ASVspoof 5 Evaluation Plan*

Héctor Delgado, Nicholas Evans, Jee-weon Jung, Tomi Kinnunen, Ivan Kukanov, Kong Aik Lee, Xuechen Liu, Hye-jin Shim, Md Sahidullah, Hemlata Tak, Massimiliano Todisco, Xin Wang, Junichi Yamagishi
ASVspoof consortium
http://www.asvspoof.org/

June 13, 2024

TL;DR for new participants

- ASVspoof is centred around the challenges to design: (1) spoofing-robust automatic speaker verification (SASV) solutions; (2) stand-alone speech deepfake detectors.
- Join us as challenge participants. Challenge participants receive a database consisting of bona fide and spoofed data collected from a wide range of speakers and an experimental protocol. They apply their deepfake detection and SASV solutions to produce a set of scores and prepare a system description.
- The organisers analyse and rank the results, and present a summary at a future ASVspoof workshop to which challenge participants are encouraged to submit a paper.

... and for readers familiar with ASVspoof

- ASVspoof 5 involves two phases:
  - Phase 1 (closed by January 1st, 2024): we collected spoofing/deepfake attack data from external data contributors who were provided with a spoofing data generation protocol and access to surrogate automatic speaker verification (ASV) and spoofing countermeasure (CM) sub-systems. Spoofing/deepfake attacks are expected to be more adversarial than in previous editions of ASVspoof and to fool both ASV and CM surrogate sub-systems.
  - Phase 2: the traditional challenge has evolved to encompass two tracks. **Track 1** robust speech deepfake (DF) detection in a stand-alone setting without speaker verification, similar in spirit to the DF task in ASVspoof 2021. **Track 2** the development of an SASV system, i.e. an **ASV** system which is robust to spoofing attacks. Participants can either develop a CM system and combine it with a pre-trained ASV system provided by the organisers (like for previous ASVspoof editions) or, alternatively, develop single classifiers or fused ASV and CM systems.

*Document Version 0.4 (June 13, 2024). The order of challenge co-organiser names is alphabetical.

1 Introduction

The automatic speaker verification spoofing and countermeasures (ASVspoof) challenge series is a community-led initiative which aims to promote the consideration of spoofing and speech deepfakes in addition to the development of countermeasures. ASVspoof 5 is the fifth edition in a series of previously-biennial, competitive challenges [1]–[4]. The common goal of each edition is to foster progress in the development of countermeasures, also referred to as spoofing detection and presentation attack detection (PAD) solutions which are capable of discriminating between bona fide and spoofed or deepfake speech utterances.

While previous challenge editions have incorporated distinct logical access (LA), physical access (PA), and speech deepfake (DF) sub-tasks [5], ASVspoof 5 has evolved to encompass two tracks: (i) stand-alone spoofing and speech deepfake detection (no ASV) (ii) spoofing-robust automatic speaker verification (SASV). The latter shares the same goal as the LA sub-task of previous ASVspoof editions, but is extended to allow participants to develop single classifiers (as per the previous SASV challenge) or the fusion of separate ASV and CM sub-systems. As for previous ASVspoof editions, participants may use a pre-trained ASV sub-system provided by the organisers. The key technologies which can be used to generate spoofed/deepfake speech, e.g. speech synthesis and voice conversion, continue to evolve, with tremendous advances having been made in the last few years. It is essential that progress in spoofing/deepfake detection keeps apace.

ASVspoof 5 has two phases: Phase 1 involves the collection of spoofed data, and Phase 2 corresponds to the detection of spoofed data. An overview of ASVspoof 5 phases is illustrated in Figure 1. Phase 1 has been closed by January 1st, 2024, the details of which are presented in the evaluation plan version 0.2. The focus of the current document is upon Phase 2. It will evolve in the coming months to include a description of the evaluation data, metrics, and other details. The challenge aims to benchmark the latest detection solutions in the face of more diverse and adversarial attacks which are designed to compromise not just an ASV sub-system, but also a CM sub-system.

2 Technical objectives

ASVspoof 5 maintains the pursuit of *generalised* countermeasures. These entail spoofing/deepfake detection solutions which perform reliably even in the face of utterances generated with new or previously unseen spoofing attack algorithms and methods. Specific technical objectives for ASVspoof 5 are to:

- evaluate the threat of spoofing attacks when spoofed utterances are generated using non-studio-quality data;
- improve reliability in the face of stronger attacks specifically designed, adapted or optimised to compromise not just ASV sub-systems, but also CM sub-systems;
- facilitate the comparison of separate ASV/CM sub-systems (separately or jointly optimised) and single/integrated SASV solutions;
- evaluate reliability when training protocol restrictions are relaxed to allow the use of external data (disjoint from the predefined test dataset) or pre-trained models learned with external data.

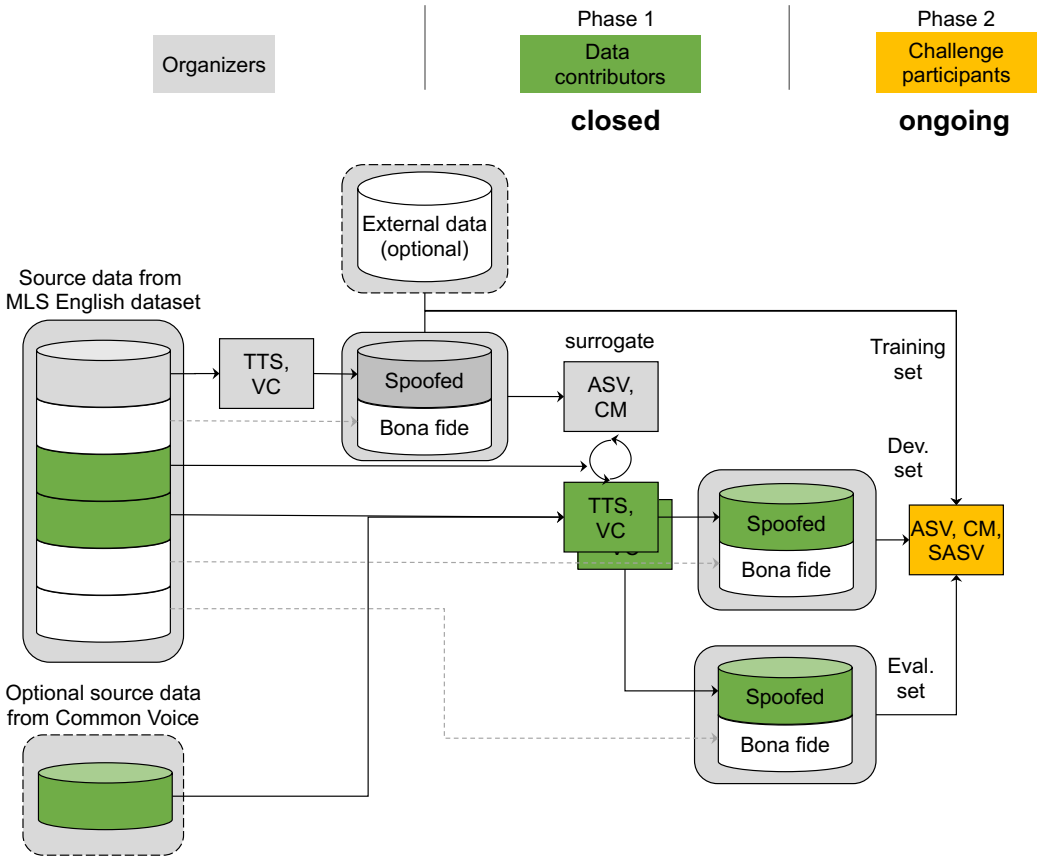

Figure 1: An overview of ASVspoof 5 phases reflecting the roles of organisers (gray-coloured components), data contributors (green) and challenge participants (yellow).

3 Source database

To support the objectives outlined above, ASVspoof 5 transitions to a new source dataset, namely the *Multilingual Librispeech* (MLS) dataset (English-language subset). It contains speech recordings collected from a large number of speakers in a variety of different recording conditions [6]. The MLS dataset is partitioned into disjoint subsets which support the development of text-to-speech (TTS) and voice conversion (VC) models as well as CM, ASV and SASV systems. Phase 1 data contributors may also utilise a subset of utterances selected from the English Common Voice Corpus 11.0 [7] for the training of speaker encoders.

4 Guidelines for Phase 2 challenge participants

4.1 Tracks and metrics

ASVspoof entails two challenge tracks (summarized in Table 1).

Track 1 consists of a stand-alone speech deepfake (bonafide vs spoof) detection task. In continuation of the task for ASVspoof 2021 [8], the vision is to address the generalization and robustness

of spoofed and deepfake speech detection. Evaluation utterances will be of varied technical quality stemming from the potential application of speech coding, audio compression, bandlimiting or other processing algorithms. Participants are tasked with the design of detection systems which should assign a single real-valued detection score to each unlabelled audio utterance. Track 1 evaluation metrics include the *minimum detection cost function* (minDCF), the *cost of log-likelihood ratio* (C_{llr}) [9] and the *equal error rate* (EER). The costs and priors for calculation of the minDCF will be specified in a forthcoming update to this document.

Track 2 entails a spoofing-robust automatic speaker verification (SASV) task [10]. SASV systems must compare an unlabeled probe (test) utterance to an enrollment utterance(s) of known target speakers. SASV systems developed by participants will be benchmarked using a mix of three types of trials—*target bonafide* (target), *bonafide non-target* (non-target), and *spoofed non-target* (spoof). SASV systems should accept target trials only. The primary evaluation metric for Track 2 is the new *agnostic* DCF (a-DCF) [11] which assigns a cost to a system, taking into account the miss rate and two false alarm rates (relating to non-target or spoof trials). As for Track 1, costs and priors for calculation of the a-DCF will be specified in a forthcoming update to this document. Track 2 participants may take part in two different ways:

1. **Tandem SASV using a reference ASV sub-system** – Participants develop stand-alone spoofing detectors (countermeasure) which will be combined with a reference ASV sub-system provided by the challenge organisers. CM and ASV sub-systems are combined using a tandem approach. In addition to the primary metric, a-DCF, submissions will also be evaluated using the *minimum tandem detection function* (t-DCF) [12] and the recently-introduced *tandem equal error rate* (t-EER) [13]. This is the familiar *logical access* (LA) sub-task of previous ASVspoof editions.

2. **Arbitrary SASV systems** – Participants develop a full SASV system using any custom/preferred architecture (tandem, score fusion, embedding fusion, end-to-end, *etc*). SASV systems of this variety should output a single detection score which reflects support for the target speaker hypothesis. Submissions will be evaluated using the a-DCF.

4.2 Conditions

There are two conditions, *open* and *closed*, for both **Track 1** and **Track 2**:

1. **Closed condition** – Participants commit to using data within the ASVspoof 5 training partition and Voxceleb2 dataset only:

   - **Track 1:** The ASVspoof 5 training partition must be used for the training of spoofing/deepfake detection systems.
   - **Track 2:** The ASVspoof 5 training partition and the Voxceleb2 dataset can be used for the training of SASV systems.

2. **Open condition** – Subject to specific conditions, participants may, in addition to the data available under the closed condition, use external data and pre-trained models. The use of any such additional resources **must** be described clearly in the system description accompanying any submission. Participants must **not** use under any circumstances any data or systems trained using data which overlaps with that used in the creation of ASVspoof 5 data. This restriction applies to the use of identical utterances (speech recordings), and also applies to use of non-identical utterances collected from the same speakers. It also applies to the use of any external data or resources for which there is a data overlap, or for which the potential overlap

	Track 1	Track 2
Task	Stand-alone deepfake detection	Spoofing-robust ASV
Scenario	Generic	Telephony/VoIP
Classes	`bonafide (real)*, spoof (deepfake)`	`target*, nontarget, spoof`
Decisions	`ACCEPT, REJECT`	`ACCEPT, REJECT`
Metrics	minDCF (**primary**), actDCF, C_{llr} [9], EER	min a-DCF [11] (**primary**), min t-DCF [12], t-EER [13]
Required scores	1 (deepfake detection)	1 (SASV accept/reject)
Optional scores	—	CM, ASV subsystem scores

Table 1: Summary of the detection scenarios and system assumptions in the two main tracks of ASVspoof 5. For 'classes', star (*) indicates the 'positive' class, associated with larger detection score values. Ideally, the positive class instances should be accepted, the other classes rejected.

Table 2: Phase 2 protocol and meta data files

	File name	Usage
Training	`ASVspoof5.train.metadata.txt`	Meta-data of training set utterances
Development	`ASVspoof5.dev.metadata.txt`	Meta-data of development set utterances
	`ASVspoof5.dev.trial.txt`	ASV protocol
	`ASVspoof5.dev.enroll.txt`	ASV target speaker enrollment data list
Evaluation	to be released	to be released

with ASVspoof 5 data cannot be ascertained. Use of any such data or resources is strictly prohibited. **Prohibited** data resources include (but are not limited to) any data contained within the LibriLight or MLS English datasets, and the MUSAN corpus speech subset (as they overlap with the evaluation data); this restriction applies to any other dataset, pre-trained model, or any other resource derived from these datasets. **Permitted** data resources include, e.g., CommonVoice and previous ASVspoof challenge databases generated from VCTK source databases. In both cases, there is no utterance or speaker overlap with ASVspoof 5 data. Use of the LibriSpeech dataset **is** permitted, as is the use of models pre-trained using the LibriSpeech dataset. Use of LibriSpeech data and derived resources is only permitted because, by design, there is no speaker overlap with data contained within the ASVspoof 5 database.

4.3 Rules

- The pooling of data within training and development partitions, e.g. to create a larger training set, is strictly prohibited.
- For the **closed condition**, the use of external non-speech and non-voice resources (e.g. noise samples and impulse responses), speech codecs and audio compression tools for training or data augmentation IS permitted, under the condition that their use is detailed in the system description accompanying any submission. Any audio samples originating from human vocal production organs (e.g. singing, hum, cry, laughter, whistling, laughter, coughs, sneezing etc) as well as their combinations (e.g. choir, simulated babble) are NOT permitted.
- For the **open condition**, additional speech data for training or data augmentation purposes IS allowed, subject to the conditions specified in Section 4.2.
- Each test sample must be scored independently of each other; techniques such as domain

adaption using the evaluation segments, or any use of evaluation data for normalisation (across multiple or all test samples) purposes is NOT allowed.

- Participants are required to submit scores for development and evaluation data on the CodaLab platform. Details will be added in a forthcoming update to this document.
- Teams are required to submit a detailed system description relating to their final submission (1 system description per team). A system description template in the form of a high-level annotation will be provided and will serve as an indicator of the level of detail required. Compliance with the template and high-level annotation is mandatory and is in everyone's interest. System descriptions in such a common format will be essential to meaningful post-evaluation analysis and will help everyone to put different approaches into context with one another (to promote open science and a better, common understanding).
- Participants must not make public comparisons to the results or rankings of competing teams. This rule applies also to the disclosure of a participant's own *ranking* which is, by definition, a comparison to the results and rankings of competing teams. It applies also to claims of being the 'challenge winner', 'top-ranked system', 'leading team', or any similar expressions which may be interpreted in any way as a comparison to the results of other participants. Participants can, however, publish their *own results*. A summary of the challenge results will also be prepared by the organisers. Any participant wishing to retain anonymity should provide an anonymous team identifier with their registration. The organisers reserve the right to adjust the policies described above at any point until the ASVspoof workshop.
- Participants commit to respecting the guidelines detailed earlier in Sections 4.1 and 4.2. Track, condition and data/resource use guidelines should also be interpreted as challenge rules. If a participants is uncertain as to whether any particular data or resource is permitted, they are advised to seek clarification from the organisers.

The organisers reserve the right to exclude systems from ranking and to exclude teams from future participation in ASVspoof in the event that the above rules are not adhered to.

5 Baseline systems

Baseline systems will be provided for both tracks. The organisers will also provide a reference ASV system for optional use in Track 2. Details will be added to a forthcoming update to this document.

6 Ethics

ASVspoof 5 is committed to upholding ethical standards and to responsible research practices. Our objective is to enhance the security and reliability of automatic speaker verification (ASV) technology by promoting collaboration and progress in the development of robust spoofing/deepfake detection solutions, to promote participation and to protect the interests of stakeholders.

Data contributors and challenge participants are requested to adhere to local data protection regulations when processing speech data. We ask that you conduct your research and development activities in a responsible manner and be mindful of the potential for the misuse of software solutions and results. You are expected to disclose identified vulnerabilities or weaknesses in ASV technology in a responsible manner. The prompt and appropriate reporting of such findings is the shared

responsibility of all in our community, contributes to the improvement of security systems and protects against potential misuse.

The ASVspoof organisers explicitly disassociate themselves from any association with or endorsement of hacking activities, unauthorized access attempts, or the creation of spoofs/deepfakes for malicious purposes or personal gain. We strictly condemn any misuse of the knowledge and tools developed through ASVspoof. Any malicious use of challenge outcomes, results or findings is strictly prohibited.

Please note that the ISCA Code of Ethics for Authors, available at https://isca-speech.org/Code-of-Ethics-for-Authors, applies to all research publications and reports originating from the ASVspoof initiative and challenge series.

7 General mailing list

Phase 2 participants and team members are encouraged to subscribe to the general mailing list. Subscribe by sending an email to:

sympa@lists.asvspoof.org

with "subscribe ASVspoof5" as the subject line. Successful subscriptions are confirmed by return email. In the case that confirmation emails are not received, please add both of the following email addresses to your whitelist:

asvspoof5-request@lists.asvspoof.org
asvspoof5@lists.asvspoof.org

To post messages to the mailing list itself, emails should be addressed to:

asvspoof5@lists.asvspoof.org

Subscribers can unsubscribe at any time by following the instructions in the subscription confirmation email.

8 Registration and submission of results

8.1 Registration

Challenge participants are required to register their interest through the following form:

https://shorturl.at/cqrtK

After we have confirmed the registration, a link to download the data package will be sent to the registrant.

8.2 Results submission

The ASVspoof 5 evaluation platform is implemented using CodaLab [14],[1] an open-source web platform for the organisation and hosting of competitions and challenges in data science. Participants will need to upload score files to the CodaLab website. Upon successful uploads, participants will

[1] https://codalab.org

receive a detailed report of system performance. Details will be announced upon the opening of the CodaLab platform (see schedule in Section 9). The two challenge tracks and the open/closed conditions will be organised under separate CodaLab competition sites. Access links will be shared with registered participants.

In order to access each site, participants will be required to create an account on the CodaLab website. This will enable you to request access to each competition site. **Teams should make only one request per task.** Since CodaLab access requests will be manually validated by the challenge organisers, **requests should be made in the name of the team's registered contact person**. Similar to the preceding challenge edition, the leaderboard is anonymous; usernames will not be disclosed to other participants.

The challenge will run in two phases, a *progress* phase and the main *evaluation* phase. During the progress phase, each team may make up to 3 submissions per day. Results determined from a subset of trials will be made available for each submission. During the evaluation phase, participants will be allowed only a single submission for which results will be determined from the full set of evaluation trials. The schedule is shown in Section 9. Results for 4 reproducible baselines will be displayed on the CodaLab site.

9 Schedule

A tentative schedule is as follows:

Challenge

- Initial release of eval plan: May 20, 2024
- Phase 2
  - registration opens: May 20, 2024
  - training and development data released: May 20, 2024
  - Phase 2 CodaLab platform opens: June 05, 2024
  - evaluation data released: June 17, 2024
  - registration deadline: July 10, 2024
  - deadline to submit scores: July 17, 2024

Workshop

- Paper submission: July 31, 2024
- Acceptance notifications: August 10, 2024
- ASVspoof 5 Workshop, Interspeech 2024: August 31, 2024

10 Glossary

Generally, the terminologies of automatic speaker verification are consistent with that in the NIST speaker recognition evaluation. Terminologies more specific to spoofing and countermeasure assessment are listed as follows:

Spoofing attack: An adversary, also named impostor, attempts to deceive an automatic speaker verification system by impersonating another enrolled user in order to manipulate speaker verification results.

Anti-Spoofing: Also known as countermeasure. It is a technique to countering spoofing attacks to secure automatic speaker verification.

Bona fide trial: A trial in which the speech signal is recorded from a live human being without any modification.

Spoof trial: In the case of the physical access, a spoofing trial means a trial in which an authentic human speech signal is first played back through an digital-to-analog conversion process and then re-recorded again through analog-to-digital channel; an example would be using smartphone A to replay an authentic target speaker recording through the loudspeaker of A to the microphone of smartphone B that acts as the end-user terminal of an ASV system. In the case of the logical access, a spoofing trial means a trial in which the original, genuine speech signal is modified automatically in order to manipulate ASV.

11 Acknowledgements

The ASVspoof 5 consortium expresses its gratitude to the following individuals who generated spoofed data: Cheng Gong, Tianjin University, China; Guo Hanjie and Liping Chen, University of Science and Technology of China, China; Myeonghun Jeong, Seoul National University, South Korea; Florian Lux, University of Stuttgart, Germany; Sebastien Le Maguer, University of Helsinki, Finland; Nicolas Muller, Fraunhofer AISEC, Germany; Soumi Maiti, Carnegie Mellon University, USA; Vishwanath Pratap Singh, University of Estern Finland, Finland; Chengzhe Sun, Shuwei Hou, Siwei Lyu, University at Buffalo, State University of New York, USA; Yihan Wu, Renmin University of China, China; Ge Zhu, University of Rochester, USA; Wangyou Zhang, Shaohai Jiaotong University, China.

References

[1] Z. Wu, T. Kinnunen, N. Evans, *et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041. DOI: 10.21437/Interspeech.2015-462.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017.

[3] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019, pp. 1008–1012. DOI: 10.21437/Interspeech.2019-2249.

[4] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54. DOI: 10.21437/ASVSPOOF.2021-8.

[5] X. Liu, X. Wang, M. Sahidullah, *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–14, 2023. DOI: 10.1109/TASLP.2023.3285283.

[6] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Proc. Interspeech*, 2020, pp. 2757–2761. DOI: 10.21437/Interspeech.2020-2826.

[7] R. Ardila, M. Branson, K. Davis, *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proc. LREC*, May 2020, pp. 4218–4222.

[8] X. Liu, X. Wang, M. Sahidullah, *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023. DOI: 10.1109/TASLP.2023.3285283.

[9] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006, Odyssey 2004: The speaker and Language Recognition Workshop, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2005.08.001.

[10] J.-w. Jung, H. Tak, H.-j. Shim, *et al.*, "SASV 2022: The first spoofing-aware speaker verification challenge," in *Proc. Interspeech*, 2022.

[11] H.-j. Shim, J.-w. Jung, T. Kinnunen, *et al.*, "a-DCF: An architecture agnostic metric with application to spoofing-robust speaker verification," in *Proc. Speaker Odyssey*, To appear, 2024. arXiv: 2403.01355 [eess.AS].

[12] T. Kinnunen, H. Delgado, N. Evans, *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. DOI: 10.1109/TASLP.2020.3009494.

[13] T. H. Kinnunen, K. A. Lee, H. Tak, *et al.*, "t-EER: Parameter-free tandem evaluation of countermeasures and biometric comparators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2622–2637, 2024. DOI: 10.1109/TPAMI.2023.3313648.

[14] A. Pavao, I. Guyon, A.-C. Letournel, *et al.*, "Codalab competitions: An open source platform to organize scientific challenges," *Technical report*, 2022. [Online]. Available: https://hal.inria.fr/hal-03629462v1.