referred to as *probabilistic embeddings*, namely latent (inferred, rather than observable) representations that lie in a lower dimensional subspace that is more immune to nuisance variation. In group (2), joint factor analysis (JFA) (Kenny, 2005; Kenny et al., 2007;

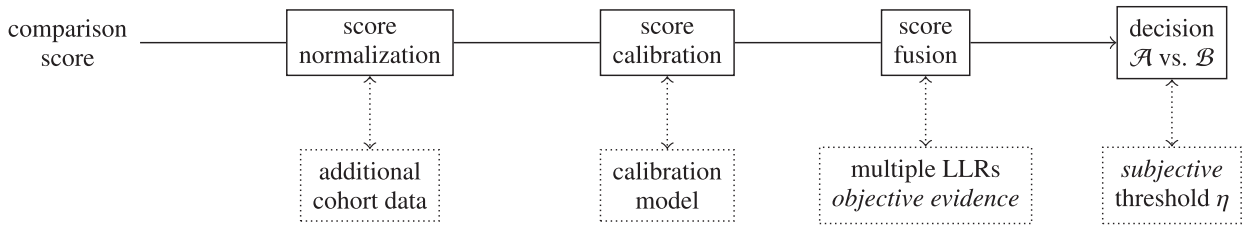


Fig. 4. Overview on score processing for the purpose of making good informed decisions on average.

Glembek et al., 2009) explicitly yields *biometric* and *non-biometric* embeddings; scoring the likelihood of non-biometric factors given biometric factors as reference. By contrast, group (3) yields a *total variability subspace*, referring to factors as intermediate-sized vectors (*i-vectors*) (Dehak et al., 2011). Pairs of reference − probe i-vector embeddings are compared by cosine distance similarity, and probabilistic linear discriminant analysis (PLDA) (Prince and Elder, 2007; Prince, 2012; Garcia-Romero and Epsy-Wilson, 2011; Cumani et al., 2013). Groups (4,5) encompass *discriminative embeddings*, which are derived using deep learning techniques (referred to as *x-vectors*) (Snyder et al., 2016; 2017; 2018b), are also compared by PLDA. In the case of automatic speaker recognition, *biometric information* in the form of *templates* are generally point estimates (usually of high quality), e.g., supervectors, JFA factors, or embeddings (without uncertainty propagation). However, when facing variable quality conditions, uncertainty needs to be propagated in a principled manner, e.g., i-vector embeddings estimated alongside their uncertainty are a *model* rather than a *template*.[25] Group (5) relates to the extraction of *meta-embeddings* (Brümmer et al., 2017; 2018) which are capable of principled uncertainty propagation.

Whatever the approach to automatic speaker recognition, some form of normalisation (Kinnunen and Li, 2010) is usually applied to compensate for nuisance variation, e.g., cepstral mean and variance normalisation that marginalises microphone and other channel effects via normally distributed data. Other reasons to normalise scores are to improve system callibration and fusion. A general approach to score post-processing is illustrated in Fig. 4. A large, auxiliary set of cohort data is often used in conjunction with references and probes to normalise the scores produced by some of the systems described above. Score calibration is often applied to transform scores into log-likelihood ratios (LLRs). However, score fusion techniques can be applied to improve reliability by combining the scores produced by different automatic speaker recognition systems, e.g., in a way that they produce LLRs that reflect the weight of evidence for a given probe in a given reference − probe comparison. The use of LLRs, rather than raw scores, has distinct advantages, e.g., (a) the Bayes decision risk is minimised, and (b) decision making is solely inferred from biometric information, namely the proportion of *mated* and *non-mated* reference − probe pairs (ideally, encoded by a score *in its value*: the LLR).

The preservation of privacy in speaker characterisation applications presents a formidable challenge, for privacy must be preserved *throughout the full processing chains* illustrated in Figs. 3 and 4. In groups (1−5) of Fig. 3, *biometric information* is represented by *models* rather than templates. This places greater demands on privacy preservation or, more specifically, the employed cryptography algorithms that must operate upon them. Algorithmically, templates are simply expected values of some form (i.e., averages of high precision), whereas more complex models centre data (i.e., remove averages of low precision) in order to comparatively report on the remaining uncertainty in the biometric similarity relative to the remaining uncertainty in the biometric dissimilarity (i.e., in the form of an LLR). Algebraically, groups (1−4) rely on *logsums*, and *inner products* (i.e., *dot products*), where group (5) and end-to-end uncertainty propagation also rely on *matrix inversions* and *log-determinants* (relying on *floats*, not *integers*). Privacy must often be preserved in all of these computations. One of the principal difficulties of bringing privacy preservation to automatic speaker characterisation lies in the very nature of speech signals. Whereas modulo arithmetic operating on *integers* is fundamental to most encryption techniques, the representation of speech signals and computation upon them involve *floating point* operations. Furthermore, the application of cryptography to speech signals must account for variable signal quality and, therefore, the propagation of uncertainty. The latter translates to the computation of *matrix inversions* and *log-determinants*, typically expensive computations that are ill-suited to being performed in the encrypted domain. Finally, score normalisation, calibration, and fusion must also be performed in the encrypted domain. This only exacerbates the computational demands involved in only a single reference − probe comparison.

The difficulty in applying privacy preservation to automatic speaker characterisation also requires reflection upon the system architecture at a higher level. At the highest level, there is a need to *(i)* preserve the privacy of *data subjects* and *(ii)* protect the data of *comparator vendors*. Regarding *data subjects*, data protection considers the commonly understood data privacy. Biometric references are stored as point estimates (i.e., templates; assuming no uncertainty) or as models (i.e., with the uncertainty about the point estimate). In the case of a *data breach*, the outsourced protected information (e.g., protected biometric references) should not be linkable by any adversary.

[25] The vast majority of the literature assumes high precision after embedding extraction (for the sake of computational effort), such that embeddings are treated as templates (approximated from models) and the uncertainty of the feature estimate is not further considered. By contrast, Cumani (2015) propagates the uncertainty of i-vector embeddings (as models) in a principled manner throughout the PLDA comparison.

Regarding *comparator vendors*, data protection considers data security of model parameters, such as of UBM, JFA, and PLDA. To compute LLRs, biometric discriminant features need to be extracted, and compared distinguishing between the variability within and between samples of speakers. The development of such systems (i.e., the model parameter training) requires data, such that comparator vendors are interested in preserving the security of their data from the perspective of a company (despite being non-sensitive data in terms of legislation, the data is sensitive to a company).

6. Privacy-preserving speaker characterisation

Specific privacy-preserving techniques are needed to deliver biometric information protection for speaker characterisation applications. In particular, the current techniques developed for other biometrics must be adapted from the protection of *templates* to the protection of *models*. This, of course, must be accomplished without degradation of recognition performance. This Section describes a number of potential techniques. They facilitate the storage of biometric voice references in an encrypted format which nonetheless supports comparisons between references and probes being made in the encrypted domain. During enrolment, privacy preservation concerns the protection of the reference (model), whereas during recognition, it also concerns the protection of the probe. Drawing upon the techniques presented in Section 4, suitable solutions are based upon homomorphic encryption, secure two-party computation, string representation comparisons and template/model binarisation techniques. These solutions are described in the following subsections.

6.1. Paillier cryptosystem-based methods

HE based encryption methods can be used to preserve privacy in both references and probes. The Paillier cryptosystem, a partially homomorphic encryption scheme, is well-suited to preserve privacy in vector comparisons in GMM−UBM approaches, as well as in the comparison of embeddings (*i-vector/x-vector*).

6.1.1. GMM−UBM comparison (supervectors)

The work in Pathak and Raj (2011, 2013) and Pathak et al. (2012) reports the adaptation of a standard GMM−UBM comparator to support privacy preservation using Paillier encryption. During enrolment, the UBM is adapted to the training data in the usual way (without encryption) to derive the biometric reference in the form of a speaker-specific GMM. This is then encrypted and stored in the enrolment database (on the server), while the associated secret key is retained by the end-user (on the client). Protocols are based on HE and secure two-party computation for both the enrolment and the verification phase in speaker recognition.

Therefore, the log likelihood of a Gaussian $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\Sigma)$ given observations $\boldsymbol{x}$ is formulated as: $\log\mathcal{N}(\boldsymbol{x}\mid\boldsymbol{\mu},\Sigma) = \widetilde{\boldsymbol{x}}^{\top}\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}}$ where $\widetilde{\boldsymbol{x}}$ is an extended vector obtained by concatenating $\boldsymbol{1}$ and $\boldsymbol{x}$, and $\widetilde{\boldsymbol{W}}$ is the parameters of the Gaussian distribution in the form,

$$\widetilde{\boldsymbol{W}} = \begin{bmatrix} -\frac{1}{2}\Sigma^{-1} & \Sigma^{-1}\boldsymbol{\mu} \\ \boldsymbol{0} & \boldsymbol{w}^{*} \end{bmatrix} \qquad \text{with} \qquad \boldsymbol{w}^{*} = -\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\log|\Sigma|\,. \tag{3}$$

By denoting $\boldsymbol{x}$ as all pairwise product terms $\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_j$, the log-likelihood results as: $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}$, where $\boldsymbol{W}$ is the vectorised form of $\widetilde{\boldsymbol{W}}$. Hence, the parameters of both, a UBM and an adapted model, can be represented as a sequence of $\boldsymbol{W}$ matrices.

During verification, encrypted acoustic features are sent to the server, which computes the log-likelihoods in the encrypted domain for the UBM and the components of the encrypted reference GMM (representing the claimed identity). The comparison score is obtained from component-wise encrypted log-likelihoods by the logsum protocol, which requires additional communication between the server and the client.

6.1.2. Cosine comparisons of embeddings (i-vector/x-vector)

The cosine distance similarity is often used to compare embeddings in terms of correlation, usually after the application of various pre-processing operations, e.g., linear discriminant analysis and whitening. During

enrolment, length-normalised i-vector features are stored as encrypted data. In verification, protected references are loaded by the client, such that comparisons with the length-normalised probe i-vectors are carried out under encryption. The score is computed in the encrypted domain, sent to an authentication server, which decrypts the encrypted score in order to make a verification decision. The cosine comparison of F-dimensional embeddings $\boldsymbol{X} = \{x_1, \ldots, x_F\}, \boldsymbol{Y} = \{y_1, \ldots, y_F\}$ (with reference $\boldsymbol{Y}$ and probe $\boldsymbol{X}$) is derived as (Gomez-Barrero et al., 2016; 2017b; Nautsch et al., 2018):

$$S_{cos}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}}{\|\boldsymbol{X}\|\ \|\boldsymbol{Y}\|} = \sum_{f=1}^{F} \frac{x_f}{\|\boldsymbol{X}\|}\frac{y_f}{\|\boldsymbol{Y}\|}\,, \qquad enc_{pk}(S_{cos}(\boldsymbol{X},\boldsymbol{Y})) = \prod_{f=1}^{F} enc_{pk}\left(\frac{y_f}{\|\boldsymbol{Y}\|}\right)^{\frac{x_f}{\|\boldsymbol{X}\|}}, \tag{4}$$

where the protected reference $\boldsymbol{Y}_{cos}^{enc_{pk}}$ is defined for length/normalised features as: $\boldsymbol{Y}_{cos}^{enc_{pk}} = \left(\left(enc_{pk}(y_f)\right)_{f=1}^{F}\right) = enc_{pk}(\boldsymbol{Y})$.

Fig. 5 illustrates the deployable distributed architecture: a client C extracts the probe feature vector $\boldsymbol{X}$ and requests the encrypted reference feature vector $enc_{pk}(\boldsymbol{Y})$ from the database $\mathrm{DB}_{controller}$. Scores are calculated on the client and sent to the authentication server $\mathrm{AS}_{operator}$, holding the key pair (pk, sk). The $\mathrm{AS}_{operator}$ outputs the decision D of whether the decrypted score S_{cos} is greater or equal to a threshold η, or not. Ideally, $\mathrm{DB}_{controller}$ is in the domain of an independent data controller, restricting access to operators.

6.1.3. PLDA comparisons of embeddings (i-vector/x-vector)

In contrast to the cosine comparator, comparators of the PLDA family compute LLRs S_{PLDA}. Therefore, log-expectations are compared, assuming reference and probe data originate from the same speaker, such that stacked embeddings correlate within class variance $\boldsymbol{\Sigma}_{within}$, whereas if they stem from different speakers, solely the total variability $\boldsymbol{\Sigma}_{total}$ is modelled (i.e., $\Sigma_{within} = \boldsymbol{0}$). For centred embeddings, PLDA scores are computed as (Garcia-Romero and Epsy-Wilson, 2011):

$$S_{\mathrm{PLDA}}(\boldsymbol{X},\boldsymbol{Y}) = \log\mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\middle|\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\Sigma_{total} & \Sigma_{within}\\ \Sigma_{within} & \Sigma_{total}\end{bmatrix}\right) - \log\mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\middle|\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\Sigma_{total} & \boldsymbol{0}\\ \boldsymbol{0} & \Sigma_{total}\end{bmatrix}\right),$$

$$S_{\mathrm{PLDA}}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X} + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{Y} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X} + const$$

$$with \quad \boldsymbol{Q} = \Sigma_{total}^{-1} - \left(\Sigma_{total} - \Sigma_{within}\,\Sigma_{total}^{-1}\,\Sigma_{within}\right)^{-1}, \qquad \boldsymbol{P} = \Sigma_{total}^{-1}\,\Sigma_{within}\left(\Sigma_{total} - \Sigma_{within}\,\Sigma_{total}^{-1}\,\Sigma_{within}\right)^{-1}. \tag{5}$$

By employing the *Frobenius inner product* (denoting the operation $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\,\boldsymbol{y} = \mathrm{vec}(\boldsymbol{A})^{\mathrm{T}}\,\mathrm{vec}(\boldsymbol{x}\,\boldsymbol{y}^{\mathrm{T}})$ with the operator $\mathrm{vec}(\cdot)$ stacking matrices into vectors), the PLDA score is expressed in terms of a dot product (Cumani et al., 2013):

$$S_{\mathrm{PLDA}}(\boldsymbol{X},\boldsymbol{Y}) = \begin{bmatrix}\mathrm{vec}(\boldsymbol{Q})\\ \mathrm{vec}(\boldsymbol{P})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathrm{vec}(\boldsymbol{X}\,\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{Y}\,\boldsymbol{Y}^{\mathrm{T}})\\ \mathrm{vec}(\boldsymbol{X}\,\boldsymbol{Y}^{\mathrm{T}} + \boldsymbol{Y}\,\boldsymbol{X}^{\mathrm{T}})\end{bmatrix} + const\,, \tag{6}$$

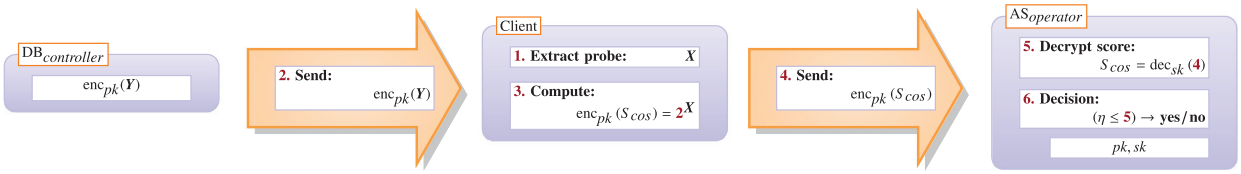


Fig. 5. Architecture: HE-protected cosine similarity comparison for length-normalised features, from Gomez-Barrero et al. (2016) and Nautsch et al. (2018), with client, servers (blue) and communication channels (orange). Notably, the depicted architecture is the HE solution for the *dot product*. The steps in the protocol are: 1. the extraction of the probe is being embedded, 2. the encrypted reference is sent to the client device, 3. the encrypted score is computed via Eq. (4) (the equation is symbolically indicated in the form of $2^{\boldsymbol{X}}$ with 2 representing the data from step 2 and $\boldsymbol{X}$ the probe data), 4. the encrypted score is sent to the authentication server, 5. the encrypted score is decrypted (symbolically indicated by $dec_{sk}(4)$ with 4 representing the data of step 4), and 6. the plaintext score is compared to a threshold η to make a decision (with 5 indicating the data of step 5). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

such that the same encryption scheme illustrated in Fig. 5 can be employed to preserve the *data privacy of individuals*. Notably, terms solely depending on the reference $\boldsymbol{Y}$ are storable pre-computed (Nautsch et al., 2018).

For the purpose of preserving *data security of vendors* (i.e., the vec($\boldsymbol{P}$), vec($\boldsymbol{Q}$) terms), the encryption scheme is employed twice: firstly, auxiliary data aggregates the *biometric information* in the encrypted domain with key pairs generated by the *operator*. Secondly, these aggregates are decrypted to plaintext, such that *vendor* model parameters $\boldsymbol{P}$, $\boldsymbol{Q}$ (protected by a second key pair generated by the vendor) are employed to compute LLRs (Nautsch et al., 2018).

6.2. GMMs using secure two-party computation

Privacy-preserving speaker verification systems that rely on secure two-party computation have also been proposed in the literature. The garbled circuit protocol (see Section 4.2) has been used in a privacy-preserving GMM−UBM-based speaker verification system proposed by Portêlo et al. (2014). During verification, the user, who is already enrolled in the system, is responsible for generating the GCs for the verification computation while the system is responsible for evaluating them and deciding on whether or not to authenticate the user. In the particular case of GMM models, there are two critical operations that must be encoded in a GC: the scalar product (i.e., inner product or dot product) and the logsum operation. The former is a simple linear operation consisting of multiplications and additions and is therefore straightforward to implement. The logsum is required to compute the occurrence probabilities of events and is a nonlinear operation. The work in Portêlo et al. (2015b) considers linear piecewise approximations for implementing logsums in GCs.

Given these two components, performing a GMM evaluation becomes a problem feasible to solve using GCs. This approach is able to obtain similar results to the non-private counterpart, but at the same time guarantees that none of the participants in the protocol reveal their private information. A major advantage is that this approach is quite efficient in terms of execution times, especially when it is possible to evaluate the circuit in parallel (under the UBM−GMM model, different audio frames are considered independent, so their evaluation can be computed in parallel). However, the efficiency can be affected significantly if the system has computational constraints because the execution time scales linearly with the number of GMM components.

A relevant drawback of this scheme is the fact that the verification system needs to know both the UBM and the adapted model in plain. This last component represents a privacy leakage from the user's perspective and, even though the system does not have access to a raw audio signal or a sequence of feature vectors needed to make the acceptance decision, the system is in possession of a characterisation of the client's voice given by the parameters of the adapted model. A noteworthy advantage of this scheme is that it only needs two parties compared to the HE-based ones presented in Sections 6.1.2 and 6.1.3.

This approach keeps the inputs private because the underlying general-purpose STPC protocol is provably secure (Lindell and Pinkas, 2009). Another line of research applies special-purpose STPC protocols based on Lu et al. (2014) to GMMs (Rahulamathavan et al., 2018) and i-vectors (Rahulamathavan et al., 2019). However, their underlying protocols do not have a rigorous security proof.

6.3. Distance-preserving hashing techniques

Akin to the literature on *cancelable biometrics*, hashing techniques are also employable to speaker characterisation. Two groups of hashing methods are presented, based on string comparison, and on binary embeddings and secure modular hashing.

6.3.1. Speaker comparison as string comparison

Pathak and Raj (2012b) proposed a method for speaker verification that requires minimal computational overhead to satisfy privacy constraints. The basic idea is to convert the speaker verification task to a string comparison. In this case, the utterances are converted into *supervector* features that are invariant with the length of the utterance. The verification can be based on nearest-neighbour classification, which is reduced to a string comparison transforming these features applying a Locality Sensitive Hashing (LSH) transformation. To prevent the system from gaining information about supervectors, the LSH transformation is converted into an obfuscated string applying a cryptographic hash function, e.g., SHA-256. Cryptographic hash functions satisfy the property that they can be computed

efficiently but are computationally hard to invert. Moreover, they are orders of magnitude faster to compute compared to homomorphic encryption schemes.

However, the attempt of protecting the LSH transformation using a cryptographic hash can be frustrated just applying brute-force search. Indeed, as the possible values of the LSH transformation lie in a relatively small set (by crytographic standards), a dictionary attack is entirely feasible. To mitigate this vulnerability, Pathak and Raj (2012b) propose to concatenate a long random string r_i to the LSH transformation before applying the cryptographic hash. This string r_i, referred to as a *salt*, must be unique to the user i and the system. Consequently, the user must keep it private so as to prevent the system from gaining information. The server never observes any LSH key before a salted cryptographic hash function is applied, and the end-user does not need to store any speech data on their device.

6.3.2. Binary embeddings and secure modular hashing

Secure binary embedding (SBE) (Boufounos and Rane, 2011) is a scheme for preserving privacy based on LSH, that uses a quantised random projection. Using the Hamming distance metric, the embeddings can be used to determine whether or not two points are close enough to meet a dissimilarity threshold. Moreover, if the distance between the reference and probe as points is large, the leakage of information (i.e., mutual information between embeddings) is negligible. Later, Jiménez et al. (2015) provided a generalisation of this method, named Secure Modular Hashing (SMH), by extending the quantisation function and showing that the collision probability of two randomly chosen embeddings depends on the Euclidean distance between the original points. Jiménez and Raj (2017a) show that it is possible to obtain an accurate estimation of the Euclidean distance between points using a *modular distance*, as long as the original points are close enough. Therefore, hashes hide the original vectors, which contain private information, but they allow the estimation of distances as long as the original vectors are close enough.

The application of both SBE and SMH to speaker verification systems is straightforward: if a classifier could be made to operate on SBE hashes of i-vectors rather than on the i-vectors themselves, speaker verification may be performed without exposing speaker data. In Portêlo et al. (2013) and Jiménez et al. (2015), SVM classifiers were considered as non-private reference speaker verification systems for speaker modelling. Under this privacy-preserving scheme, SVM kernels must be modified to work with Hamming distances between hashes. In particular, the authors propose to use a modified version of the Radial Bases Kernel, where the Euclidean distance is replaced by the mean Hamming distance between hashes.

Jiménez and Raj (2017b) propose the modification of the distribution of the random parameters in the hash function to change the dependency on the Euclidean distance by the Manhattan distance. Similar properties with respect to the mutual information between hashes are shown; the mutual information is negligible if the ℓ_1 distance is large. The technique is applied to the speaker verification task using nearest neighbour classification. In addition, it is shown that properties of cancelable biometrics are satisfied by this kind of transformation. Moreover, Portêlo et al. (2015a) propose the use of SBE as a mechanism to hide information but show that it is possible to apply Dynamic Time Wrapping directly to embeddings so as to retrieve speech data. Although this scheme provides information-theoretic security on its own, there is no guarantee that an attacker is not able to infer any information about the plaintext vectors if he manages to obtain the secret keys, or if he has some prior knowledge about the plaintext vectors.

6.4. Binarisation of templates, models, and features

Biometric data binarisation (of templates, models, or features) is a pre-processing step that is required by many cryptographic protection mechanisms. We first describe several binary representations that were developed originally for biometric speaker characterisation rather than for privacy preservation. Binarisation schemes developed specifically for template protection are reviewed subsequently.

6.4.1. Binary approaches to speaker characterisation

Many voice biometric applications require lightweight processing. For such scenarios, binarisation has been explored as a means to speed-up computation. In a speaker identification framework based on the GMM–UBM paradigm, the work in Billeb et al. (2014), proposed a computationally efficient, two stage identification system which uses binarised voice biometric templates to pre-screen a large database before performing the final identification

step using the original template. The number of full comparisons is reduced to a fraction of the original number, while the original speaker identification performance is preserved.

Another method referred to as *binary key speaker modelling* was proposed as an efficient method to represent speaker-discriminative information in the form of single binary vectors. It was initially proposed in Anguera and Bonastre (2010) for speakers recognition and further improved in Bonastre et al. (2011) and Hernandez-Sierra et al. (2014), and applied to speaker diarization in Anguera and Bonastre (2011), Delgado et al. (2015), and Patino et al. (2018). The same technique has also been applied to model other voice characteristics, such as vocal emotion (Luque and Anguera, 2014). From a privacy perspective, this representation has the advantage of being binary by definition. While binary keys require any additional pre-processing step, it is thus ideally suited to being coupled with traditional protection mechanisms which require a binary input. The method relies on a generator model composed of a collection of single Gaussian components, usually derived from a UBM trained on traditional acoustic features.

Fig. 6 depicts the process by which a sequence of acoustic features (e.g., MFCCs) is transformed into a single binary vector referred to as a binary key (BK). It uses a generator model composed of N components, whose size dictates the dimension of the final binary representation. The input features (1) are processed frame-wise by evaluating their likelihoods, given all the Gaussian components of the generator model (2). An initial feature-level binarisation (3) is obtained by activating the bits corresponding to the positions of the top-M scoring Gaussians indicated in solid blue color in (2). Next, a sequence-level accumulation is performed by the row-wise sum of the binarised features (4), thereby resulting in what is referred to as a cumulative vector (CV). Finally, an additional quantisation is performed to generate the final BK by activating the bits corresponding to the top M positions in the cumulative vector (5).

As described above, the method provides binary representations of a sequence of feature vectors corresponding to a speech utterance at two levels, namely at the feature level and the utterance level. Acoustic processing and matching is performed only at the beginning of the process when computing likelihoods of acoustic features given the Gaussian components of the generator model. The verification stage is performed by thresholding a similarity measure between the pair of enrolment-test BKs. Binary metrics such as the Jaccard similarity are an appropriate means to compare binary vectors. In addition, well known session compensation methods, such as nuisance attribute projection, have been proven to be effective when operating upon binary representations Hernandez-Sierra et al. (2014).

6.4.2. *Binarisation for protecting speaker templates and models*

A suitable binary representation is required by many protection mechanisms. Based upon a GMM–UBM framework, Billeb et al. (2015) propose a binarisation technique, which is used to extract scalable high-entropy binary voice reference data (templates) from speaker models. Binary feature vectors are then protected with a fuzzy commitment scheme which uses error correction list-decoding to overcome the high intra-class variance in voice samples. Experimental work showed that the system achieved privacy protection at a negligible loss of recognition performance.

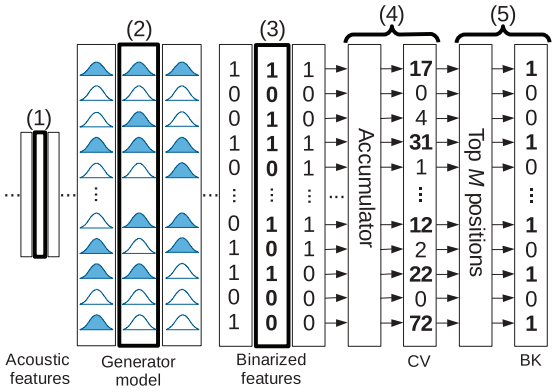


Fig. 6. Binary key extraction process. CV and BK refer to cumulative vector and binary key, respectively. Figure adapted from Patino et al. (2018).

Paulini et al. (2016) propose a multi-bit allocation binarisation technique, also based upon the GMM–UBM paradigm. The proposed scheme is designed to extract discriminative compact binary feature vectors to be applied in a voice biometric template protection algorithm. Their binarisation acts over GMM supervectors estimated over MFCC features. The feature space is generically divided into intervals, which are encoded with multiple bits using a Gray code. In addition, relevant projections are employed to detect most discriminative and stable supervector elements per enroled subject. Experimental work shows that the resulting binary representation causes only a marginal decrease in recognition performance compared to the baseline system.

Based on binary key speaker modelling by Anguera and Bonastre (2010), a speaker template protection approach for speaker verification was proposed in Mtibaa et al. (2018). A binary voice vector (template) is extracted from a speaker model (GMM) and is then transformed using a shuffling scheme inspired by Kanade et al. (2012). The shuffled binary voice vector represents the cancelable template. By applying the shuffling scheme, this architecture has the desired properties of renewability and unlinkability (see Section 3). If the stored shuffled template is compromised, the user can cancel the old template and issue a new one by changing the shuffling key. Additionally, the authors showed that the protected system even outperforms the unprotected system. And, even if the biometric data is compromised, the EER of the system in such a scenario remains lower than that of the baseline biometric system. The disadvantage of this approach is that the user needs to permanently store their GMM, implying that it is vulnerable to recovery attacks.

6.5. Summary

The data needing protection within a speaker characterisation system fundamentally defines the data privacy safeguards employed. Homomorphic crypotosystems can be used to compute the crucial inner product and logsum operations in a privacy-preserving fashion. In particular, the Paillier cryptosystem is a popular technique for preserving privacy in GMM–UBM systems as well as in cosine, and PLDA comparators (i-vector and x-vector embeddings). Notably, the Paillier-based architectures can also be applied to protect PLDA model parameters and thus protect vendor data as well as end-user data. Also, while the additively homomorphic Paillier cryptosystem is the most popular in the literature, other homomorphic approaches might be more promising, especially when targeting *post-quantum secure* encryption. The garbled circuits STPC protocol can be used for GMM–UBM comparisons. For supervectors, string comparison and binarisation schemes are employed, allowing to sustain data privacy via conventional hashing approaches (such as for passwords). However, speech data variability reduces the biometric recognition performance, though methods like error correcting codes are capable of recompensation.

Cryptographic approaches trade off efficiency for privacy, resulting in varying overheads of computation, communication, and rounds of interaction. In general, HE-based solutions are computationally heavy, while STPC-based solutions impose some communication overhead. An issue is that the underlying cryptographic primitives operate upon integer representations while speaker characterisation requires floating point operations. Some privacy-preserving methods are capable of preserving privacy without degrading biometric recognition performance. While others do not sustain score discrimination or calibration properties, they can therefore be more efficient. Approaches to privacy preservation reported thus far do not cover the secure score computation of HMM, SVM, JFA, and meta-embedding approaches to speaker characterisation. Furthermore, they neither address end-to-end, probabilistic signal processing, nor score normalisation, calibration, and fusion. It is also worth mentioning that, although there has been little research on the topic of privacy-preserving language recognition, the techniques described in this section are easily adaptable to that topic. The reason for this is the fact that most of the state-of-the-art techniques applied in speaker characterisation are akin to those used in spoken language recognition, such as GMM–UBM, supervectors, i-vectors and x-vectors (González et al., 2011; Snyder et al., 2018a).

7. Preserving privacy in non-biometric speech applications

Extending the perspective of biometric privacy using voice data to non-biometric characterisation tasks, this Section provides an overview not only on secure computation methods, but also on preserving privacy in paralinguistics. On the one hand, methods are easily transferable to speaker characterisation, and on the other hand, we provide an overview on characterising speech without characterising speakers. Particularly, this section addresses methods such that biometric identities cannot be inferred by singling them out (see Section 2). From the perspective of speaker

characterisation, speech application as *non-biometrics* is an ample field. The following subsections summarise their parts in five groups on: homomorphic encryption, secure modular hashing, differential privacy, hardware-assisted privacy, and voice-activated functions. Thereby, the presented applications range over speech alignment, paralinguistics, and the recognition of speech, languages, and emotions. Regarding technology, we address the learning of speech characterisation systems, and deep neural networks.

7.1. Homomorphic encryption and secure two-party computation based methods

HE- and STPC-based methods can be used to preserve privacy in non-biometric speech characterisation applications. In the following, we outline methods for two technologies: hidden Markov models (HMMs) and deep learning.

7.1.1. HMM−UBM and speech alignment with homomorphic encryption

In signal processing, especially in applications related to speech processing, classification based on HMMs is a very common task. Using a client-server model, we consider that the client has a signal that s/he wants to analyse, while the server possesses accurate HMMs obtained via extensive training. The client wants to get the outcome of the HMM-based classifier provided by the server. In the context of this paper, we need to consider privacy constraints. Hence, a client device needs to prevent exposure of the data to the server and a server cannot share the HMMs with a client device. In Pathak et al. (2011), the client encrypts his/her data and provides it to a server which performs the HMM-based probabilistic inferences, using a protocol based on partially homomorphic encryption (see Section 4.1) along with 1-of-n Oblivious Transfer. More particular, the protocol for *Secure Forward Algorithm* uses a Secure Logarithm, a Secure Exponent and a Secure LogSum Protocol. All of these protocols depend on random masking and homomorphic operations.

7.1.2. Deep neural networks using homomorphic encryption and secure two-party computation

First proposed in Gilad-Bachrach et al. (2016) and later expanded by Chabanne et al. (2017), Hesamifard et al. (2017), Bourse et al. (2017), and Sanyal et al. (2018), this approach consists of replacing all operations in the inference stage of a Deep Neural Network (DNN), or Convolutional Neural Network (CNN), by their FHE or SHE counterparts (see Section 4.1). In this method, the DNN receives as input an encrypted feature vector, and computes an encrypted prediction that only the user can decrypt. Although currently limited in the number and size of the layers in the DNN due to the computational complexity of FHE and SHE schemes, this method has the advantage of protecting the user's privacy, as well as the privacy of the model's architecture and parameters. In addition, since the user does not take part in the computation, only two rounds of communication are necessary between the user and the service provider.

When using SHE, this approach requires replacing the activation functions of the network with polynomials. The reason for this is that most SHE schemes are not able to perform operations other than additions and multiplications. As such, regular activation functions (i.e., ReLU, Sigmoid), are replaced with polynomial approximations. However, polynomial approximations have small convergence intervals, outside of which they diverge quickly. To prevent this, a Batch Normalisation layer, e.g., see Ioffe and Szegedy (2015), may be introduced before each activation function, to guarantee that most of the activation's inputs fall within its convergence interval (Chabanne et al., 2017). Nevertheless, using polynomial activation functions may induce some problems when training the network, since these functions have unbounded derivatives, which in turn may cause the adjusted model to have some accuracy degradation when compared to a model with regular activation functions. On the other hand, some FHE cryptosystems allow bit-wise operations to be performed efficiently between ciphertexts, and thus to implement piece wise functions, such as the ReLU and Sign functions. Nonetheless, the efficiency of these schemes depends on the maximum absolute value that needs to be computed by the network, demanding a trade-off between the scheme's efficiency and the precision of the weights and inputs of the network (Bourse et al., 2017; Sanyal et al., 2018).

As previously stated in Section 4.7, most cryptosystems are based on the modulo operation, working only with integers. This raises an issue with the inputs and parameters of the DNN that are, for most applications, real numbers. Encoding techniques are often incompatible with approaches that might otherwise improve the efficiency of the scheme, such as HE *batching*, which enables Single Instruction Multiple Data (SIMD) operations. For this reason, in the literature, networks are often "discretised", either by scaling or through quantisation. However, these processes reduce the precision of the weights and input features of the DNN, which may lead to further accuracy degradation.

Overall, combining HE with DNNs can have particularly useful applications in the context of paralinguistic tasks, where potentially sensitive information may be extracted from the speaker's data. This has been shown in Dias et al. (2018) for emotion recognition, and in Teixeira et al. (2018), for the detection of voice-affecting diseases such as Cold, Depression and Parkinson's. Both examples used SEAL's (Microsoft Research, Redmond, WA., 2018) implementation of a SHE cryptosystem.

STPC has also been used to securely evaluate DNNs (Sadeghi and Schneider, 2008; Barni et al., 2011; Mohassel and Zhang, 2017; Liu et al., 2017; Riazi et al., 2018; Juvekar et al., 2018; Riazi et al., 2019). While usually revealing the DNN topology to the client, STPC-based secure evaluation of DNNs can achieve faster execution times and, in most cases, lower bandwidths than the HE-based approaches summarised above. Additionally, the DNN topology can also be hidden with STPC, but this incurs a logarithmic overhead (Sadeghi and Schneider, 2008). Although easily applicable, STPC-based secure evaluation of DNNs has not yet been used for speech tasks.

7.2. Secure modular hashing for private emotion recognition

As described in Section 6.3.2, Secure Modular Hashing (SMH) is a scheme based on Locality Sensitive Hashing (LSH) that has been applied to speaker verification systems based on nearest neighbour and SVM classifiers. In the context of paralinguistics, SVM classifiers have a widespread use, in part due to the advantages they pose when training data is scarce, which is more often than not the case. Following this line of reasoning, Dias et al. (2018) made a proof of concept on how SMH together with SVMs (using the same modified kernel as in Jiménez et al., 2015) can be applied to emotion recognition with little to no accuracy degradation when compared to a baseline (non-privacy-preserving) SVM classifier. This work left open the possibility of applying the same scheme to other paralinguistic tasks. However, in this framework, the resulting predictions are obtained in the clear (i.e., the predicted value is not private), which may be a breach of the speaker's privacy for some applications (e.g., health-related applications).

7.3. Differentially private recognition systems

In recognition systems, the query mechanism can be thought of as an algorithm learning the classification rule which is evaluated over the training data set. The output of an algorithm satisfying differential privacy is likely to be the same when the value of any single data set instance is modified, and therefore, no additional information can be obtained about any individual training instances with certainty by observing the output of the learning algorithm.

In addition, the ε-differential privacy model limits the information that an adversary can gain about a particular private value, by observing an entire function learned from a data set containing that value, even if s/he knows every other value in the database. Naturally, if we want to apply this method to the recognition problem (a.k.a. machine learning), we need to be aware of the trade-off between *privacy* and *learnability*.

Most of the training methods for a recognition system are based on optimisation problems. In practice, there are two strategies for inserting differential privacy on the learning process of a recognition system. The first type adds noise to the execution process of the corresponding optimisation algorithm. The second type makes a perturbation directly to the objective function, typically adding a differentially private noise before the learning procedure. Both strategies have been explored in different models, which we will now briefly mention.

1. *Large margin classifiers*
   Pathak and Raj (2012a) present an algorithm for learning a discriminatively trained multiclass Gaussian mixture model-based classifier that preserves differential privacy using a large margin loss function. The solution involves adding a perturbation term to the objective function. The authors show that differential privacy is satisfied and they establish a bound on the excess risk of the classifier learned which is directly proportional to the number of classes and inversely proportional to the privacy parameter reflecting a trade-off between privacy and utility.
2. *Multi-party classifiers with differential privacy*
   In many cases it is common to have the problem of learning a classifier from a multi-party collection of private data. The goal is to learn a classifier from the union of all the data. In Pathak et al. (2010), the imposed conditions are that: (a) none of the parties are willing to share the data with one another or with any third party (e.g., a

curator). (b) The computed classifier cannot be reverse engineered to learn about any individual data instance possessed by any contributing party.

3. *Differentially private deep learning*

   Deep learning has been one of the most successful techniques in machine learning and signal processing. The basic idea is to apply a multiple-layer structure to extract complex features from high-dimensional data and use them to build models. Each layer has a set of parameters that must be learned from training samples. The minimisation of a loss function, which depends on these parameters, is attempted. In practice, stochastic gradient descent (SGD) methods are used to achieve this objective.

   As with any machine learning, the knowledge of the model or its use on particular samples can leak information about the training data, which results in privacy risks. To deal with this issue, deep learning models can be adapted to reach differential privacy guarantees. Shokri and Shmatikov (2015) have designed a differentially private SGD algorithm by introducing a sparse vector technique. Similarly, Abadi et al. (2016) designed a differentially private SGD relying on a Gaussian Mechanism (Dwork et al., 2014). On the other hand, Phan et al. (2016) considered a perturbation on the objective function of a deep auto-encoder to achieve differential privacy.

Differential privacy can be used as a publication mechanism for different speech models, such as GMMs and HMMs. In fact, some of the protocols discussed in the previous sections need the system to send part of its recognition model to the client, for example, the UBM model. If the model is trained using publicly available data there are no privacy concerns. However, that is an uncommon situation, since in general most of these systems are trained using private speech data. Therefore, even though the system operators are willing to expose the model structure, it is necessary to protect the training data from any leakage through the model parameters.

Moreover, while homomorphic encryption and secure two-party computation solve the problem of private inference, these techniques do not deal with the problem of potential leakage of information using the system responses. In fact, an attacker may try to use the responses of the system to infer model parameters and eventually learn information about training samples. In this context, differential privacy can be added to prevent those types of attack.

7.4. Hardware-assisted privacy-preserving speech processing

The solutions outlined so far achieve privacy-preserving speaker and speech characterisation using purely cryptographic primitives. Hardware-assisted security (see Section 4.6) for private speech processing has been explored by Brasser et al. (2018) in an Intel SGX-based architecture called *VoiceGuard*. This solution essentially computes the speech characterisation inside an SGX enclave, thereby revealing no information about the client's data to the server or about the server's model to the client. This is more efficient than pure cryptographic solutions, as it leverages the lower computation and communication overhead of native SGX instructions.

Fig. 7 provides an overview of the VoiceGuard architecture. The figure displays the case of three involved parties with the user U supplying its speech input (e.g., an i-vector), the vendor V providing the speech characterisation through an acoustic model AM (e.g., a DNN) and a language model LM (usually a decoding graph); and the service provider P securely applying V's models on U's input without learning any information about the models or the input. A solution involving only U and V, with V also providing the SGX computation, is analogous. In the first phase, the preparation phase, U and V agree on what kind of speech computation code should be executed, and make sure that the code contains no sensitive outputs or other unwanted instructions that could leak sensitive data (steps 1 and 2 in the figure). In the following initialisation phase, the SGX enclave is created using the agreed upon code and is measured by the SGX architecture (step 3). The enclave code is, then, executed. First, a public key PK is created, which then is sent and attested to both U and V by employing the remote attestation (RA) feature of SGX (step 4). This proves that the enclave code measurement M is valid, that PK belongs to the correct enclave, and that messages encrypted with PK can only be decrypted inside that enclave. Upon verification, both U and V open a secure channel by sending the symmetric encryption keys K_U and K_V encrypted under PK to the enclave (step 5). Finally, in the operation phase, U provides its speech input using its secure channel (step 6). V's model inputs are already stored in an encrypted form, as on-demand loading could leak sensitive information through the access pattern. Using K_U and K_V, the generic speech recognition (SR-) engine is executed inside the enclave on the decrypted inputs. The resulting output can then be sent to U (step 7), to V, or to both (e.g., in the case of speaker verification). Repeated executions only require repetitions of the operation phase.

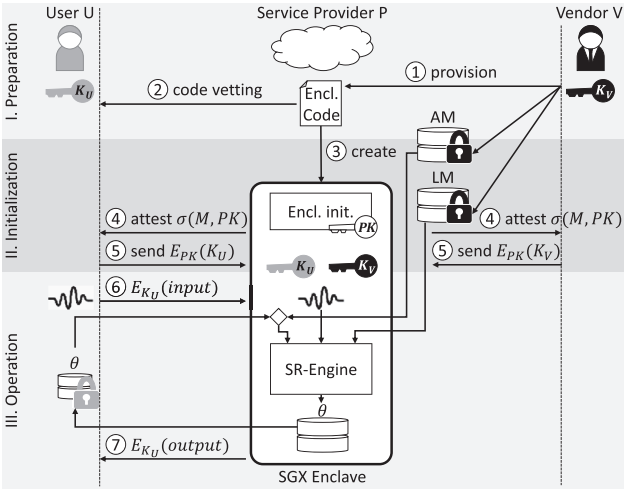


Fig. 7. VoiceGuard architecture using Intel SGX, taken from Brasser et al. (2018).

Brasser et al. (2018) show that VoiceGuard can achieve practical speech recognition performance when run on DNNs trained with various i-vector training sets. As it is a general architecture for speech characterisation tasks, applying VoiceGuard to other uses, such as speaker characterisation, is straightforward, and in theory any (speech) characterisation computation can be executed in a privacy-preserving way. However, one has to keep in mind the side-channel vulnerabilities of Intel SGX mentioned in Section 4.6. A limitation of VoiceGuard is that *AM* and *LM* need to be loaded into the enclave. For speaker verification, this is not an issue, as models are usually small in size, but the larger models of state-of-the-art speech recognition systems might require repeated loading of smaller chunks.

7.5. Achieving privacy in voice-activated applications

The following solutions allow the use of intelligent digital assistants whilst at the same time achieving some measure of privacy. One possible solution is an on-device speech recognition system such as the one proposed in Glackin et al. (2017). In this case, the intelligent digital assistant device could be the user's smartphone, laptop or desktop computer. The on-device solution avoids the data-in-use protection needed when performing computation in the cloud. It is more suitable for voice-activated applications since they normally process short-duration voice data in real time.

Performing speech recognition offline on the client side rather than on the cloud side means that at a minimum, the corresponding transcription hides the speakers' biological and environmental voice features, as noted above, and only reveals the transcribed texts to the cloud server to enable the server to respond to the speakers' queries. Some very private tasks can be done locally on the user side without using a cloud server such as making phone calls, home management, and calendar management. However, most tasks have to use an internet search engine to respond to queries depending on dynamic data such as news headlines, weather forecasts, travel information, shopping, etc. Most of the search engines store search data forever which is an obvious privacy threat. The only way towards full privacy is to develop and deploy a private search engine via the internet. Fortunately, the deployment of Intel SGX (see Section 4.6) by many cloud providers such as Alibaba Cloud, Azure, Baidu and IBM Cloud shows that SGX-based solutions will indeed be a reality in the near future. Therefore, rather than using encrypted search solutions such as fully homomorphic encryption (Gentry, 2009) or private information retrieval solutions (Lindell and Waisbard, 2010) to build private search engines that will hardly be scalable, an alternative solution will be to build an SGX-based private-search engine solution, such as the one proposed in Mokhtar et al. (2017). To achieve full privacy on voice-activated applications, the SGX-based private search engine can either be combined with an on-device speech recognition solution (Glackin et al., 2017) or SGX-based solution such as the one described above.

7.6. Summary

Preserving privacy in non-biometric speech technology is an obligation not only of the broader speech community as a whole, but in particular for research to speaker characterisation. On the infrastructure level, the Intel SGX technology encapsulates computations into a secured domain, where Intel is to be entrusted, moving hardware concerns regarding security of many entities. Such infrastructures are also relevant to the deployment of speaker characterisation systems. On training algorithms (also in distributed systems), differential privacy explicitly enforces algorithms to learn non-biometric feature spaces; a cross-validation fashioned learning procedure ensures ϵ-indifferential privacy, that is an additional training parameter, summarising the trading-in of acceptable privacy loss (which should be low). Notably, an inversion of such feature space training could yield a solely biometric feature space (on human speech signals, glancing at presentation attack detection). On signal processing, HMM speech alignment is secured for preserving privacy in *speech characterisation*; an alike data privacy preserving solution is directly employable to the HMM–UBM comparison in *speaker characterisation* (as the technology used is the same). It has also been shown that DNNs can be made secure through the use of HE, and that this framework can be applied to paralinguistic tasks.

Similarly to what was stated in the summary of Section 6, the techniques described in this Section can be extended to cover spoken language recognition tasks, in particular, the HE- or STPC-based DNN schemes, as well as hardware techniques, such as Intel SGX. However, to achieve state-of-the-art results, larger and deeper networks are necessary. For this reason, since the HE- or STPC-based DNN schemes are limited in the number and size of layers, further improvements are required before they can be applied to these tasks. Alternatively, hardware-based techniques could be directly applied to spoken language recognition.

Technology ensuring data privacy in non-biometric characterisation tasks are posing *anti-biometrics* as a new challenge to the speaker characterisation community. Moreover, methods originally proposed for privacy-preserving speech characterisation are immediately transferable to technology concerning speaker characterisation. Privacy needs to be preserved regarding *biometrics* and *anti-biometrics*, where future work needs to embrace common measures to evaluate the extent of preserved data privacy.

8. Evaluation measures on data privacy

Evaluating the privacy provided by an automated characterisation system, one needs to distinguish between *biometric* and *non-biometric* characterisation tasks. For the former, the privacy preservation of the biometric information is of relevance, whereas for the latter, the sanitisation of the data from biometric information is of relevance. Due to employing privacy safeguards, performance losses might arise in terms of *biometric discrimination* or *score calibration*. We suggest to report on $C_{\text{llr}}^{\text{min}}$ and C_{llr} (Brümmer and du Preez, 2008; Brümmer, 2010; Brümmer and de Villiers, 2011)—both metrics[26] emerged for measuring the performance of speaker verification systems, see Section 5—as well as on established performance measures within biometric standardisation (ISO/IEC JTC1 SC27 Security Techniques, 2011; ISO/IEC JTC1 SC37 Biometrics, 2017a). This section addresses the question: *how good is privacy preserved in biometric or non-biometric algorithms* instead. To ensure future work to report on concise metrics, this Section defines three metrics in more detail, particularly: the *unlinkability* in the biometric recognition task; the *degree of privacy* in non-biometric characterisation tasks; and the *degree of biometric evidence* in non-biometric characterisation tasks, quantifying the privacy adversary.

[26] For a binary classifier (decisions on classes $\mathcal{A}, \mathcal{B}$, e.g., same and different speaker) with score sets $S_{\mathcal{A}}, S_{\mathcal{B}}$, C_{llr} reports on the *goodness of LLRs* (Brümmer and du Preez, 2008; Brümmer, 2010; Brümmer and de Villiers, 2011). This embodies different interpretations: *(i)* C_{llr} represents the application-independent Bayes risk, *(ii)* measures the accuracy of the probabilistic class prediction as a strictly proper scoring rule, i.e., the accuracy of encoding all information of a comparison in the value of an LLR (the posterior score distribution is the same as the posterior evidence distribution), and *(iii)* is the generalised empirical cross-entropy (Brümmer and du Preez, 2008; Brümmer, 2010; Brümmer and de Villiers, 2011). For empirical cross-entropy between the output of a classifier and the ground-of-truth under maximum uncertainty, the term *generalised* refers here to a class prior probability of $\frac{1}{2}$, which is also known as *maximum uncertainty* (Brümmer and du Preez, 2008; Ramos and Gonzalez-Rodrigues, 2008). C_{llr} is defined via the sigmoid function $sigmoid(x) = (1 + e^{-x})^{-1}$, and the expectation operator $\langle \cdot \rangle$ (the mean value over a set of scores): $C_{\text{llr}}(S_{\mathcal{A}}, S_{\mathcal{B}}) = \frac{\langle -\log\, sigmoid(a) \rangle_{a \in S_{\mathcal{A}}} + \langle -\log\, sigmoid(-b) \rangle_{b \in S_{\mathcal{B}}}}{2 \log(2)}$. If all scores are well-calibrated, C_{llr} reaches its minimum $C_{\text{llr}}^{\text{min}}$.

8.1. Unlinkability in the biometric recognition task

As highlighted in Section 3, biometric information protection (BIP) schemes should not only maintain recognition accuracy with respect to their unprotected counterparts, but also fulfill three main requirements for protected information (ISO/IEC JTC1 SC27 Security Techniques, 2011), namely: (i) irreversibility, (ii) renewability, and (iii) unlinkability. Therefore, a standardised benchmark protocol for BIP schemes, in terms of recognition accuracy, security and privacy, is necessary to properly evaluate the performance of these techniques (Rane, 2014). Whereas the irreversibility has been widely studied in the literature, there is a lack of standardised metrics for the latter (e.g., no unlinkability metric was included in the recent ISO/IEC 30136 on performance testing of BIP schemes, ISO/IEC JTC1 SC37 Biometrics, 2018). To fill in this gap, a general framework for unlinkability evaluation is proposed (Gomez-Barrero et al., 2018). The main motivation behind the proposal in Gomez-Barrero et al. (2018) was to address the shortcomings of the existing approaches to measure the unlinkability of biometric information in terms of point estimates (i.e., templates, *not models*), which can be summarised as follows:

- unrealistic assumptions on uniformity of biometric data and the development of non general approaches for specific systems, based, for instance, on the determination of the adversary's advantage under the unlinkability (or indistinguishability) scenario (Simoens et al., 2009; Buhan et al., 2009; 2010),
- the consideration of linkability as a binary decision in contrast to a gradual property (Kholmatov and Yanikoglu, 2008; Simoens et al., 2009; Buhan et al., 2009; Kelkboom et al., 2011; Nagar et al., 2010; Piciucco et al., 2016),
- the lack of a quantitative measure, necessary for a comparative benchmark (Ferrara et al., 2014; Bringer et al., 2015; Wang and Hu, 2014), or
- the use of metrics employed for verification accuracy evaluations, not suitable for the linkability evaluation (Buhan et al., 2009; 2010; Kelkboom et al., 2011; Nagar et al., 2010; Piciucco et al., 2016; Rua et al., 2012), as it was shown in Gomez-Barrero et al. (2018).

Taking those remarks into account, both a local $D_{\leftrightarrow}(s)$ and a global $D_{\leftrightarrow}^{sys}$ linkability metric are proposed, for which a Python implementation is available.[27] $D_{\leftrightarrow}(s) \in [0, 1]$ evaluates the linkability of a system for each *specific linkage score* $s = LS(\mathbf{T}_1, \mathbf{T}_2)$. As such, this metric is appropriate to analyse within one system in which parts of the linkage score domain it fails to provide unlinkability. If for a specific score s_1, a system yields $D_{\leftrightarrow}(s_1) = 1$, it means that, in case the linkage function produced s_1, we would be able to link both templates $\mathbf{T}_1$ and $\mathbf{T}_2$ to the same instance with almost all certainty. On the other hand, $D_{\leftrightarrow}(s_0) = 0$ should be interpreted as full unlinkability for that particular score s_0. In other words, *if* s_0 were produced by the linkage function, it would be more likely that both templates stemmed from different instances, hence failing to link them to a single data subject. All intermediate values of $D_{\leftrightarrow}(s)$ between 0 and 1 report an increasing degree of linkability.

The key on the success of linking to templates lies on determining whether, given a score s, it is more likely that two templates stem from samples acquired from the same instance (i.e., mated samples, H_m) than from samples acquired from different instances (i.e., non-mated samples, H_{nm}): $p(H_m|s) > p(H_{nm}|s)$. Therefore, such linkability can be accounted for in terms of the following difference of conditional probabilities:

$$D_{\leftrightarrow}(s) = p(H_m|s) - p(H_{nm}|s)$$

However, these two conditional probabilities are unknown. Hence, the computation is carried out in terms of the likelihood ratio $LR(s) = p(s|H_m)/p(s|H_{nm})$ between the known probabilities, to reach the final metric definition:

$$D_{\leftrightarrow}(s) = \begin{cases} 0 & \text{if } LR(s)\cdot\omega \leq 1 \\ 2\dfrac{LR(s)\cdot\omega}{1+LR(s)\cdot\omega} - 1 & \text{if } LR(s)\cdot\omega > 1 \end{cases}$$

where $\omega = p(H_m)/p(H_{nm})$ denotes the ratio between the unknown prior probabilities of the *mated samples* and *non-mated samples* distributions. Thus, $D_{\leftrightarrow}(s) = 0$ refers to unlinkable score values where $p(H_m|s) \leq p(H_{nm}|s)$.

As mentioned above, it is also useful to have an estimation of the *unlinkability of the whole system*, which may allow a fairer benchmark of the unlinkability level of two or more systems. For this purpose, the global metric $D_{\leftrightarrow}^{sys} \in [0, 1]$ is

[27] https://dasec.h-da.de/research/biometrics/unlinkability and https://github.com/dasec/unlinkability-metric

introduced, which gives an estimation of the global linkability of a system, *independently* of the score. This way, if a system has $D_{\leftrightarrow}^{sys} = 1$ (i.e., case in which both the *mated samples* and *non-mated samples* distributions have no overlap), it means that it is fully linkable for all the scores of the *mated samples* distribution domain. Similarly, $D_{\leftrightarrow}^{sys} = 0$ means that the system is fully unlinkable for the whole score domain (i.e., full overlap of the distributions). In other words, independently of the score produced by the linkage function, it is equally probable that the two templates stem from the same instance (H_m) than from different instances (H_{nm}). All intermediate values of $D_{\leftrightarrow}^{sys}$ between 0 and 1 report an increasing degree of linkability.

Therefore, we are interested on measuring how likely it is to get a score stemming from the *mated samples* distribution. This can be achieved computing the difference $p(H_m \cap s) - p(H_{nm} \cap s)$ and integrating it. Regarding the success on linking templates, we are only interested in the probabilities stemming from the *mated samples* distribution, and two templates can be linked only if $p(H_m|s) > p(H_{nm}|s)$. Hence, $D_{\leftrightarrow}^{sys}$ is defined in terms of the local metric as:

$$D_{\leftrightarrow}^{sys} = \int p(s|H_m) \cdot D_{\leftrightarrow}(s) \mathrm{d}s$$

Using both metrics and the evaluation guidelines outlined in Gomez-Barrero et al. (2018), a quantitative and fair analysis of the unlinkability of the templates may be carried out. Arguably, the depicted measure cannot satisfy all demands of speaker characterisation technology, see Section 5 (e.g., on meta-embeddings), as this unlinkability measure is solely proposed for templates, and not for models. Notably, a definition accounting for feature precision is at-hand, as the outlined metric is based on LRs (also being the basis of speaker recognition methods relying on models rather than templates to characterise biometric information in speech data).

8.2. Degree of privacy in non-biometric characterisation tasks

Considering privacy in non-biometric characterisation tasks, where a generic model is trained to give a response for some task, the participation of a particular individual should not affect the output of a training algorithm A. Thus, differential privacy provides a good framework to quantify the degree of privacy, that is when the algorithm training is completed and the converged model is put under a data privacy test. In fact, we would like the output model given using the data set $\mathcal{D}$ to be indistinguishable from the output model obtained from a data set by augmenting training set $\mathcal{D}$ with the data of another individual $\mathcal{I}$ (i.e., $\mathcal{D} + \mathcal{I}$). For achieving differential privacy in machine learning algorithms, we need to introduce stochastic components to the algorithm A, so that the output model of the algorithm can be described as a probability distribution. With this, two outputs are indistinguishable if their corresponding probability distribution are. In other words, $\Pr(A(\mathcal{D}))$ should marginally differ from $\Pr(A(\mathcal{D} + \mathcal{I}))$. Analysing different possible configurations of $\mathcal{D}$ and $\mathcal{D} + \mathcal{I}$, it is possible to quantify algorithms A by moving the purpose of the inequality in differential privacy from training to testing (i.e., to an equality):

$$\epsilon = \max_{\mathcal{D}, \mathcal{D}+\mathcal{I}} \log \frac{p(A(\mathcal{D}))}{p(A(\mathcal{D} + \mathcal{I}))} \,. \tag{7}$$

The parameter ϵ can then be used to quantify the degree of privacy of the model given by algorithm A. Some examples of these methods are presented in Section 7.3.

In addition, thanks to the composition theorem, the degree of privacy remains (at the moment of performing a prediction). The leakage can then be controlled just based on the training procedure. This means that we cannot distinguish between $\mathcal{D}$ and $\mathcal{D} + \mathcal{I}$ analysis evaluation of the outputs models. In other words, algorithms are ϵ-*indistinguishable* with low values indicating a high degree of privacy.

8.3. Degree of biometric evidence in non-biometric characterisation tasks, quantifying the privacy adversary

The aim of data Privacy by Design in characterisation systems is to yield features, discriminative for a non-biometric task but not conveying any biometric information. With the goal of examining to which extent data of non-biometric applications is biometrically linkable, we propose to employ a biometric adversary, whose score outputs are measured regarding their evidential support towards making biometric decisions whether two feature representations are *mated* (proposition $\mathcal{A}$) or *non-mated* (proposition $\mathcal{B}$). By *biometric evidence*, we refer to *LLR scores*

$S = \log \Pr(E \mid \frac{\mathcal{A})}{\Pr(E} \mid \mathcal{B})$ yielded by a biometric system. Ideally, the biometric adversary estimates LLRs directly by a subspace identity model, such as Gaussian PLDA (Garcia-Romero and Epsy-Wilson, 2011), see Section 5; its Gaussian assumption is for the latent subspace, not necessarily for the observed data. By employing another biometric adversary, its scores are convertible to LLRs by *ideal score calibration* (Brümmer and du Preez, 2009). Features of non-biometric algorithms (preserving Privacy by Design) cannot yield *biometric evidence*, i.e., result in *zero-evidence* LLRs: $S \approx 0$, such that biometric adversaries are *decepted*. In systems not preserving privacy, decisions about $\mathcal{A}, \mathcal{B}$ are inferable in terms of LLRs either supporting $\mathcal{A}$ or $\mathcal{B}$, and, by examining multiple LLRs for a dataset, clusters of mated and non-mated features become inferable; being a data privacy exploit. Thus, the maximum LLR depicts the *empirical upper bound* to the *degree of biometric evidence* still present in a non-biometric system's features. For the sake of a unified reporting, we propose to employ the reference implementation of Gaussian PLDA (Garcia-Romero and Epsy-Wilson, 2011) directly on the assessed data to derive ideal LLRs. Thereby, positive LLRs support proposition $\mathcal{A}$ and negative LLRs support proposition $\mathcal{B}$; thus, the maximum of the absolute LLR values serves as the upper bound (i.e., $\max |S|$). One may transform this bound, living in $(0, +\infty)$, to the [0,1] domain by employing the *sigmoid* function. The choice of the sigmoid function is motivated by the nature of LLR thresholds (Brümmer, 2010; Brümmer and de Villiers, 2011), which are transformed probabilities (that model the degree of belief in a decision requirement) represented as *logistic odds* (i.e., *log-odds*). Probabilities are transformed to log-odds by the *logit* function with the *sigmoid* function being its inverse function (i.e., the required evidence support a LLR satisfies is expressed in probabilistic terms by the sigmoid-transformed LLR: *sigmoid*(S)). The *degree of biometric evidence* $\varsigma \in [0, 1]$ is:

$$\varsigma = sigmoid\left(\max \left| \log \frac{\Pr(E \mid \mathcal{A})}{\Pr(E \mid \mathcal{B})} \right| \right). \tag{8}$$

In other words, *given a biometric adversary* e.g., Garcia-Romero and Epsy-Wilson (2011), ς is a probabilistic representation of the maximal *degree of biometric evidence* still persistent within the features of a non-biometric system. Low ς values represent a low degree of biometric evidence (i.e., high data privacy in non-biometric applications).

9. Discussion and conclusion

Voice data can be used not only for purposes such as *speech* characterisation, but also for applications that characterise the *speaker*. As such, all speech technologies should preserve privacy to accommodate latest EU but also US legislation (e.g., see the GDPR, but also the Illinois Law, the California Consumer Privacy Act as described in Section 2.2.1); speaker characterisation systems should protect biometric information while speech characterisation systems demand signal pre-processing techniques that suppress biometric information. Many different approaches to privacy preservation have been proposed in the context of speaker characterisation, ranging from voice binarisation to homomorphic encryption, with each technique offering different compromises between data security and privacy on the one hand, and memory requirements and computational and communication complexity on the other hand. Approaches to privacy preservation in the case of speech characterisation include computation in entrusted units, fully homomorphic inference systems, and differential private learning. Be it speaker or speech characterisation, the concept of *Privacy by Design* should be applied by default so that privacy is preserved in the full set of processing components that make up a given application (rather than being applied to a subset of processing components). Only this approach can ensure reliable *data protection* and hence the preservation of *subjects' privacy*.

Biometrics researchers have made great strides in meeting the requirements for privacy preservation. This progress is largely based upon the application of encryption and secure computation technologies borrowed from the cryptography community. Standard approaches to data protection have proved readily applicable to some forms of biometrics, especially those that have the benefit of being able to rely upon the extraction or estimation of high-quality, consistent biometric templates, e.g., fingerprints, for which intersession variation is typically low. Unfortunately, the application of these same techniques to speech signals has proven to be considerably more challenging. The difficulty stems principally from the use of features and models that are designed to accommodate the inherent variability in speech signals. As a result, most speaker and speech characterisation solutions rarely utilise anything resembling a template or *print* in the same sense as other biometric technologies. Rather than making decisions in the observation domain, decisions are instead inferred from *models* estimated not from a single observation (speech utterance), but

from a number of such observations that can be used to accommodate observation variability. As a result of this inherent variability, standard approaches to cryptography must be adapted so that they can be applied to preserve privacy in speech signals. Clearly, the development of privacy-preserving solutions for speech signals is extremely complex. It also requires a reflection upon other perspectives beyond pure technology.

This survey on privacy preservation for speaker and speech characterisation captures the perspectives of the legal, biometrics, cryptography, speaker, and speech characterisation communities. The collection of all of these different perspectives in data Privacy by Design, accompanied with a proposal for harmonised evaluation measures (on the biometric unlinkability, the degree of privacy, and the remaining degree of biometric evidence), aims to stimulate the cross-fertilisation of ideas and collaborative research that will be needed in the future to meet the challenges of delivering privacy preservation for speech signals. Arguably, progress will only emerge from more intensive collaboration between the cryptography and speech technology communities. In the first instance, this effort should establish a common understanding in terms of vocabulary harmonisation and technology tutorials. Technology must also be made more transparent so as to support legal research. All of this work will help to avoid the pitfalls leading to misunderstanding and confusion that currently hamper progress (e.g., references in the wider, non-specialist literature to *voiceprints*).

In terms of technology, future research should prioritise the development of computationally feasible solutions to *matrix inversion* and *log-determinants* and the definition of effective *data protection safeguards*. Considering the progress of the cryptography community towards *quantum computing* (which ultimately would make *integer factorisation* computationally feasible, i.e., the presented approaches are linearly crackable), the need for *post-quantum secure* safeguards is inevitable. Moreover, privacy-preserving comparison and inference systems will need to meet real-time requirements. This challenge is compounded by the use of cohort score normalisation, e.g., in speaker verification, and is especially relevant to biometric identification applications, two examples in which a single decision requires multiple biometric comparison operations. The development of more efficient and usable secure floating point computation protocols will be crucial to meeting this goal. Eventually, the consideration of data privacy must be accompanied by a consideration of security. Security will be critical, e.g., in an effort to prevent attacks aimed at gaining access to an end-user's voice capturing device. The penetration of biometric systems at this level may leave them vulnerable to the injection of encrypted comparison scores in a similar manner as existing systems can be vulnerable to voice presentation attacks. In a similar vein, system protocols need to accommodate not only the data privacy of end-users and the security of vendor model parameters, but also the integrity of comparison scores computed in distributed systems.

Further work is also needed in the legal dimension. Beyond what is said and who says it, several additional levels of information are embedded in voice signals. Examples include *paralinguistic* or *extralinguistic* information. While the concept of *sensitive data* may have a clear legal definition, the scope can be somewhat limited, e.g., identity, racial or ethnic origin, political opinions, religious or philosophical beliefs and health-related information. The new era of artificial intelligence is seeing the emergence of diverse, new applications that exploit voice recordings. Many of these may not fall under the existing definition of sensitive data. Examples of such applications[28] and their potential impact upon privacy call for the existing definitions of privacy and protection to be revisited.

Acknowledgements

This work is partially supported by: the German Federal Ministry of Education and Research (BMBF) and the Hessen State Ministry for Higher Education, Research and the Arts within the German National Research Center for Applied Cybersecurity (www.crisp-da.de); the *BioMobile II* project (no. 518/16-30); the DFG as part of project E4 within the CRC 1119 CROSSING and project A.1 within the RTG 2050 Privacy and Trust for Mobile Users; the Horizon 2020 Victoria research project (grant agreement SEC-740754); research funds received from the French Agence Nationale de la Recherche-pl2X-sim-(ANR) in connection with the bilateral VoicePersonae (with JST CREST in Japan); RESPECT (with DFG in Germany) collaborative research projects; national funds in Portugal through Fundaçao para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013; the SpeechXRays project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no 653586; the Privacy Preserving Secure Speech Recognition (PPSSR) project which had received

[28] "The Dangerous Junk Science of Vocal Risk Assessment", https://theintercept.com/2018/11/25/voice-risk-analysis-ac-global/

funding from the European Union's H2020-INNOSUP-02-2016 (Innovation Associate Programme) under grant agreement no 739767; and by Omilia – Conversational Intelligence.

References

Abad, A., Ribeiro, E., Kepler, F., Astudillo, R.F., Trancoso, I., 2016. Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 308–318.

Adjedj, M., Bringer, J., Chabanne, H., Kindarji, B., 2009. Biometric identification over encrypted data made feasible. In: Proceedings of the International Conference on Information Systems Security (ICISS), pp. 86–100.

Adler, A., 2003. Sample images can be independently restored from face recognition templates. In: Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1163–1166.

Agrawal, S., Gorbunov, S., Vaitkuntanathan, V., Wee, H., 2013. Functional encryption: new perspectives and lower bounds. In: Proceedings of the Annual International Cryptology Conference (CRYPTO), pp. 500–518.

Aguilar-Melchor, C., Fau, S., Fontaine, C., Gogniat, G., Sirdey, R., 2013. Recent advances in homomorphic encryption: a possible future for signal processing in the encrypted domain. IEEE Signal Process. Mag. 30, 108–117.

Aliasgari, M., Blanton, M., Zhang, Y., Steele, A., 2013. Secure computation on floating point numbers. In: Proceedings of the Network and Distributed System Security Symposium (NDSS).

Anguera, X., Bonastre, J.-F., 2010. A novel speaker binary key derived from anchor models. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2118–2121.

Anguera, X., Bonastre, J.-F., 2011. Fast speaker diarization based on binary keys. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4428–4431.

Asharov, G., Lindell, Y., Schneider, T., Zohner, M., 2013. More efficient oblivious transfer and extensions for faster secure computation. In: Proceedings of the ACM SIGSAC Conference on Computer & Communications Security (CCS), pp. 535–548.

Bahmani, R., Barbosa, M., Brasser, F., Portela, B., Sadeghi, A.-R., Scerri, G., Warinschi, B., 2017. Secure multiparty computation from SGX. In: Proceedings of the Financial Cryptography and Data Security (FC), pp. 477–497.

Barak, A., Hirt, M., Koskas, L., Lindell, Y., 2018. An end-to-end system for large scale P2P MPC-as-a-service and low-bandwidth MPC for weak participants. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 695–712 https://github.com/cryptobiu/MATRIX .

Barni, M., Bianchi, T., Catalano, D., et al., 2010. A privacy-compliant fingerprint recognition system based on homomorphic encryption and fingercode templates. In: Proceedings of the International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 1–7.

Barni, M., Failla, P., Lazzeretti, R., Sadeghi, A.-R., Schneider, T., 2011. Privacy-preserving ECG classification with branching programs and neural networks. IEEE Trans. Inf. Forensics Secur. (TIFS) 6 (2), 452–468.

Bellare, M., Desai, A., Pointcheval, D., Rogaway, P., 1998. Relations among notions of security for public-key encryption schemes. In: Proceedings of the Annual International Cryptology Conference (CRYPTO), pp. 26–45.

Bellare, M., Hoang, V.T., Keelveedhi, S., Rogaway, P., 2013. Efficient garbling from a fixed-key blockcipher. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P). IEEE, pp. 478–492.

Bernstein, D.J., Buchmann, J., Dahmen, E., 2009. Post-Quantum Cryptography. Springer Science & Business Media.

Bianchi, T., Turchi, S., Piva, A., et al., 2010. Implementing fingercode-based identity matching in the encrypted domain. In: Proceedings of the Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), pp. 15–21.

Billeb, S., Rathgeb, C., Buschbeck, M., Reininger, H., Kasper, K., 2014. Efficient two-stage speaker identification based on universal background models. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–6.

Billeb, S., Rathgeb, C., Reininger, H., Kasper, K., Busch, C., 2015. Biometric template protection for speaker recognition based on universal background models. IET Biomet. 4 (2), 116–126.

Bimbot, F., Chollet, G., 1997. Assessment of Speaker Verification Systems. De Gruyter, pp. 408–480.

Bishop, A., Jain, A., Kowalczyk, L., 2015. Function-hiding inner product encryption. In: Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 470–491.

Blanton, M., Aliasgari, M., 2012. Secure outsourced computation of iris matching. J. Comput. Secur. (JoCS) 20 (2-3), 259–305.

Blanton, M., Gasti, P., 2011. Secure and efficient protocols for iris and fingerprint identification. In: Proceedings of the European Symposium on Research in Computer Security (ESORICS), pp. 190–209.

Bonastre, J.F., Bousquet, P.M., Matrouf, D., Anguera, X., 2011. Discriminant binary data representation for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5284–5287. doi: 10.1109/ICASSP.2011.5947550.

Boneh, D., Di Crescenzo, G., Ostrovsky, R., Persiano, G., 2004. Public key encryption with keyword search. In: Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), pp. 506–522.

Boneh, D., Franklin, M., 2001. Identity-based encryption from the Weil pairing. In: Proceedings of the Annual International Cryptology Conference (CRYPTO), pp. 213–229.

Boneh, D., Sahai, A., Waters, B., 2011. Functional encryption: definitions and challenges. In: Proceedings of the Theory of Cryptography Conference (TCC), pp. 253–273.

Boneh, D., Waters, B., 2007. Conjunctive, subset, and range queries on encrypted data. In: Proceedings of the Theory of Cryptography Conference (TCC), pp. 535–554.
Boufounos, P., Rane, S., 2011. Secure binary embeddings for privacy preserving nearest neighbors. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS).
Bourse, F., Minelli, M., Minihold, M., Paillier, P., 2017. Fast homomorphic evaluation of deep discretized neural networks. IACR Cryptol. ePrint Arch. 2017, 1114.
Boë, L.-J., 2000. Forensic voice identification in France. Speech Commun. 31, 205–224.
Brasser, F., Frassetto, T., Riedhammer, K., Sadeghi, A.-R., Schneider, T., Weinert, C., 2018. VoiceGuard: secure and private speech processing. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1303–1307.
Bridle, J.S., Brown, M.D., 1974. An Experimental Automatic Word-Recognition System. JSRU Report. Joint Speech Research Unit, Ruislip, England.1003
Bringer, J., Chabanne, H., Favre, M., Patey, A., Schneider, T., Zohner, M., 2014. GSHADE: faster privacy-preserving distance computation and biometric identification. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), pp. 187–198.
Bringer, J., Morel, C., Rathgeb, C., 2015. Security analysis of bloom filter-based iris biometric template protection. In: Proceedings of the IEEE International Conference on Biometrics (ICB), pp. 527–534.
Brümmer, N., 2010. Measuring, Refining and Calibrating Speaker and Language Information Extracted From Speech. University of Stellenbosch Ph.D. thesis.
Brümmer, N., Burget, L., Garcia, P., Plchot, O., Rhodin, J., et al., 2017. Meta-Embeddings: A Probabilistic Generalization of Embeddings in Machine Learning. Technical Report. JHU HLTCOE 2017 SCALE Workshop. https://github.com/bsxfan/meta-embeddings/tree/master/theory.
Brümmer, N., de Villiers, E., 2011. The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing. Technical Report. AGNITIO Research, South Africa.
Brümmer, N., du Preez, J., 2008. Application-independent evaluation of speaker detection. Comput. Speech Lang. (CSL) 20 (2), 230–275.
Brümmer, N., du Preez, J., 2009. The PAV Algorithm Optimizes Binary Proper Scoring Rules. Technical Report. Agnitio. http://niko.brummer.googlepages.com.
Brümmer, N., Silnova, A., Burget, L., Stafylakis, T., 2018. Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model. In: Proceedings of the Odyssey 2018: The Speaker and Language Recognition Workshop, pp. 349–356.
Buhan, I., Breebaart, J., Guajardo, J., et al., 2009. A quantitative analysis of indistinguishability for a continuous domain biometric cryptosystem. In: Proceedings of the International Conference on Data Privacy Management and Autonomous Spontaneous Security (DPM/SETOP), pp. 78–92.
Buhan, I., Merchan, J., Kelkboom, E., 2010. Efficient strategies for playing the indistinguishability game for fuzzy sketches. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS).
Security and Privacy in Biometrics. In: Campisi, P. (Ed.), Springer.
Cappelli, R., Maio, D., Lumini, A., Maltoni, D., 2007. Fingerprint image reconstruction from standard templates. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 29 (9), 1489–1503.
Cash, D., Grubbs, P., Perry, J., Ristenpart, T., 2015. Leakage-abuse attacks against searchable encryption. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS).
Cash, D., Jaeger, J., Jarecki, S., Jutla, C., Krawczyk, H., Rosu, M., Steiner, M., 2014. Dynamic searchable encryption in very-large databases: data structures and implementation. In: Proceedings of the Network and Distributed System Security Symposium (NDSS).
Cash, D., Jarecki, S., Jutla, C., Krawczyk, H., Roşu, M., Steiner, M., 2013. Highly-scalable searchable symmetric encryption with support for boolean queries. In: Proceedings of the Annual Cryptology Conf. (CRYPTO).
Cavoukian, A., Stoianov, A., 2011. Biometric encryption. In: Encyclopedia of Cryptography and Security. Springer, pp. 90–98.
Chabanne, H., de Wargny, A., Milgram, J., Morel, C., et al., 2017. Privacy-preserving classification on deep neural network. IACR Cryptol. ePrint Arch. 2017, 35.
Chun, H., Elmehdwi, Y., Li, F., Bhattacharya, P., Jiang, W., 2014. Outsourceable two-party privacy-preserving biometric authentication. In: Proceedings of the ACM ASIA Conference on Computer and Communications Security (ASIACCS), pp. 401–412.
Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., Wang, T., 2018. Privacy at scale: local differential privacy in practice. In: Proceedings of the ACM SIGMOD/PODS International Conference on Management of Data (SIGMOD/PODS), pp. 1655–1658.
Costa, L., Poullet, Y., 2012. Privacy and the regulation of 2012. Comput. Law Secur. Rev. 28 (3), 254–262.
Costan, V., Devadas, S., 2016. Intel SGX explained. IACR Cryptol. ePrint Arch. 2016, 086.
Cumani, S., 2015. Fast scoring of full posterior PLDA models. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 23 (11), 2036–2045.
Cumani, S., Brümmer, N., Burget, L., Laface, P., Plchot, O., Vasilakakis, V., 2013. Pairwise discriminative speaker verification in the i-vector space. IEEE Trans. Audio Speech Lang. Process. (TASLP) 21 (6), 1217–1227.
Curtmola, R., Garay, J., Kamara, S., Ostrovsky, R., 2006. Searchable symmetric encryption: Improved definitions and efficient constructions. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Securitys (CCS).
Damgård, I., Jurik, M., 2001. A generalisation, a simplification and some applications of Paillier's probabilistic public-key system. In: Proceedings of the International Workshop on Practice and Theory in Public Key Cryptosystems (PKC).
Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences. Trans. Acoust. Speech Signal Process. (ASSP) 28 (4), 357–366.
Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. (TASLP) 19 (4), 788–798.
Delgado, H., Anguera, X., Fredouille, C., Serrano, J., 2015. Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEE Trans. Audio Speech Lang. Process. (TASLP) 23 (12), 2286–2297.

Demmler, D., Dessouky, G., Koushanfar, F., Sadeghi, A.-R., Schneider, T., Zeitouni, S., 2015a. Automated synthesis of optimized circuits for secure computation. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 1504–1517.

Demmler, D., Schneider, T., Zohner, M., 2015b. ABY − a framework for efficient mixed-protocol secure two-party computation. In: Proceedings of the Network and Distributed System Security Symposium (NDSS).

Dias, M., Abad, A., Trancoso, I., 2018. Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Dwork, C., 2006. Differential privacy. In: Proceedings of the International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006), pp. 1–12.

Dwork, C., Roth, A., et al., 2014. The algorithmic foundations of differential privacy. Found. Trends® in Theor. Comput. Sci. (TCS) 9 (3–4), 211–407.

ElGamal, T., 1984. A public key cryptosystem and a signature scheme based on discrete logarithms. In: Proceedings of the Workshop on the Theory and Application of Cryptographic Techniques (ASIACRYPT), pp. 10–18.

Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I., Toft, T., 2009. Privacy-preserving face recognition. In: Proceedings of the International Symposium on Privacy Enhancing Technologies Symposium (PETS), pp. 235–253.

European Parliament and Council, 2016a. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

European Parliament and Council, 2016b. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons 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.

Evans, D., Huang, Y., Katz, J., Malka, L., 2011. Efficient privacy-preserving biometric identification. In: Proceedings of the Network and Distributed System Security Symposium (NDSS).

Ferrara, M., Maltoni, D., Cappelli, R., 2014. A two-factor protection scheme for MCC fingerprint templates. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG).

Galbally, J., Ross, A., Gomez-Barrero, M., Fierrez, J., Ortega-Garcia, J., 2013. Iris image reconstruction from binary templates: an efficient probabilistic approach based on genetic algorithms. Comput. Vis. Image Underst. (CVIU) 117 (10), 1512–1525.

García, J.A.G., Moro-Velázquez, L., Godino-Llorente, J.I., Castellanos-Domínguez, G., 2015. Automatic age detection in normal and pathological voice. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Garcia-Romero, D., Epsy-Wilson, C., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, pp. 249–252.

Gentry, C., 2009. Fully homomorphic encryption using ideal lattices. In: Proceedings of the ACM Symposium on Theory of Computing (STOC), pp. 169–178.

Gilad-Bachrach, R., Dowlin, N., Laine, K., et al., 2016. CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In: Proceedings of the JMLR International Conference on Machine Learning (ICML), 48, pp. 201–210.

Glackin, C., Chollet, G., Dugan, N., Cannings, N., Wall, J., Tahir, S., Ray, I.G., Rajarajan, M., 2017. Privacy preserving encrypted phonetic search of speech data. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6414–6418.

Glembek, O., Burget, L., Dehak, N., Bümmer, N., Kenny, P., 2009. Comparison of scoring methods used in speaker recognition with joint factor analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4057–4060.

Gomez-Barrero, M., Fierrez, J., Galbally, J., Maiorana, E., Campisi, P., 2016. Implementation of fixed length template protection based on homomorphic encryption with application to signature biometrics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 191–198.

Gomez-Barrero, M., Galbally, J., Morales, A., Ferrer, M.A., Fierrez, J., Ortega-Garcia, J., 2014. A novel hand reconstruction approach and its application to vulnerability assessment. Inf. Sci. 268, 103–121.

Gomez-Barrero, M., Galbally, J., Morales, A., Fierrez, J., 2017a. Privacy-preserving comparison of variable-length data with application to biometric template protection. IEEE Access 5 (1), 8606–8619.

Gomez-Barrero, M., Galbally, J., Rathgeb, C., Busch, C., 2018. General framework to evaluate unlinkability in biometric template protection systems. IEEE Trans. Inf. Forensics Secur. (TIFS) 3 (6), 1406–1420.

Gomez-Barrero, M., Maiorana, E., Galbally, J., Campisi, P., Fierrez, J., 2017b. Multi-biometric template protection based on homomorphic encryption. Pattern Recognit. 67, 149–163.

González, D.M., Plchot, O., Burget, L., Glembek, O., Matejka, P., 2011. Language recognition in i-vectors space. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Gupta, D., Mood, B., Feigenbaum, J., Butler, K., Traynor, P., 2016. Using Intel software guard extensions for efficient two-party secure function evaluation. In: Proceedings of the Workshop on Encrypted Computing and Applied Homomorphic Cryptography (WAHC).

Gürses, S., Troncoso, C., Diaz, C., 2011. Engineering privacy by design. In: Proceedings of the Computers, Privacy and Data Protection (CPDP).

Haderlein, T., Middag, F., Hönig, C., Martens, J.-P., Döllinger, M., Schützenberger, A., Nöth, E., 2015. Language-Independent Age Estimation From Speech Using Phonological and Phonemic Features. Springer Lecture Notes in Artificial Intelligence (LNAI), 9302. Springer, pp. 165–173.

Hansen, J.H.L., Hasan, T., 2015. Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process. Mag. 32 (6), 74–99.

Harb, H., Chen, L., 2005. Voice-based gender identification in multimedia applications. J. Intell. Inf. Syst. (JIIS) 24 (2), 179–198.

Hastings, M., Hemenway, B., Noble, D., Zdancewic, S., 2019. SoK: general-purpose compilers for secure multi-party computation. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P). to appear.

Hernandez-Sierra, G., Calvo, J.R., Bonastre, J.-F., Bousquet, P.-M., 2014. Session compensation using binary speech representation for speaker recognition. Pattern Recognit. Lett. 49, 17–23.

Hesamifard, E., Takabi, H., Ghasemi, M., 2017. CryptoDL: deep neural networks over encrypted data. Comput. Res. Repos. (CoRR) abs/1711.05189.
Hoepman, J.H., 2013. Privacy design strategies. In: Proceedings of the Privacy Law Scholars Conference (PLSC).
Hoffstein, J., Pipher, J., Silverman, J.H., 1998. NTRU: a ring-based public key cryptosystem. In: Proceedings of the International Algorithmic Number Theory Symposium (ANTS), pp. 267–288.
Hu, S., Li, M., Wang, Q., Chow, S.S., Du, M., 2018. Outsourced biometric identification with privacy. IEEE Trans. Inf. Forensics Secur. (TIFS) 13 (10), 2448–2463.
IEEE Standards Association, 2008. 754-2008 IEEE standard for Floating-Point Arithmetic.
Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. Comput. Res. Repos. (CoRR) abs/1502.03167.
Ishai, Y., Kilian, J., Nissim, K., Petrank, E., 2003. Extending oblivious transfers efficiently. In: Proceedings of the Annual International Cryptology Conference (CRYPTO), pp. 145–161.
Islam, M.S., Kuzu, M., Kantarcioglu, M., 2012. Access pattern disclosure on searchable encryption: ramification, attack and mitigation. In: Proceedings of the Network and Distributed System Security Symposium (NDSS).
ISO/CASCO Committee on Conformity Assessment, 2017. ISO/IEC 17025:2017. General Requirements for the Competence of Testing and Calibration Laboratories. International Organization for Standardization.
ISO/IEC JTC1 SC27 Security Techniques, 2011. ISO/IEC 24745:2011. Information Technology − Security Techniques − Biometric Information Protection. International Organization for Standardization.
ISO/IEC JTC1 SC37 Biometrics, 2017. ISO/IEC 19795-1:2017. Information Technology − Biometric Performance Testing and Reporting − Part 1: Principles and Framework. International Organization for Standardization and International Electrotechnical Committee.
ISO/IEC JTC1 SC37 Biometrics, 2017. ISO/IEC 2382-37:2017 Information Technology − Vocabulary − Part 37: Biometrics. International Organization for Standardization.
ISO/IEC JTC1 SC37 Biometrics, 2018. ISO/IEC 30136:2018. Information Technology − Performance Testing of Biometric Template Protection schemes. International Organization for Standardization.
Jasserand, C., 2016. Legal nature of biometric data: from 'generic' personal data to sensitive data: which changes does the new data protection framework introduce? Eur. Data Protect. Law Rev. 2 (3), 297–311.
Jiménez, A., Raj, B., 2017a. Privacy preserving distance computation using somewhat-trusted third parties. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6399–6403.
Jiménez, A., Raj, B., 2017b. A two factor transformation for speaker verification through ℓ_1 comparison. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6.
Jiménez, A., Raj, B., Portêlo, J., Trancoso, I., 2015. Secure modular hashing. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6.
Juvekar, C., Vaikuntanathan, V., Chandrakasan, A., 2018. GAZELLE: a low latency framework for secure neural network inference. In: Proceedings of the USENIX Security Symposium (USENIX Security).
Kamara, S., Papamanthou, C., Roeder, T., 2012. Dynamic searchable symmetric encryption. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS).
Kanade, S.G., Petrovska-Delacrétaz, D., Dorizzi, B., 2012. Enhancing information security and privacy by combining biometrics with cryptography. Synth. Lect. Inf. Secur. Priv. Trust (SPT) 3 (1), 1–140.
Katz, J., Lindell, Y., 2014. Introduction to Modern Cryptography. Chapman and Hall/CRC.
Katz, J., Sahai, A., Waters, B., 2008. Predicate encryption supporting disjunctions, polynomial equations, and inner products. In: Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), pp. 146–162.
Kelkboom, E.J., Breebaart, J., Kevenaar, T.A., Buhan, I., Veldhuis, R.N., 2011. Preventing the decodability attack based cross-matching in a fuzzy commitment scheme. IEEE Trans. Inf. Forensics Secur. (TIFS) 6 (1), 107–121.
Kenny, P., 2005. Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Technical Report. CRIM, Montreal. CRIM-06/08-13
Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process. (TASLP) 15 (4), 1435–1447.
Kholmatov, A., Yanikoglu, B., 2008. Realization of correlation attack against the fuzzy vault scheme. In: Proceedings of the SPIE Security, Forensics, Steganography, and Watermarking of Multimedia Contents X.
Kim, S., Lewi, K., Mandal, A., Montgomery, H., Roy, A., Wu, D.J., 2018. Function-hiding inner product encryption is practical. In: Proceedings of the International Conference on Security and Cryptography for Networks (SCN), pp. 544–562.
Kindt, E., 2018. Having yes, using no? About the new legal regime for biometric data. Comput. Law Secur. Rev. 34 (3), 523–538.
Kindt, E., 2019. A legal perspective on the relevance of biometric presentation attack detection (PAD) for payment services under PSDII and the GDPR. In: Marcel, S., Nixon, M., Fierrez, J., Evans, N. (Eds.), Handbook of Biometric Anti-Spoofing − Presentation Attack Detection. second ed. Advances in Computer Vision and Pattern Recognition. Springer 26. 17 pages (in print).
Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. 52 (1), 12–40.
Klitou, D., 2014. Privacy-Invading Technologies and Privacy by Design − Safeguarding Privacy, Liberty and Security in the 21st Century. Springer.
Koeberl, P., Phegade, V., Rajan, A., Schneider, T., Schulz, S., Zhdanova, M., 2015. Time to rethink: trust brokerage using trusted execution environments. In: Proceedings of the Trust and Trustworthy Computing (TRUST), pp. 181–190.
Kolesnikov, V., Schneider, T., 2008. Improved garbled circuit: free XOR gates and applications. In: Proceedings of the International Colloquium on Automata, Languages, and Programming (ICALP), pp. 486–498.

Lessig, L., 2006. Code, Version 2.0. Basic Books.
Lindell, Y., 2017. How to simulate it – a tutorial on the simulation proof technique. Tutorials on the Foundations of Cryptography. Springer, pp. 277–346.
Lindell, Y., Pinkas, B., 2009. A proof of security of Yao's protocol for two-party computation. J. Cryptol. (JoC) 161–188.
Lindell, Y., Pinkas, B., 2012. Secure two-party computation via cut-and-choose oblivious transfer. J. of Cryptol. (JoC) 25 (4), 680–722.
Lindell, Y., Waisbard, E., 2010. Private web search with malicious adversaries. In: Proceedings of the International Symposium on Privacy Enhancing Technologies Symposium (PETS). Springer, pp. 220–235.
Liu, J., Juuti, M., Lu, Y., Asokan, N., 2017. Oblivious neural network predictions via MiniONN transformations. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 619–631.
Lu, R., Zhu, H., Liu, X., Liu, J.K., Shao, J., 2014. Toward efficient and privacy-preserving computing in big data era. IEEE Netw. 28 (4), 46–50.
Luque, J., Anguera, X., 2014. On the modeling of natural vocal emotion expressions through binary key. In: Proceedings of the European Signal Processing Conference (EUSIPCO), pp. 1562–1566.
Malkhi, D., Nisan, N., Pinkas, B., Sella, Y., et al., 2004. Fairplay – a secure two-party computation system. In: Proceedings of the USENIX Security Symposium.
McKeen, F., Alexandrovich, I., Berenzon, A., Rozas, C.V., Shafi, H., Shanbhogue, V., Savagaonkar, U.R., 2013. Innovative instructions and software model for isolated execution. In: Proceedings of the Workshop on Hardware and Architectural Support for Security and Privacy (HASP).
Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., Natale, C.D., 2014. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowl.-Based Syst. 63, 68–81.
Meuwly, D., Ramos, D., Haraksim, R., 2017. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci. Int. 276, 142–153.
Microsoft ResearchRedmond, WA., 2018. Simple Encrypted Arithmetic Library (release 3.0.0). http://sealcrypto.org.
Mohassel, P., Zhang, Y., 2017. SecureML: a system for scalable privacy-preserving machine learning. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 19–38.
Mokhtar, S.B., Boutet, A., Felber, P., Pasin, M., Pires, R., Schiavoni, V., 2017. X-search: revisiting private web search using Intel SGX. In: Proceedings of the ACM/IFIP/USENIX Middleware Conference, pp. 198–208.
Mtibaa, A., Petrovska-Delacretaz, D., Hamida, A.B., 2018. Cancelable speaker verification system based on binary Gaussian mixtures. In: Proceedings of the Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1–6.
Nagar, A., Nandakumar, K., Jain, A.K., 2010. Biometric template transformation: a security analysis. In: Proceedings of the SPIE Media Forensics and Security II.
Nautsch, A., Isadskiy, S., Kolberg, J., Gomez-Barrero, M., Busch, C., 2018. Homomorphic encryption for speaker recognition: protection of biometric templates and vendor model parameters. In: Proceedings of the Speaker and Language Recognition Workshop (Odyssey). ISCA, pp. 16–23.
Oppenheim, A.V., Schafer, R.W., 1968. Homomorphic analysis of speech. IEEE Trans. Audio Electroacoust. (AU) 16 (2), 221–226.
Osadchy, M., Pinkas, B., Jarrous, A., Moskovich, B., 2010. SCiFi – a system for secure face identification. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 239–254.
Paillier, P., 1999. Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT). Springer, pp. 223–238.
Paillier, P., Pointcheval, D., 1999. Efficient public-key cryptosystems provably secure against active adversaries. In: Proceedings of the Annual International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 165–179.
Patel, V.M., Ratha, N., Chellappa, R., 2015. Cancelable biometrics: a review. IEEE Signal Process. Mag. 32 (5), 54–65.
Pathak, M., Portêlo, J., Raj, B., Trancoso, I., 2012. Privacy-preserving speaker authentication. In: Proceedings of the International Conference on Information Security (ISC), pp. 1–22.
Pathak, M., Raj, B., 2011. Privacy preserving speaker verification using adapted GMMs. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).
Pathak, M., Raj, B., 2012a. Large margin multiclass Gaussian mixture models with differential privacy. IEEE Trans. Depend. Secur. Comput. (TDSC) 9 (4), 463–469.
Pathak, M., Raj, B., 2012b. Privacy preserving speaker verification as password matching. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Pathak, M., Raj, B., 2013. Privacy-preserving speaker verification and identification using Gaussian mixture models. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 21 (2), 397–406.
Pathak, M., Rane, S., Raj, B., 2010. Multiparty differential privacy via aggregation of locally trained classifiers. In: Proceedings of the Neural Information Processing Systems (NIPS), pp. 1876–1884.
Pathak, M.A., Rane, S., Sun, W., Raj, B., 2011. Privacy preserving probabilistic inference with hidden Markov models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5868–5871.
Patino, J., Delgado, H., Evans, N., 2018. The EURECOM submission to the first DIHARD challenge. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2813–2817.
Patsakis, C., van Rest, J., Chorań, M., Bouroche, M., 2015. Privacy-preserving biometric authentication and matching via lattice-based encryption. In: Proceedings of the International Workshop on Data Privacy Management (DPM), pp. 169–182.
Paulini, M., Rathgeb, C., Nautsch, A., Reichau, H., Reininger, H., Busch, C., 2016. Multi-bit allocation: preparing voice biometrics for template protection. In: Proceedings of the Speaker and Language Recognition Workshop (Odyssey), pp. 291–296.
Phan, N., Wang, Y., Wu, X., Dou, D., 2016. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, 16, pp. 1309–1316.

Piciucco, E., Maiorana, E., et al., 2016. Cancelable biometrics for finger vein recognition. In: Proceedings of the IEEE International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), pp. 1–5.
Pinkas, B., Reinman, T., 2010. Oblivious RAM revisited. In: Proceedings of the Annual Cryptology Conference (CRYPTO), pp. 502–519.
Portêlo, J., Abad, A., Raj, B., Trancoso, I., 2015a. Privacy-preserving query-by-example speech search. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Portêlo, J., Raj, B., Abad, A., Trancoso, I., 2014. Privacy-preserving speaker verification using garbled GMMs. In: Proceedings of the European Signal Processing Conference (EUSIPCO), pp. 2070–2074.
Portêlo, J., Raj, B., Boufounos, P., Trancoso, I., Abad, A., 2013. Speaker verification using secure binary embeddings. In: Proceedings of the European Signal Processing Conference (EUSIPCO).
Portêlo, J., Raj, B., Trancoso, I., 2015b. Logsum using garbled circuits. Publ. Libr. Sci. (PloS One) 10 (3), e0122236.
Prabhakar, S., Pankanti, S., Jain, A.K., 2003. Biometric recognition: security and privacy concerns. IEEE Secur. Priv. (SECPRIV) 99, 33–42.
Prince, S.J.D., 2012. Computer Vision: Models, Learning and Inference. Cambridge University Press.
Prince, S.J.D., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the International Conference on Computer Vision (ICCV). CVF.
Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.
Rahulamathavan, S., Yao, X., Yogachandran, R., Cumanan, K., Rajarajan, M., 2018. Redesign of Gaussian mixture model for efficient and privacy-preserving speaker recognition. In: Proceedings of the International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA). IEEE, pp. 1–8.
Rahulamathavan, Y., Sutharsini, K.R., Ray, I.G., Lu, R., Rajarajan, M., 2019. Privacy-preserving i-vector based speaker verification. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 27 (3), 496–506.
Ramos, D., Gonzalez-Rodrigues, J., 2008. Cross-entropy analysis of the information in forensic speaker recognition. In: Proceedings of the IEEE Odyssey.
Rane, S., 2014. Standardization of biometric template protection. IEEE Multimed. 21 (4), 94–99.
Rathgeb, C., Uhl, A., 2011. A survey on biometric cryptosystems and cancelable biometrics. EURASIP J. Inf. Secur. (JIS) 3.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Conversat. Speech Digit. Signal Process. 10, 19–41.
Riazi, M.S., Samragh, M., Chen, H., Laine, K., Lauter, K., Koushanfar, F., 2019. Xonn: Xnor-based oblivious deep neural network inference. In: Proceedings of the USENIX Security Symposium. (to appear)
Riazi, M.S., Weinert, C., Tkachenko, O., Songhori, E.M., Schneider, T., Koushanfar, F., 2018. Chameleon: a hybrid secure computation framework for machine learning applications. In: Proceedings of the ACM ASIA Conference on Computer and Communications Security (ASIACCS), pp. 707–721.
Rua, E.A., Maiorana, E., Castro, J.L.A., Campisi, P., 2012. Biometric template protection using universal background models: an application to online signature. IEEE Trans. Inf. Forensics Secur. (TIFS) 7 (1), 269–282.
Rubinstein, I., Good, N., 2013. Privacy by design: a counterfactual analysis of Google and Facebook incidents. Berkeley Technol. Law J. 28, 1133–1413.
Sadeghi, A.-R., Schneider, T., 2008. Generalized universal circuits for secure evaluation of private functions with application to data classification. In: Proceedings of the International Conference on Information Security and Cryptology (ICISC). Springer, pp. 336–353.
Sadeghi, A.-R., Schneider, T., Wehrenberg, I., 2009. Efficient privacy-preserving face recognition. In: Proceedings of the International Conference on Information Security and Cryptology (ICISC), pp. 229–244.
Sadjadi, S.O., Ganapathy, S., Pelecanos, J.W., 2016. Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5040–5044.
Sanyal, A., Kusner, M.J., Gascón, A., Kanade, V., 2018. TAPAS: tricks to accelerate (encrypted) prediction as a service. Comput. Res. Repos. (CoRR) abs/1806.03461.
Schuller, B., Batliner, A., 2013. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. John Wiley & Sons.
Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 31–36.
Shen, E., Shi, E., Waters, B., 2009. Predicate privacy in encryption systems. In: Proceedings of the Theory of Cryptography Conference (TCC), pp. 457–473.
Shokri, R., Shmatikov, V., 2015. Privacy-preserving deep learning. In: Proceedings of the ACM SIGSAC conference on Computer and Communications Security (CCS), pp. 1310–1321.
Simoens, K., Tuyls, P., Preneel, B., 2009. Privacy weaknesses in biometric sketches. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 188–203.
Simoens, K., Yang, B., Zhou, X., Beato, F., Busch, C., et al., 2012. Criteria towards metrics for benchmarking template protection algorithms. In: Proceedings of the IAPR International Conference on Biometrics (ICB). IAPR, pp. 498–505.
Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., Khudanpur, S., 2018a. Spoken language recognition using x-vectors. In: Proceedings of the Odyssey 2014: The Speaker and Language Recognition Workshop, pp. 105–111.
Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S., 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 165–170.
Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S., 2017. Deep neural network embeddings for text-independent speaker verification. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, pp. 999–1003.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018b. X-vectors: robust DNN embeddings for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329–5333.

Song, D.X., Wagner, D., Perrig, A., 2000. Practical techniques for searches on encrypted data. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 44–55.

Spiekermann, S., Crannor, L.F., 2009. Engineering privacy. IEEE Trans. Softw. Eng. (TSE) 35 (1), 67–82.

Stehlé, D., Steinfeld, R., Tanaka, K., Xagawa, K., 2009. Efficient public key encryption based on ideal lattices. In: Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 617–635.

Stevens, S.S., Volkmann, J., Newman, E.B., 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. (JASA) 8 (3), 185–190.

Teixeira, F., Abad, A., Trancoso, I., 2018. Patient privacy in paralinguistic tasks. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 3428–3432.

Thorne, B., 2017. Python Paillier. https://github.com/n1analytics/python-paillier/. Accessed: 2018-01-11.

Tkachenko, O., Weinert, C., Schneider, T., Hamacher, K., 2018. Large-scale privacy-preserving statistical computations for distributed genome-wide association studies. In: Proceedings of the ACM ASIA Conference on Computer and Communications Security (ASIACCS), pp. 221–235.

Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. Trans. Acoust. Speech Lang. Process. (TASLP) 15 (8), 2222–2235.

Vilda, P.G., Fernández-Baíllo, R., Biarge, M.V.R., Lluis, V.N., Marquina, A.Á., Mazaira-Fernández, L.M., Martínez-Olalla, R., Godino-Llorente, J.I., 2009. Glottal source biometrical signature for voice pathology detection. Speech Commun. 51 (9), 759–781.

Wang, Q., Hu, S., Ren, K., He, M., Du, M., Wang, Z., 2015. CloudBI: practical privacy-preserving outsourcing of biometric identification in the cloud. In: Proceedings of the European Symposium on Research in Computer Security (ESORICS), pp. 186–205.

Wang, S., Hu, J., 2014. Design of alignment-free cancelable fingerprint templates via curtailed circular convolution. Pattern Recognit. 47 (3), 1321–1329.

Xu, Y., Cui, W., Peinado, M., 2015. Controlled-channel attacks: deterministic side channels for untrusted operating systems. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 640–656.

Yao, A.C., 1982. Protocols for secure computations. In: Proceedings of the Annual Symposium on Foundations of Computer Science (SFCS), pp. 160–164.

Yasuda, M., Shimoyama, T., Kogure, J., Yokoyama, K., Koshiba, T., 2013. Packed homomorphic encryption based on ideal lattices and its application to biometrics. In: Proceedings of the International Conference on Availability, Reliability, and Security (ARES), pp. 55–74.

Yasuda, M., Shimoyama, T., Kogure, J., Yokoyama, K., Koshiba, T., 2015. New packing method in somewhat homomorphic encryption and its applications. Secur. Commun. Netw. 8 (13), 2194–2213.

Zahur, S., Rosulek, M., Evans, D., 2015. Two halves make a whole: reducing data transfer in garbled circuits using half gates. In: Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), pp. 220–250.