Benchmarking and challenges in security and privacy for voice biometrics

Jean-Francois Bonastre, Héctor Delgado, Nicholas Evans, Tomi Kinnunen,
Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Paul-Gauthier Noé, Jose Patino,
Md Sahidullah, Brij Mohan Lal Srivastava, Massimiliano Todisco,
Natalia Tomashenko, Emmanuel Vincent, Xin Wang, Junichi Yamagishi

ASVspoof and VoicePrivacy organising committees

organisers@asvspoof.org, organisers@voiceprivacychallenge.org

Abstract

For many decades, research in speech technologies has focused upon improving reliability. With this now meeting user expectations for a range of diverse applications, speech technology is today omni-present. As result, a focus on security and privacy has now come to the fore. Here, the research effort is in its relative infancy and progress calls for greater, multidisciplinary collaboration with security, privacy, legal and ethical experts among others. Such collaboration is now underway. To help catalyse the efforts, this paper provides a high-level overview of some related research. It targets the non-speech audience and describes the benchmarking methodology that has spearheaded progress in traditional research and which now drives recent security and privacy initiatives related to voice biometrics. We describe: the ASVspoof challenge relating to the development of spoofing countermeasures; the VoicePrivacy initiative which promotes research in anonymisation for privacy preservation.

1. Introduction

The voice is among the most natural and convenient means to human-machine interaction and biometric authentication. In some scenarios, particularly telephony or teleconferencing applications, voice can be the *only* available biometric. While automatic speaker verification (ASV) systems can provide a reliable means to authentication, like all biometric technologies, it is not without security and privacy concerns. Arguably, these concerns are potentially greater for voice biometrics than they can be for authentication systems that use alternative biometric characteristics.

Security concerns relate to the potential for ASV systems to be manipulated by adversaries through spoofing attacks [1], now referred to as presentation attacks [2]. Fraudsters can launch spoofing attacks to gain illegitimate access to protected services or resources by presenting to the ASV system a speech recording which has been manipulated to *sound*[1] like another speaker. Without adequate protection, spoofing attacks can substantially degrade the reliability of almost any ASV system.

While not related specifically to voice biometrics, but to speech technology more generally, privacy concerns relate to the potential for speech data to be exploited for purposes other than those to which an individual might have given consent [3]. Speech signals are a rich source of personal, private information. In providing recordings of speech to a particular voice

---

[1]We refer to machine perception rather than human perception; humans and machines do not *hear* in the same way.

service, the speaker usually furnishes the service provider with much more information than is strictly necessary in order to perform the expected task, hence the need for privacy preservation.

In this article we describe two specific benchmarking challenges launched by the speech processing research community to expedite solutions to security and privacy concerns. The first involves solutions to protect ASV systems from being manipulated by spoofing in the form of spoofing countermeasures or presentation attack detection systems [2] whose development is spearheaded through the ASVspoof initiative launched in 2015 [4, 5, 6]. The second relates to the VoicePrivacy initiative, launched in 2020 [7], which aims to promote the development of privacy preservation solutions for speech technology. The inaugural VoicePrivacy challenge focused upon anonymisation, namely techniques to manipulate speech data in order that it cannot be used with ASV systems to recognise the speaker.

The article is intended as a contribution to the efforts to build bridges between the speech community and, e.g., the legal and ethical communities and, in particular, to support the recently formed Security and Privacy in Speech Communication (SPSC) special interest group of the International Speech Communication Association (ISCA). It targets the non-specialist and is hence intentionally high-level, with a focus upon the essentials and clarity, rather than upon scientific and technical rigour. We describe the importance of benchmarking campaigns, classifier fundamentals and the early, classical approaches to performance estimation. After presenting a high-level overview of ASV systems, the remainder of the article introduces the ASVspoof and VoicePrivacy initiatives.

2. Evaluation-driven research

In the early days, researchers collected their own datasets to develop methods (algorithms and software). To make technological progress in complex tasks such as speech or speaker recognition, and to be able to meaningfully compare different methods, the need for commensurable performance benchmarking was quickly recognised. From the mid-1990s, the National Institute of Standards and Technology (NIST) — a US-based standardisation body — pioneered *evaluation-driven* research [8, 9]. The key ingredients are: (1) commonly agreed (often public) data, evaluation rules, and performance metrics; (2) disentangled roles for researchers (evaluees) and evaluators. Performance claims should not be reported by evaluees but instead by independent evaluators who provide common infrastructure (data, rules, metrics) in the form of an *evaluation campaign* or *challenge*. Only with such a level playing field can competing methods can be meaningfully compared. There are many such
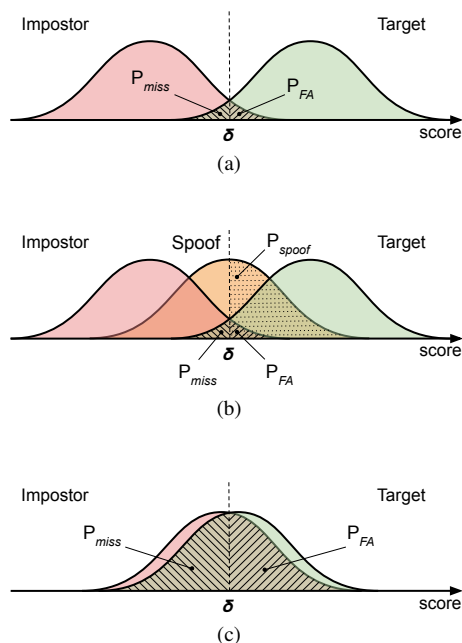
Figure 1: Illustrative score distributions for (a) automatic speaker verification (ASV), (b) an ASV system subjected to spoofing attacks and (c) an ASV system presented with anonymised data.

challenges [10, 11] within the speech field. Typically, they require participants [12, 13] to design special-purpose software to solve some specific tasks. For many, the software takes the form of a *binary classifier*, namely a system which must choose between two mutually exclusive hypotheses.

Even if classifiers are based on well-established statistical principles and are trained objectively via numerical optimisation techniques, they can (and do) make errors. Errors are the result of over-simplified modelling assumptions, natural variability in the input data, sources of external nuisance variation, or as a result of limited training data, etc. The design of ever-more reliable classifiers is the traditional bread and butter of machine learning and speech research.

A binary classifier makes two types of errors: *misses* (false rejections) and *false alarms* (false acceptances). Misses imply that the classifier rejects a positive input that should be accepted. A false alarm results from the classifier accepting a negative input that should be rejected. Classifier performance can be estimated by dividing the number of misses and false alarms by the number of tested positive and negative cases respectively. The resulting *miss rate* (P_{miss}) and *false alarm rate* (P_{FA}) can be seen as proxies for user convenience and security.

Binary classifier decisions are derived in two stages. The first involves computation of a 'soft decision' (the *score*) — a real number that expresses the classifier's confidence in the positive case. Example raw score *distributions* for such a binary classifier are illustrated in Fig. 1(a) were the positive case is the target class and the negative case is the impostor class. In practice, the score is often a *logarithmic likelihood ratio* (LLR) and expresses the relative strength between the two competing hypotheses. Second, the score is compared with a pre-set *threshold* δ. Scores above the threshold imply 'accept', whereas scores below the threshold imply 'reject'. By increasing the

threshold, the false alarm rate (P_{FA}, red shaded area in Fig. 1) can be reduced at the expense of a higher miss rate (P_{miss}, green shaded area) and vice versa. The trade-off can be readily visualised in so-called detection error trade-off (DET) plots [14] such as that illustrated in Fig. 2 (described below).

Since there are two error rates, which do we report? Or do we report both? What threshold / how do we set it? The answer to these questions is rather subtle and a detailed treatment is outside the scope of this paper. One particular metric, namely the *equal error rate* (EER) is adopted for a broad range of tasks. It is computed using the threshold that makes miss and false alarm rates equal (see Fig. 2) — thereby yielding a single number to report. The *lower* the EER, the more reliable the classifier.

The authors acknowledge that, even if the EER is deprecated in ISO standards [15], it provides a compact summary of the *discrimination* capabilities of a classifier — how well it is capable of observing (in speech, 'hearing') differences between positive and negative inputs. In practice, however, the EER does not provide a full picture. More comprehensive approaches to assessment have been developed and have been broadly adopted by the community. These provide a view of classifier performance through the lens of a formal *decision policy* which represents the effective trade-off between the two decision outcomes [16].

3. Automatic speaker verification

ASV systems provide one of the most natural and convenient means to biometric person authentication. Test recordings (probes) are compared with enrolment recordings (references) to verify (or not) a claimed identity. The reference is used to create a model[2] which is stored in a reference database. At test time, the model corresponding to the claimed identity is compared to the test utterance resulting in a soft score. A hard accept/reject decision can be fully automated (e.g., online banking) or semi-automated (with some human intervention, e.g., when forensic practitioners present voice evidence in court). Scores should reflect reliably the extent to which the strength-of-evidence supports the same/different identity propositions: the higher the score, the greater the similarity and vice versa.

Detection error trade-off plots for two different ASV systems assessed using the ASVspoof 2019 logical access database and the VoicePrivacy 2020 database are illustrated in Figs. 2 and 4 and show EERs of 2.5% and 1.1% respectively (blue profiles). Each point on any one profile corresponds to a different decision threshold (operating point). The profiles show how misses can be traded off against false alarms in order to meet different application requirements.

4. Security vulnerabilities: ASVspoof

Without adequate protections, the reliability of ASV systems can be compromised by the presentation of synthetic or converted voice, replayed speech and impersonation [6]. Generated automatically from a text input, today's state-of-the-art synthesis systems are capable of producing speech that the human cannot distinguish from bona fide speech [17]. Voice conversion systems operate directly upon an input speech signal and alter the voice to that of another speaker [18]. Unlike synthetic and converted voice spoofing attacks, which both demand a certain technical expertise and suitable training and adaptation data,

---

[2]On account of their dynamic nature and the variability in speech signals, we refer to models, not templates.
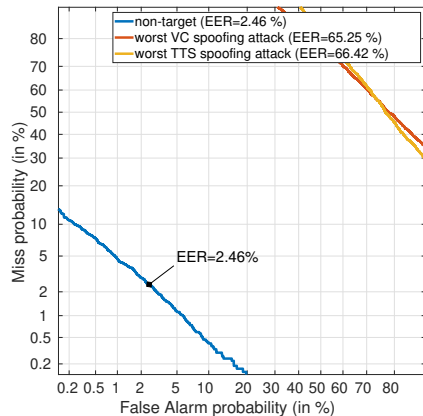
Figure 2: Detection error trade-off plot for the ASVspoof 2019 logical access task. Profiles shown for the baseline **ASV** system (blue profile) and the same system subjected to the most effective synthetic (text-to-speech, TTS) speech spoofing attack (orange profile) and the most effective voice conversion (VC) spoofing attack (red profile).

Figure 3: As for Fig. 2 except for **spoofing countermeasures**. Profiles shown for the top five performing systems.

replay attacks can be launched by the layman, requiring only consumer-grade recording and replaying devices. All can substantially degrade ASV reliability. Impersonation, while still a threat [19, 20], is less effective.

The impact of spoofing attacks upon an ASV system is illustrated in Fig. 1(b). Spoofing attacks introduce a third class of input so that the ASV system must now cope with target (matching), impostor (non-matching) and spoofed utterances. A successful spoofing attack circumvents the ASV system by provoking a score above the decision threshold. The score distribution for spoofing attacks is illustrated in the middle of Fig. 1(b). Spoofing attacks provoke a false alarm rate P_{spoof} that is greater than the original false alarm rate P_{FA}. One can think of spoofing attacks as a special case of impostors where there is a concerted effort to deceive the ASV system.

Without the capacity to distinguish between spoofed and bona fide speech, ASV reliability will degrade as a result of spoofing attacks, sometimes substantially. This degradation assessed using the ASVspoof 2019 logical access database (synthetic and converted voice spoofing attacks) is illustrated in Fig. 2. The baseline EER (with no no spoofing attacks) of 2.5% increases to over 50% when impostor trials are replaced by the most effective synthetic and converted voice spoofing attacks. These results should be interpreted with caution, however. In practice, one must consider the relative likelihood of the ASV system being presented with target and impostor trials, versus that of spoofing attacks. Without this consideration, the results in Fig. 2 might convey an overly pessimistic view of the vulnerabilities to spoofing. There is, in any case, potential to detect attacks automatically using countermeasures.

The series of ASVspoof challenges held bi-annually since 2015 have spearheaded the development of spoofing countermeasures or presentation attack detection (PAD) solutions for ASV. Spoofing countermeasures can be applied prior to ASV in order to detect attacks and prevent them from reaching the ASV system. The most recently completed challenge was held in 2019 and included separate logical access (synthetic and converted voice attacks) and physical access (replay attacks)
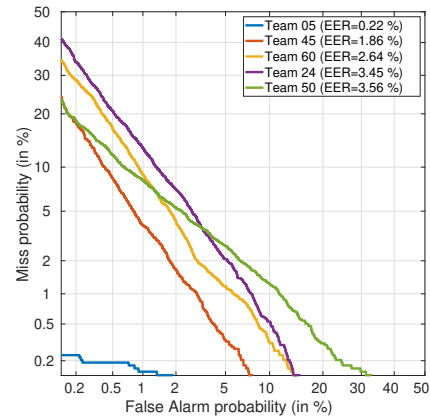
tasks [6]. DET plots for the top-five performing spoofing countermeasures for the ASVspoof 2019 logical access task are shown in Fig. 3. They show EERs of as low as 0.2%.

Thus, while spoofing attacks can present a substantial threat to reliability, and while human listeners may not be able to detect spoofing attacks, countermeasures can be effective. Even so, while ASV countermeasures have proven potential to detect attacks, they can also degrade usability; they can erroneously classify bona fide speech as spoofed speech. Assessment and performance estimates should hence reflect the impact of both spoofing and countermeasures *upon the ASV system*; countermeasures should be assessed in tandem with ASV. Such more elaborate approaches to assessment, e.g. [21], are outside the scope of the current article.

5. Privacy implications: VoicePrivacy

Speech signals contain much more than just the spoken message or the voice identity [22, 23]. The speaker's sex/gender, age, socio-economic and geographical background, emotion and health condition etc. can all be estimated automatically using recordings of speech [24, 25]. With much of this information being personal and private there is hence an interest to develop privacy safeguards for speech technology. This is the goal of the VoicePrivacy initiative, founded in 2020. While solutions to privacy preservation can take many different forms, and while VoicePrivacy may explore different approaches in the future, the inaugural challenge focused upon the development of *anonymization* solutions [7].

The idea is to protect privacy by distorting speech data such that it cannot be used by ASV systems to recognise the speaker. Anonymisation is achieved by suppressing the information in speech signals that is typically used by machines to infer identity. On its own, this is relatively straightforward, e.g. by adding sufficient levels of background noise, or by replacing speech with silence. The challenge comes from the requirement to suppress personally identifiable attributes contained within the speech signal while leaving all other attributes intact. These requirements imply that anonymisation should not interfere with the application of some down stream tasks such as automatic speech recognition, nor should it introduce processing artefacts that might degrade subjective intelligibility or naturalness. Last, anonymised voices should remain distinctive, meaning the task
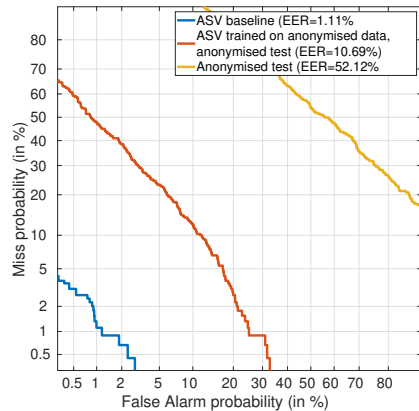
Figure 4: Detection error trade-off plot for the VoicePrivacy 2020 challenge (LibriSpeech test, male trials). Profiles shown for the baseline ASV system (blue profile) and the same system presented with anonymised test utterances (orange profile) and anonymised test utterances when the ASV system is re-trained using similarly anonymised training data (red profile).

might better be referred to as pseudonymisation.

Rather than operating upon the speech signal so that it reflects the voice of another, specific speaker, anonymisation aims to prevent the speaker identity from being recognised. Anonymisation acts to increase the confusion between utterances produced by the same and different speakers. For a perfect anonymisation system, the score distributions corresponding to impostor and target trials overlap. The desired effect of anonymisation upon an ASV system is illustrated in Fig. 1(c). In this case, the ASV system cannot produce simultaneously both a low false alarm rate and a low miss rate, no matter what the decision threshold, and the EER is 50%. Real anonymisation solutions are less effective.

The first VoicePrivacy challenge was held in 2020 and attracted submissions from 7 independent teams. A DET plot for the primary challenge baseline [7] is illustrated in Fig. 4. It shows an increase in the EER from a baseline of 1% (blue profile) to 52% after anonymisation (orange profile). The EER of over 50% suggests that the anonymisation goal is met. However, if an anonymisation adversary were to adapt the ASV system in light of anonymisation, then performance is less effective; anonymised utterances still contain some personally identifiable information (PII) and the potential to re-identify the speaker remains. When the ASV system is re-trained or adapted using similarly anonymised training data, the result is a lower EER of 10.7% (red profile); true anonymisation remains elusive.

The above treatment does not reflect impacts upon intelligibility/naturalness, nor voice distinctiveness. In practice, the multiple objectives and complex nature of anonymisation means several different metrics are used in practice. The development of more suitable approaches to assessment are the focus of current research [26, 27, 28, 29].

6. Reflections and further considerations

The community has made substantial progress to address the security and privacy implications of speech technology and rapid progress has been made in the last half-decade. We have: proposed definitions for a number of distinct tasks and solutions; established evaluation driven research initiatives as a vehicle to successfully raise the profile of security and privacy research and to build new research communities; developed protocols, criteria and metrics for the assessment of security and privacy safeguards. Nonetheless, we are perhaps still far from having a comprehensive appreciation of the implications as well as the potential of safeguards.

Will spoofing ever be a solved problem? Possibly not. Speech synthesis and voice conversion are long-established research fields with genuine applications, such as anonymisation, yet the same technology poses a threat to the security of ASV. The progress speech synthesis and voice conversion in recent years has been impressive. Today's technology produces synthetic speech that humans cannot distinguish from bona fide speech. Whether or not similar advances may one day result in machines that produce synthetic speech that other machines cannot detect is an intriguing question. For the time being it is clear that, if we seek reliable, secure approaches to person authentication using ASV, we must intensify our efforts in anti-spoofing. Future directions include a focus on more adversarial attacks, the development of countermeasures that function reliably in the wild, e.g. in the face of background noise and other sources of nuisance variation such as bandwidth and channel variability which typify telephony ASV scenarios, as well as the approach used to estimate performance.

The research effort in anonymisation is relatively embryonic. While there is an opportunity to provide some level of protection, current solutions fall short of delivering true anonymisation. Furthermore, we have observed differences in the level of privacy that a given anonymisation solution provides to different individuals. Whereas the protection for some can be strong, others are left with relatively little protection at all. Since personally identifiable information encapsulates far more than that used by most ASV systems to infer identity, since other alternative attributes can be used instead, and since current anonymisation solutions do not necessarily suppress them, is *full* anonymisation even technically possible?

Anonymisation solutions that focus only upon the attributes used by typical ASV systems already degrade intelligibility and naturalness. If *all* personally identifiable information can be successfully suppressed, then what remains? Is personally identifiable information so inextricably, indelibly embedded in speech that it cannot be (fully) removed? Even if a speech signal can be fully anonymised, will anything resembling speech remain? Or, for a given application, what level of intelligibility/naturalness can be sacrificed in order to achieve anonymisation? How should we even measure anonymisation performance and privacy? Are the methodology and metrics adequate? How does this research fit with broader solutions to privacy preservation, e.g. encryption and distributed learning?

That we have certainly raised more questions above than we have answered serves to show that our community's journey in security and privacy research is only just beginning. Our efforts must be redoubled looking to the future.

7. Acknowledgements

8. References

[1] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, "Introduction to voice presentation attack detection and recent advances," in *Handbook of Biometric Anti-Spoofing*, pp. 321–361. Springer, 2019.

[2] "ISO/IEC 30107. Information Technology – Biometric presentation attack detection," Standard, International Organization for Standardization, Geneva, Switzerland, 2016.

[3] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding," in *Proc. Interspeech 2019*, 2019, pp. 3695–3699.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017, pp. 2–6.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.

[7] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*, oct 2020, pp. 1693–1697.

[8] K. A. Lee, O. Sadjadi, H. Li, and D. Reynolds, "Two decades into speaker recognition evaluation - are we there yet?," *Computer Speech & Language*, vol. 61, pp. 101058, 2020.

[9] C. S. Greenberg, L. P. Mason, et al., "Two decades of speaker recognition evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, pp. 101032, 2020.

[10] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 2020.

[11] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV Challenge 2020: Large-Scale Evaluation of Short-Duration Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 731–735.

[12] K. A. Lee, V. Hautamaki, T. Kinnunen, H. Yamamoto, et al., "I4U submission to NIST SRE 2018: Leveraging from a decade of shared experiences," in *Proc. Interspeech*, 2019, pp. 1497–1501.

[13] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, et al., "State-of-the-art speaker recognition for telephone and video speech: the JHU-MIT submission for NIST SRE18," in *Proc. Interspeech*, 2019, pp. 1488–1492.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.

[15] "ISO/IEC IS 19795-1:2021. Information Technology – Biometric performance testing and reporting – Part 1: Principles and framework," Standard, International Organization for Standardization, Geneva, Switzerland, 2021.

[16] A. Nautsch, *Speaker Recognition in Unconstrained Environments*, Ph.D. thesis, Technische Universität Darmstadt, 2019.

[17] J. Shen, R. Pang, R. J. Weiss, et al., "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[18] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, et al., "Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[19] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry.," in *Proc. Interspeech*, 2013, pp. 930–934.

[20] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, feb 2015.

[21] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, et al., "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[22] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, et al., "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[23] COMPRISE, "Deliverable Nº5.1: Data protection and GDPR requirements [online]. available: https://www.compriseh2020.eu/files/2019/06/d5.1.pdf," 2019.

[24] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in *Proc. ASRU*, 2003, pp. 31–36.

[25] T. Schultz, "Speaker characteristics," in *Speaker classification I*, pp. 47–74. Springer, 2007.

[26] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?," in *Interspeech 2019*, sep 2019, pp. 3700–3704.

[27] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, and N. Evans, "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment," in *Proc. Interspeech*, 2020, pp. 1698–1702.

[28] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "A Comparative Study of Speech Anonymization Metrics," in *Proc. Interspeech*, 2020, pp. 1708–1712.

[29] P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, "Speech Pseudonymisation Assessment Using Voice Similarity Matrices," in *Proc. Interspeech 2020*, 2020, pp. 1718–1722.