Engineering a new scholarly ecosystem: security and privacy in speech communication

ANDREAS NAUTSCH, EURECOM, France, now with vitas.ai, Germany

1 BRIEF OVERVIEW; ON EFFORTS PREPARING A SIG COMING INTO PLACE

We can neither avoid that digital societies are becoming more complex rapidly, nor postpone the social contract further that *technology is to aid society*. Such is the responsibility of industry and academia, alike [to aid society]. We know that we need to come together from the different disciplines: but how do we do it? This position paper describes personal experience summarising a time period of about three years.[1] During this time, independent activities led to the formation of the Special Interest Group (SIG) on *Security and Privacy in Speech Communication* (SPSC) within the International Speech Communication Association (ISCA).[2]

Prior, in Interspeech proceedings (formerly named Eurospeech), ISCA research discussed security issues for voice biometrics[3], an early work on security appeared in 1999 [13]—playback and synthesis of spoken digits to subvert voice biometrics—; whereas early works on privacy appeared in 2005 [7] and 2007 [28]—concerning speech recognition from sensor level to segmentation of multi-person and situated spontaneous speech.[4] In subsequent Interspeech proceedings, speeding-up search on large-scale data, binarisation concepts were investigated: cryptographic approaches (not operating on floating point data) are enabled, e.g., by masking speech-derived signals [29] and by masking parameters of probabilistic models [2] (that could generate speech signals; not only recognise biometric identity).

In the past decade, several efforts have been undertaken to foster research on security and privacy in the setting of speech and language technology. Not all of these efforts have been successful. Of the successful initiatives within ISCA and the IEEE Signal Processing Society, which are bridging across fields within and beyond speech technology, one can mention the biannual anti-spoofing challenges coined *ASVspoof* (security of voice biometrics) [6] and the line of PhD students studying privacy-preserving voice biometrics and speech processing [11, 19, 20]. Facing GDPR adoption (in 2016), Interspeech 2015 featured the special event *Privacy Issues in Speech Data Collection and Usage*[5]; shortly after, research on patient privacy started intersecting with secure computation and cryptography leading to the 2018 papers [3, 23]. International projects followed shortly after, such as the H2020 COMPRISE (2018 start), the SECURE research project at Aalborg University (2018 start), and the JST-ANR VoicePersonae (2019 start).

The ISCA SIG-SPSC formed in 2019 at Interspeech, which featured also a special session on *Privacy in Speech and Audio Interfaces* among which's co-organisers was the inaugural chair of SPSC. The SIG is formed not only by people from the above efforts, who met at the (early) 2019 conference IEEE ICASSP. One of the igniting drives for this was created by bringing interdisciplinary experts together through the writing of [15, 16], i.e., co-authors coming from voice biometrics, study of the Law, speech and language technology, secure computation and cryptography, and border control biometrics. To foster its interdisciplinary approach in nurturing multidisciplinary skills of emerging scholars, the SIG was proposed to deliberately become a joint body with other SIGs of ISCA and beyond.

[1] This period spans wrapping up my dissertation to solidifying post-doctoral position abroad, and deciding to stop applying for faculty positions, eventually.

[2] See: www.spsc-sig.org and www.isca-speech.org—ISCA papers are freely online available at www.isca-speech.org/archive, covering decades of research on speech technology and all its technological facets.

[3] Typically, *voice biometrics* is referred to *speaker verification* and *speaker recognition*.

[4] The film industry addressed security and privacy issues regarding 'speech' way earlier; it should not be necessary to mention the Stasimuseum (www.stasimuseum.de) educating about the GDR Ministry for State Security (Stasi) 'listening-in'.

[5] During the panel, a senior researcher mentioned that she found her voice in a movie (taken from a speech synthesis database); being asked is better.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38477

2 WHAT DO I KNOW ALREADY?

My main background is in voice biometrics: are two audio files from the same speaker? Comparison results (scores) are thresholded to make yes/no decisions. Yet, antithetical methodological and diverging goal sets are at play:

- The philosophical debate in statistics on quantifying epistemic uncertainty and aleatory uncertainty [18] results in antithetical perspectives on what performance is. In one, error trade-offs are reported and technology integrators are externalised (e.g., [8]). In the other, the perspective on error trade-offs are the basis to cost and information models (e.g., [5, 21]). The latter framework is also used by forensic practitioners to validate how well they prepare evidence when reporting to a judge/jury (regardless of which beliefs a judge/jury might have) [27].
- In ASVspoof (automatic speaker verification anti-spoofing), fake audio detection is investigated for strengthening voice biometrics. There, tandem systems are composed (biometrics with anti-spoofing) [24]. Going beyond related standards [10], tandem performance assessment interlinks subsystem contribution [12].
- Privacy solutions rooted in cryptography and secure multiparty computation have different drawbacks in computational time and time taken to exchange data between servers in IT infrastructures [26]. This contours the real-time demands posed to modern speech technology. Quantisation of evidence representation is a consequence: down to one decision threshold only is facilitated to maintain usability.
- In the VoicePrivacy challenge [25], the privacy-preserving task is to modify/sanitise speech data from biometric features. From a forensic method validation perspective [17], however, when transferring Shannon's original concepts of 'perfect secrecy' [22], core concepts enabling modern cryptography might not hold for statistics assumed inherently known are what speech experts research on since decades.
- Mindsets outline meaning of words, and through this societal impact of technology. 'Quality' is highly contextual. In biometrics standards [9], quality is effectively viewed as functionality of some factory piece, and the conformity of incoming material. On the contrary, when systems are to compensate environmental changes to retain performance, one cannot avoid taking a holistic approach for making design participatory and anticipatory.

To bring experts from different fields together to foster multidisciplinary skill development meets hesitation: while neat on paper, this is entirely antithetical to the academic economy and the core principles of its currencies (paper citations and research grants for very discrete work, not holistic theory). A new scholarly ecosystem is needed that is not only capable, due its experts talking with one another (not only to), but moreover: one that is productive. The pandemic revealed limiting social dynamics *incentivised*. Beneficial social dynamics in counterplay led to slowly building the SPSC community with its first main event in November 2021.[6]

3 HOW DO I STUDY THE PHENOMENON?

One needs to diverge from the norm, and seek to constantly improve. (Seasonally show-casing yet another 5% improvement while tempting does not cause shifts.) Getting active early on in more than two research communities enables personal growth, only through which collectives can grow, then societies. Settings are necessary that allow for constant dialog across disciplines—each expert has core interests placed differently through their individual experience traces.

One can but bring people at a table. For example, the term *biometric data* was put in late into the GDPR which had technological far reaching impacts—when is speech none biometric, if it is without voice or if it is not influenced by our habits how we speak and like to talk about? The European Data Protection Supervisor (EDPS) published several TechDispatches since July 2019 (public EDPS opinions for technologists, and society at large), and its first one addresses

[6] https://spsc-symposium2021.de/

Smart Speakers and Virtual Assistants[7]. In early 2020, Thomas Zerdick (head of unit *Technology and Privacy* at the EDPS) gave the keynote talk at our concept workshop *Privacy: Speech meets Legal Experts.*[8] In March 2021, the European Data Protection Board (EDPB) put their draft *Guidelines 02/2021 on Virtual Voice Assistants* to public consultation; we commented on this as SIG-SPSC based on interdisciplinary expert discussions. In an upcoming event at the Lorentz Center *Speech as Personal Identifiable Information*, we seek to further bridge between communities, namely, usability, speech and language technology, IT-security, policy and governance, and anthropology.

4 WHAT I WOULD LIKE TO KNOW?

How can we come together and make a difference? As a first step, we might inquire information and knowledge (analysis), yet, what we might be seeking actually is understanding and wisdom (synthesis). How can we nourish one another through mutual care that is productive and efficient for lifelong learners?

5 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

Which methodologies are there to learn from? I want to develop more capacities to better understand human communication, how technology can provide aid, so we can enable progress in societies—not to enforce them to run in circles through making people fit to machine operation. This can be achieved through solving specific problems; these problems can be hypothetical: how to figure out counterfactuals otherwise? For enriching systemic thinking: why and what for attack disciplines problems in their way?

6 WHAT DO I WANT TO TEACH OTHER DISCIPLINES?

We need to substitute teaching with play to stimulate learning through curiosity. Systemic thinking minds are punished by and not rewarded by the present day education megastructure. Compare systems thinking pioneer Russel Ackoff [1]:

- *Creativity is actively suppressed and in most schools conformity—which is anathema to creativity—is valued instead.*
- *Problems do not "belong" to any discipline. [...] The distinction made between science and the humanities (which include the arts) probably does the most harm. [...] They can be viewed and discussed separately, but they cannot be separated.*
- *We should seek wisdom more than anything else—the ability to make value judgments, to know the consequences of our actions, and to learn from our mistakes.*

More crisp in the words of the film critic Wolfgang M. Schmitt: *We only watch, but we do not see.*

A new scholarly ecosystem is in demand. One where young minds are treated as equals. There is no luxury left to afford in unproductive time spent in incentivised first and human rights second activities. *Security and privacy in speech communication* is—as the discrete topic it is—an onboarding; SPSC cannot become truly holistic (it is but speech).

Yet, remedy lies in Deming's Plan-Do-Study-Act cycle [4, 14]. Academic research today got stuck in the Do step, barely completing any analysis of the data that all disciplines generated (the begin of the Study step). How to Act next? To break up silos, to facilitate organising a body of knowledge holistically, a new scholarly ecosystem must emerge.

ACKNOWLEDGMENTS

This work was partially funded by the ANR/JST VoicePersonae project (grant No. JPMJCR18A6).

Thanks to who helped me on this journey, be it providing access to their networks, their collaboration, keeping up regular exchange across the fields, supporting the formation of SPSC, stepping up for service, and leadership.

[7] https://data.europa.eu/doi/10.2804/755512
[8] https://www.spsc-sig.org/2020-01-29-speech-legal-workshop

REFERENCES

[1] R. Ackoff and D. Greenberg. 2008. *Turning Learning Right Side Up: Putting Education Back on Track.* FT Press.

[2] X. Anguera and J.-F. Bonastre. 2010. A Novel Speaker Binary Key Derived from Anchor Models. In *Proc. Interspeech.*

[3] Ferdinand Brasser, Tommaso Frassetto, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, and Christian Weinert. 2018. VoiceGuard: Secure and Private Speech Processing. In *Proc. Interspeech.*

[4] W. E. Deming, K. E. Cahill, and K. L. Allan. 2018. *Out of the Crisis.* The MIT Press.

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

[6] N. W. D. Evans, J. Yamagishi, and T. Kinnunen. 2013. Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics. In *IEEE Signal Processing Society Newsletter.*

[7] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano. 2005. Applications of NAM Microphones in Speech Recognition for Privacy in Human-Machine Communication. In *Proc. Interspeech.*

[8] ISO/IEC JTC1 SC37 Biometrics. 2006. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework.* International Organization for Standardization and International Electrotechnical Committee. confirmed in 2011 and in 2016.

[9] ISO/IEC JTC1 SC37 Biometrics. 2017. *ISO/IEC 2382-37:2017 Information Technology - Vocabulary - Part 37: Biometrics.* International Organization for Standardization.

[10] ISO/IEC JTC1 SC37 Biometrics. 2017. *ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting.* International Organization for Standardization.

[11] A. Jimenez. 2019. *An Information Theoretic Approach for Privacy Preservation in Distance-based Machine Learning.* Ph.D. Dissertation. Carnegie Mellon University.

[12] Tomi Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, Md Sahidullah, J. Yamagishi, and D. A. Reynolds. 2020. Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2195–2210.

[13] J. Lindberg and M. Blomberg. 1999. Vulnerability in Speaker Verification - A Study of Technical Impostor Techniques. In *Proc. Eurospeech.*

[14] R. Moen and C. Norman. 2009. The History of the PDCA Cycle. In *Proc. ANQ Congress.*

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

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

[17] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, and N. Evans. 2020. The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment. In *Proc. Interspeech.* ISCA, 1698–1702. https://doi.org/10.21437/Interspeech.2020-1815

[18] T. O'Hagan. 2004. Dicing with the unknown. *Significance* 1, 3 (2004), 132–133. https://doi.org/10.1111/j.1740-9713.2004.00050.x

[19] M. Pathak. 2013. *Privacy-Preserving Machine Learning for Speech Processing.* Ph.D. Dissertation. Carnegie Mellon University.

[20] J. Portêlo. 2015. *Privacy-preserving frameworks for speech mining.* Ph.D. Dissertation. Universidade de Lisboa.

[21] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez. 2018. Deconstructing Cross-entropy for Probabilistic Binary Classifiers. *Entropy* 20, 3 (3 2018), 208.

[22] C. E. Shannon. 1949. Communication Theory of Secrecy Systems. *Bell System Technical Journal* 28, 4 (10 1949), 656–715.

[23] F. Teixeira, A. Abad, and I. Trancoso. 2018. Patient Privacy in Paralinguistic Tasks. In *Proc. Interspeech.*

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

[25] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. M. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco. 2020. Introducing the VoicePrivacy Initiative. In *Proc. Interspeech.*

[26] A. Treiber, A. Nautsch, J. Kolberg, T. Schneider, and C. Busch. 2019. Privacy-preserving PLDA speaker verification using outsourced secure computation. *Speech Communication* 114 (2019), 60–71. https://doi.org/10.1016/j.specom.2019.09.004

[27] S. E. Willis, L. Mc Kenna, S. Mc Dermott, A. Barrett, B. Rasmusson, et al. 2015. *ENFSI Guideline for Evaluative Reporting in Forensic Science.* European Network of Forensic Science Institutes. [Online] http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, accessed: 2017-05-22.

[28] D. Wyatt, T. Choudhury, and J. Bilmes. 2007. Conversation Detection and Speaker Segmentation in Privacy-Sensitive Situated Speech Data. In *Proc. Interspeech.*

[29] X. J. Zhand, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen. 2008. Frequency-Domain Parameter Estimations for Binary Masked Signals. In *Proc. Interspeech.*