Privacy-Preserving Speaker Recognition with Cohort Score Normalisation

Andreas Nautsch[1]*, Jose Patino*[1]*, Amos Treiber*[2]*, Themos Stafylakis*[3]*,*
Petr Mizera[3]*, Massimiliano Todisco*[1]*, Thomas Schneider*[2] *and Nicholas Evans*[1]

[1]Audio Security and Privacy, Digital Security Department, EURECOM, France
[2]Cryptography and Privacy Engineering, Department of Computer Science, TU Darmstadt, Germany
[3]Omilia – Conversational Intelligence, Greece

{nautsch,patino,todisco,evans}@eurecom.fr,
{treiber,schneider}@encrypto.cs.tu-darmstadt.de, {tstafylakis,pmizera}@omilia.com

Abstract

In many voice biometrics applications there is a requirement to preserve privacy, not least because of the recently enforced General Data Protection Regulation (GDPR). Though progress in bringing privacy preservation to voice biometrics is lagging behind developments in other biometrics communities, recent years have seen rapid progress, with secure computation mechanisms such as homomorphic encryption being applied successfully to speaker recognition. Even so, the computational overhead incurred by processing speech data in the encrypted domain is substantial. While still tolerable for single biometric comparisons, most state-of-the-art systems perform some form of cohort-based score normalisation, requiring *many thousands of* biometric comparisons. The computational overhead is then prohibitive, meaning that one must accept either degraded performance (no score normalisation) or potential for privacy violations. This paper proposes the first computationally feasible approach to privacy-preserving cohort score normalisation. Our solution is a cohort pruning scheme based on secure multi-party computation which enables privacy-preserving score normalisation using probabilistic linear discriminant analysis (PLDA) comparisons. The solution operates upon binary voice representations. While the binarisation is lossy in biometric rank-1 performance, it supports computationally-feasible biometric rank-n comparisons in the encrypted domain.

Index Terms: privacy, speaker recognition, score normalisation, binary keys, secure computation

1. Introduction

Today there is a growing drive to bring privacy preservation to the realm of speech processing. Following new privacy regulation such as the European GDPR [1], technology to protect sensitive data, including voice data, is attracting the attention of researchers and industrial stakeholders alike. Perhaps the most compelling argument to preserve privacy in speech signals is because they represent inherently personal and private information. Examples include paralinguistic and extralinguistic information, attributes and characteristics, e.g., gender, age, language, dialect, accent, health status, general well-being and emotional state—and the biometric identity.

This paper concerns the protection of privacy for voice biometric applications, e.g. speaker recognition. Recent years have seen rapid progress in privacy-preserving speaker recognition, e.g. [2, 3]. The most recent contribution to the field [4] reported the first i-vector-based solution using homomorphic encryption (HE). HE supports computation upon sensitive biometric voice data *in the encrypted domain* and is a popular tool for privacy preservation. However, the computational demands of HE are prohibitive. This is especially true in the case of speaker recognition systems that employ some form of cohort score normalisation. When operating in unconstrained environments cohort score normalisation is key to performance and is a feature of any state-of-the-art solution. Unfortunately, cohort score normalisation only compounds the computational burden of encryption since it typically involves many thousands of biometric comparisons in the scoring of a single utterance. The scale of the computational demands are currently a bottleneck to privacy preservation for speaker recognition.

The work reported in this paper aims to overcome this bottleneck with an alternative, efficient approach to cohort score normalisation. Using an efficient approach to speaker modelling [5], we propose to replace the speaker representation used in cohort score normalisation with an alternative *binary key* (BK) representation. As a native binary representation, BKs are readily suited to efficient computation in the encrypted domain. The paper shows that the computational overhead of operating upon encrypted representations can then be reduced greatly, meaning that probabilistic linear discriminant analysis (PLDA) comparisons can, for the first time, be performed in the encrypted domain with realistic computational resources.

This paper is organised as follows. Section 2 describes the related work in privacy-preserving speaker recognition. Section 3 describes BK voice representations. Section 4 describes the proposed efficient cohort pruning scheme using BK representations and shows how it can be employed for privacy-preserving score normalisation. Section 5 presents an experimental validation. Conclusions are provided in Section 6.

2. Preliminaries and Related Work

There is an extensive body of literature concerning the preservation of privacy in biometrics. Unfortunately, most relates not to speaker recognition, but to other biometric characteristics, e.g. fingerprint, iris, and face recognition [6, 7]. Whatever the characteristic, the requirements for effective privacy preservation are the same. These are outlined in the ISO/IEC 24745 standard [8] which stipulates that biometric information must be *unlinkable* (data of protected databases are not relatable), *irreversible* (neither embeddings nor audio can be recreated from protected data), and *renewable* (no biometric voice data needs to be recaptured to update a privacy-preservation algorithm).

The conventional approach to meet these requirements involves some form of encryption. Since the late 1980s, the focus of the cryptographic community is *secure computation* [9, 10], specifically the evaluation of a function in ways that do not reveal any information about the inputs of the involved parties,

except for the results. Secure computation mechanisms may be harnessed to retain the functionality of an application without compromising the privacy of the involved parties. The main techniques include homomorphic encryption (HE) which enables computations to be carried out on ciphertexts, and secure multi-party computation (SMPC) which allows interactive computations on data that is *secretly shared*[1] between the parties.[2]

Recent advances in state-of-the-art implementations of secure computation protocols (cf. [11]) have shown to be efficient solutions to privacy preservation in a wide variety of applications [12]. Even so, different solutions offer different levels of computational complexity. SMPC protocols typically involve multiple rounds of interaction (communications between parties involved in the secure computation). While not necessarily requiring interaction, HE usually incurs a higher computational overhead. It follows that, while deployed secure computation techniques can be highly efficient and scalable, it depends on the use case and the employed mechanisms.

Both SMPC and HE have been applied successfully to privacy-preserving speaker recognition [2, 13, 14, 15, 16]. This body of work explores privacy preservation in traditional Gaussian mixture model (GMM) and hidden Markov model (HMM) architectures. Typically, HE is used to hide biometric information, while scoring is sometimes performed using SMPC. The solution reported in [15, 16] preserves privacy in an HMM framework by storing the corresponding secret shares among multiple servers, a technique known as outsourced SMPC [17]. Of course, software solutions are not the only approach to privacy preservation. The work in [18] shows how privacy can also be preserved by using trusted execution environments such as the Intel SGX architecture [19].

Recently, in [4], an HE-based solution to privacy preservation in the form of the Paillier cryptosystem has been applied to state-of-the-art speaker recognition architectures including i-vector systems using PLDA. This work shows that a one-to-one PLDA comparison can be computed in a few hundred milliseconds, depending on whether the speaker model is also protected. Unfortunately, while the solution delivers privacy preservation with no degradation to computational precision, it does not scale well. Protection of a cohort score normalisation process which requires many thousands of comparisons is computationally prohibitive; a runtime in the order of 50 minutes would be needed to process one reference-probe comparison involving 10 000 cohort comparisons, a representative number for today's state-of-the-art techniques.

With cohort score normalisation being a feature of any state-of-the-art approach to speaker recognition, and with performance degradation being the cost of its omission, there is hence an interest to devise computationally manageable solutions. With no previous work having considered this problem thus far, this is the goal of the research reported in this paper.

[1] E.g., in the Boolean Goldreich-Micali-Wigderson (GMW) protocol [10] for two parties that we will use in our work, an input bitstring x can be secretly shared among the parties by sending a random bitstring r of the same length to one party and sending $x \oplus r$ to the other party. Then, the GMW protocol can be executed to securely compute any functionality on x using just the shares of x. The inputs stay hidden because neither r nor $x \oplus r$ reveal any information about x.

[2] Depending on the use case, a party could be a client device, an authentication, or a database/processing server. In contrast to the plaintext domain (one party is sufficient to carry out a computation), security in SMPC is established by splitting computations in a distributed system architecture, where each party computes only on secretly shared data.

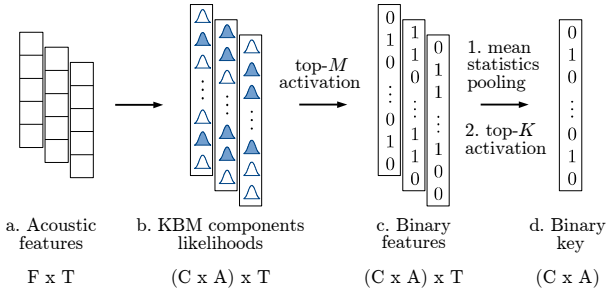

Figure 1: *BK extraction process from T frames with F-dimensional acoustic features to BKs from a KBM with A anchors for each of the C UBM components. Before setting K KBM elements as* True *at the sample level, M elements are preselected at the frame level.*

3. Binary Key Voice Representations

Binary voice representations have been reported previously in the context of privacy preservation. Cryptobiometric (extraction/binding of cryptographic keys *from* biometric data)[3] systems based upon the binarisation[4] of GMM-based supervectors are reported in [20, 3]. The work in this paper uses an alternative, more elaborate approach based upon *binary keys* (BKs), originally proposed in [5, 21]. The BK approach takes a more speaker-discriminatory approach to modelling, much like the idea behind *anchor models* [22, 23]. The same versatile approach has been applied successfully to a number of related problems including emotion recognition [24], speaker change detection [25], speaker diarization [26, 27] and privacy preservation [28] in the context of *cancelable biometric systems* (irreversible feature transforms). Full details of the implementation used in this paper can be found in [5].

The extraction of BKs is performed using a so-called binary key background model (KBM). The KBM plays a role similar to that of a conventional universal background model (UBM) but, instead of representing the acoustic space in an expected sense, it is formed from the concatenation of a number A of speaker-dependent models (learned using traditional UBM maximum a-posteriori adaptation). The role of the KBM anchors is similar in principle to a latent speaker subspace (PLDA alike; in a rough approximation), namely the extraction of discriminant BKs.

The BK extraction process is illustrated in Figure 1. From acoustic features (a), KBM component likelihoods are computed (b). Similarly to i-vector extraction [29], in which component posteriors are pooled to zero order statistics, top-M likelihoods (which at the frame level equals the top-M component posteriors) are used to determine the most frequently activated components (c); again, a rough approximation. An even more compressed speaker representation (d) is obtained with the final BK representation which indicates simply the K elements with the highest pooled mean statistics.

The research hypothesis under investigation is: the loss in precision will be tolerable given their use only for cohort pruning; their use will cause only marginal degradation to the benefit of score normalisation while nonetheless facilitating privacy.

[3] In contrast, HE uses cryptographic keys *for de-/encrypting* biometric data (*biometrics in the encrypted domain*).

[4] The term *binarisation* is potentially misleading. It refers to a *higher level* binary representation (under the acceptance of precision loss) of digital speech data (which is itself already stored in *binary* bit form).

4. Privacy-Preserving Cohort Pruning

The contribution in this paper is an efficient, privacy-preserving approach to score normalisation. It is based upon cohort pruning using BK speaker representations that allow for efficient computation in the encrypted domain. The use of HE-protected i-vectors here is too slow; unprotected i-vectors are not *unlinkable*. The following describes the approach and shows how computation is performed using SMPC while preserving the privacy of both data subjects and cohort speakers.

4.1. Score normalisation

Score normalisation is a processing step of any state-of-the-art approach to ASV. It is applied to remove nuisance bias and variation that would otherwise influence comparison scores in diverse environmental conditions. The general approach to normalisation is based upon a set of auxiliary scores resulting from comparisons between references, probes, and cohort data. A score S is normalised to S' according to $S' = \frac{S-\mu}{\sigma}$, where the mean μ and standard deviation σ are derived from (Gaussian distributed) scores of comparisons with cohort data.

In the case that comparisons involve reference data, this approach is referred to as zero normalisation (*z-norm*). In this case, cohort data characteristics are assumed to match those of the probe $\mathfrak{P}$ (which are fixed for one *quality* condition). Normalisation is then performed using the mean $\mu_{\mathcal{R}}$ and standard deviation $\sigma_{\mathcal{R}}$ derived from the set of comparison scores $\mathcal{R}$.

Normalisation can also be applied using probe data. This is known as test normalisation (*t-norm*). Here, cohort data characteristics are assumed to match those of the reference, in which case the cohort data consists of reference representations. Normalisation is then performed using the mean $\mu_{\mathcal{P}}$ and standard deviation $\sigma_{\mathcal{P}}$ derived from the set of comparison scores $\mathcal{P}$.

Typically, cohort score distributions are rarely Gaussian-distributed. *Adaptive* z-norm (az-norm) and t-norm (at-norm) are commonly applied instead in order to account for this discrepancy. In practice, normalisation is performed with only the top-n scores of $\mathcal{R}$ and $\mathcal{P}$ and, for i-vectors, both (a)z-norm and (a)t-norm are usually combined in symmetric fashion, giving (a)s-norm: $S' = \frac{1}{2}\left(\frac{S-\mu_{\mathcal{R}}}{\sigma_{\mathcal{R}}} + \frac{S-\mu_{\mathcal{P}}}{\sigma_{\mathcal{P}}}\right)$. The normalisation process too needs to preserve privacy.

4.2. Privacy preservation

By using [4], privacy-preserving score normalisation can be performed using reference, probe, and cohort embeddings, all processed in the encrypted domain via HE-based PLDA (HE-PLDA) [4]. The resulting, encrypted scores $\mathcal{R}$, S, and $\mathcal{P}$, none of which reveal any sensitive information, can then be decrypted by an authentication server in order that normalised scores S' can be computed in the plaintext domain.

Assessments of computing demands were performed with a Python implementation of HE-PLDA with two 400-dimensional embeddings, 64-bit floating point precision and a key size of 3072 bits (recommended by NIST [30] in order to support adequate security given advances in computing power until 2030 and beyond) running on an Intel Core i9-7960X CPU with 128 GB of RAM. Computations require 320 ms per comparison when only subject data is encrypted (target in this paper), and 973 109 ms per comparison when both subject data *and* PLDA model parameters are encrypted (the second architecture in [4]). Since a cohort size exceeding some few thousand voice samples is not unusual, the privacy-preserving computation of $\mathcal{R}, \mathcal{P}$ is computationally prohibitive.

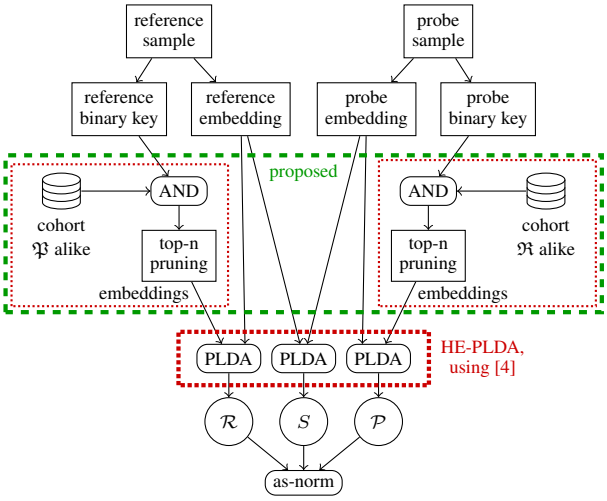

Figure 2: *Our proposed privacy-preserving as-norm protocol with cohort pruning (green dashed area). The red dotted areas indicate that operations are carried out in the encrypted domain and do not leak any information except the decryptable outputs.*

4.3. Cohort pruning

The research hypothesis under investigation here is that the selection of top-n relevant cohort comparisons can be performed more efficiently by accepting some modest degradation to computational precision, while still preserving privacy. Instead of selecting the top-n cohort comparisons using ASV scores, they are selected using measures of acoustic similarity derived from BK representations of reference, probe, and cohort samples. As illustrated in Figure 2, using BK representations, HE-PLDA can then be made efficient via secure bit-wise AND operations and top-n cohort pruning.

More precisely, we employ the Boolean GMW protocol [10] (cf. Section 2) in the case of two involved parties to securely compute our proposed cohort pruning technique. Probe, reference, and cohort BKs are secretly shared between two servers that jointly and securely compute the top-n pruning. This principle is referred to as outsourced secure computation [17]. Assuming non-colluding servers as in [4], this approach can tolerate the corruption of one server without any privacy leakage. This assumption can be seen as realistic, given that one server could be supplied by an independent provider. Since we use protocols with *semi-honest* security here, the secure pruning requires servers that honestly follow the protocol.

Using the GMW protocol, we can easily compute an AND between the secretly shared sample and all secretly shared cohort data. On the resulting shares, we then securely compute the Hamming weight using the circuit of [31] and perform a secure top-n pruning as optimised in [32]. As a result, the identifiers of the top-n embeddings are revealed and can be used for score normalisation. Apart from this information, nothing else is leaked about the sample and cohort voice data.

5. Experimental Validation

Given the research objective to demonstrate improvements in computational efficiency, rather than improved performance, only brief details of the text-independent speaker recognition system are provided here. It is based on 400-dimensional i-vectors, extracted from conventional acoustic features using time delay deep neural network (TDNN) for estimating UBM

posteriors. The TDNN is trained using the KALDI toolkit [33] with SRE'04-08 and SWBD data (not x-vectors). The Python backend is based on PLDA with mean and length normalisation, and trained with SRE'04-08 data. The KBM used for BK extraction is learned in Matlab using a 2 048-component UBM trained with conventional acoustic features and $A = 20$ anchors (10 fe/male each). KBM optimisation is performed using a subset of the cohort set containing data from 71 speakers. For BK extraction, at feature level the top $M = 1$ components are activated, while at sample level the top $K = 2048$ bits are set. The cohort set is a subset of the PLDA training set with 11 640 voice samples of 3 812 speakers. The proposed approach is evaluated on the 2010 NIST SRE common condition 5, particularly the core-core and core-10s protocols. In order to report on diverse data, we pooled the scores of both protocols (core-core/10s).

5.1. Recognition results

Results are reported in terms of $C_{\text{llr}}^{\text{min}}$, the minimum decision cost function (minDCF; effective prior 0.01) and the equal-error rate (EER). Table 1 shows that conventional as-norm gives the same or better performance than the baseline system (without any score normalisation). The proposed *privacy-preserving* as-norm solution gives slightly worse results in terms of minDCF even though, curiously, improvements are observed in terms of $C_{\text{llr}}^{\text{min}}$ and EER. For minDCF, an improvement over the baseline is also observed but without reaching the performance of the unprotected AS-norm. This result confirms our research hypothesis: privacy preservation incurs only a modest performance degradation (in the minDCF sense) and, encouragingly, also improves upon the baseline system without any score normalisation (in the $C_{\text{llr}}^{\text{min}}$ and EER sense).

5.2. Proof of biometric information protection

The sample embeddings as well as the cohort embeddings used for PLDA comparisons are protected via the original privacy-preserving PLDA system. As such, if biometric information protection in the form of unlinkability, irreversibility, and renewability is given by the original system (as in [4]), then the embeddings are protected as well. The BKs of samples and cohorts are protected by the Boolean secret sharing of the GMW protocol between two servers (cf. Section 2). Because of the information-theoretic indistinguishability of any two secret shares, unlinkablitity and irreversibility are guaranteed. Due to the nature of secret sharing, the protected data is also renewable; secret shares can be re-randomised with a new random bitstring.

5.3. Complexity analysis

We implemented our secure cohort pruning architecture using the state-of-the-art SMPC framework *ABY* [12]. We ran our implementations on two machines with Intel Core i9-7960X CPUs and 128 GBs of RAM. To simulate real-world network conditions of the involved servers, we restricted the connection between the servers to 1 Gbit/s bandwith and 1 ms round trip time. Results are presented in Table 1. Note that these are the online runtimes and that some additional input-independent precomputation is required; we account for the BK extraction time[5] (28.3 s and 3.2 s for a core-core and for a core-10s probe, respectively; for core-core/10s, the average is 16.9 s). The largest gain in real-world network conditions for privacy-preserving score normalisation are observed for small cohorts with $14\times$ to $19\times$ gains in runtimes. In other words, rather than runtimes in the order of 50 minutes only a few minutes are necessary. In the privacy-preserving cohort pruning and as-norm, the BK extraction takes 6.4-20.0% of the runtime and the GMW pruning takes 39.2-78.0% (their time share is lower on higher cohort sizes as the runtime share of HE-PLDA increases). For the GMW pruning, all privacy-preserving az/at-norm comparisons (using BKs against the entire cohort) are carried out in less than 157 s and 52 s, respectively. These times already include the sorting of the top-50 cohort indices for pruning the HE-PLDA cohort comparisons. To prune larger cohort sizes, the privacy-preserving sorting requires additional time, e.g. from top-50 to top-400 in az-norm, an additional 126 s are necessary.

[5] BKs are extracted with Matlab on a DELL R620 with two Intel Xeon E5-2630L CPUs and 128 GBs of RAM.

Table 1: *Runtimes and recognition results for the baseline system, the baseline system with conventional as-norm and the proposed privacy-preserving alternative, for different cohort sizes n. The realtime improvement results from dividing the time of scoring all cohort data with HE-PLDA by: BK extraction + GMW (scoring and top-n sorting) + top-n HE-PLDA time.*

n		50	100	150	200	250	300	400
Baseline ($C_{\text{llr}}^{\text{min}}$ / minDCF / EER)		0.161 / 0.410 / 4.6						
Runtime top-n HE-PLDA (necessary)		16 s	32 s	48 s	64 s	80 s	96 s	128 s
HE-PLDA (z-norm)		3 725 s (for all 11 640 reference-cohort comparisons)						
GMW pruning (BK: 28 s)		157 s	177 s	198 s	220 s	247 s	269 s	283 s
improvement (az-norm)		**19**×	16×	14×	12×	10×	9×	8×
HE-PLDA (t-norm)		1 220 s (for all 3 812 cohort-probe comparisons)						
GMW pruning (BK: 17 s)		52 s	59 s	66 s	73 s	82 s	89 s	94 s
improvement (at-norm)		**14**×	11×	9×	8×	7×	6×	5×
conventional (unprotected) as-norm	$C_{\text{llr}}^{\text{min}}$	0.161	0.157	0.156	0.155	0.155	0.155	0.155
	minDCF	0.390	0.376	0.374	0.373	0.374	**0.372**	0.369
	EER	4.6	4.5	4.5	4.5	4.4	4.4	4.4
proposed	$C_{\text{llr}}^{\text{min}}$	0.158	0.151	0.149	**0.147**	0.149	0.149	0.149
	minDCF	0.509	0.492	0.466	0.452	0.435	0.429	0.408
	EER	4.4	4.3	4.3	**4.1**	4.2	**4.1**	4.2

6. Conclusions

This paper reports the first approach to computationally manageable (yet demanding) privacy-preserving speaker recognition with cohort score normalisation. Prior to this work, the latter was a computational bottleneck for PLDA with Paillier homomorphic encryption, with normalisation strategies that require many thousands of biometric comparisons being computationally prohibitive when performed in the encrypted domain. The set of cohort data used for score normalisation is pruned using a native binary speaker representation. Privacy is outsourced via secure multi-party computation through which a top-n cohort set is pruned securely. Privacy-insensitive cohort scores can then be decrypted and treated in the usual way. The cohort list is revealed to the sites capturing the reference and the probe data, respectively. This could be used by a security (*not privacy*) adversary to mount hill-climbing attacks; instead, if the top-n lists are in the province of the biometric service owner, these top-n indices serve the intended recognition purpose. Future work could investigate the use of SMPC protocols for carrying out ranking and matrix operations in the protected domain (including PLDA) demanding less computational but more server communication overheads.

Acknowledgements. This work was supported by the BMBF and the HMWK within CRISP, the DFG as part of project E4 within the CRC 1119 CROSSING and A.1 within RTG 2050, by Omilia – Conversational Intelligence, and by the Voice Personae and RESPECT projects, both funded by the French ANR.

7. References

[1] European Council, "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)," April 2016.

[2] J. Portêlo, B. Raj, A. Abad, and I. Trancoso, "Privacy-preserving speaker verification using garbled GMMs," in *Proc. European Signal Processing Conf. (EUSIPCO)*. IEEE, 2014, pp. 2070–2074.

[3] M. Paulini, C. Rathgeb, A. Nautsch, H. Reichau, H. Reininger, and C. Busch, "Multi-bit allocation: Preparing voice biometrics for template protection," in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, 2016, pp. 291–296.

[4] A. Nautsch, S. Isadskiy, J. Kolberg, M. Gomez-Barrero, and C. Busch, "Homomorphic encryption for speaker recognition: Protection of biometric templates and vendor model parameters," in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*. ISCA, 2018, pp. 16–23.

[5] X. Anguera and J.-F. Bonastre, "A novel speaker binary key derived from anchor models," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*. ISCA, 2010, pp. 2118–2121.

[6] M. Blanton and P. Gasti, "Secure and efficient protocols for iris and fingerprint identification," in *Proc. European Symposium on Research in Computer Security (ESORICS)*. Springer, 2011, pp. 190–209.

[7] J. Bringer, H. Chabanne, M. Favre, A. Patey, T. Schneider, and M. Zohner, "GSHADE: faster privacy-preserving distance computation and biometric identification," in *Proc. ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*. ACM, 2014, pp. 187–198.

[8] ISO/IEC JTC1 SC27 Security Techniques, *ISO/IEC 24745:2011. Information Technology - Security Techniques - Biometric Information Protection*, International Organization for Standardization, 2011.

[9] A. C. Yao, "Protocols for secure computations," in *Proc. Annual Symposium on Foundations of Computer Science (SFCS)*. IEEE, 1982, pp. 160–164.

[10] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game," in *Proc. ACM Symposium on Theory of Computing (STOC)*. ACM, 1987, pp. 218–229.

[11] M. Hastings, B. Hemenway, D. Noble, and S. Zdancewic, "SoK: General-purpose compilers for secure multi-party computation," in *Proc. IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2019, full version: https://marsella.github.io/static/mpcsok.pdf.

[12] D. Demmler, T. Schneider, and M. Zohner, "ABY-A framework for efficient mixed-protocol secure two-party computation," in *Proc. Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2015.

[13] P. Smaragdis and M. Shashanka, "A framework for secure speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 15, no. 4, pp. 1404–1413, 2007.

[14] M. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using Gaussian mixture models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 2, pp. 397–406, 2013.

[15] M. Aliasgari and M. Blanton, "Secure computation of hidden Markov models," in *Proc. Intl. Conf. on Security and Cryptography (SECRYPT)*. IEEE, 2013, pp. 1–12.

[16] M. Aliasgari, M. Blanton, and F. Bayatbabolghani, "Secure computation of hidden Markov models and secure floating-point arithmetic in the malicious model," *Intl. Journal of Information Security*, vol. 16, no. 6, pp. 577–601, 2017.

[17] S. Kamara and M. Raykova, "Secure outsourced computation in a multi-tenant cloud," in *Proc. IBM Workshop on Cryptography and Security in Clouds*, 2011, pp. 15–16.

[18] F. Brasser, T. Frassetto, K. Riedhammer, A.-R. Sadeghi, T. Schneider, and C. Weinert, "VoiceGuard: Secure and private speech processing," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*. ISCA, 2018, pp. 1303–1307.

[19] F. McKeen, I. Alexandrovich, A. Berenzon, C. V. Rozas, H. Shafi, V. Shanbhogue, and U. R. Savagaonkar, "Innovative instructions and software model for isolated execution," in *Proc. Workshop on Hardware and Architectural Support for Security and Privacy (HASP)*. ACM, 2013.

[20] S. Billeb, C. Rathgeb, H. Reininger, K. Kasper, and C. Busch, "Biometric template protection for speaker recognition based on universal background models," *IET Biometrics*, vol. 4, no. 2, pp. 116–126, 2015.

[21] J.-F. Bonastre, P.-M. Bousquet, D. Matrouf, and X. Anguera, "Discriminant binary data representation for speaker recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5284–5287.

[22] T. Merlin, J.-F. Bonastre, and C. Fredouille, "Non directly acoustic process for costless speaker recognition and indexation," in *Proc. Intl. Workshop on Intelligent Communication Technologies and Applications*, vol. 29, 1999.

[23] Y. Mami and D. Charlet, "Speaker identification by location in an optimal space of anchor models," in *Proc. Intl. Conf. on Spoken Language Processing (ICSLP)*, 2002.

[24] J. Luque and X. Anguera, "On the modeling of natural vocal emotion expressions through binary key," in *Proc. European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 1562–1566.

[25] J. Patino, H. Delgado, and N. Evans, "Speaker change detection using binary key modelling with contextual information," in *Proc. Intl. Conf. on Statistical Language and Speech Processing (ICSLP)*. Springer, 2017, pp. 250–261.

[26] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Fast single-and cross-show speaker diarization using binary key speaker modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2286–2297, 2015.

[27] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first DIHARD challenge," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*. ISCA, 2018, pp. 2813–2817.

[28] A. Mtibaa, D. Petrovska-Delacretaz, and A. B. Hamida, "Cancelable speaker verification system based on binary Gaussian mixtures," in *Proc. Advanced Technologies for Signal and Image Processing (ATSIP)*, 2018, pp. 1–6.

[29] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 4, pp. 788–798, 2011.

[30] E. Barker, "NIST special publication 800–57 part 1, revision 4," 2016.

[31] J. Boyar and R. Peralta, "The exact multiplicative complexity of the Hamming weight function," in *Proc. Electronic Colloquium on Computational Complexity (ECCC)*, 2005.

[32] K. Järvinen, H. Leppäkoski, E. S. Lohan, P. Richter, T. Schneider, O. Tkachenko, and Z. Yang, "PILOT: Practical privacy-preserving Indoor Localization using OuTsourcing," in *Proc. IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2019, to appear. Preliminary version: https://encrypto.de/papers/JLLRSTY19.pdf.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.