ASVspoof 5 Evaluation Plan*

Héctor Delgado, Nicholas Evans, Jeeweon 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/

September 28, 2023

TL;DR for new participants

- ASVspoof is centered around the challenges to design: (1) spoofing-robust automatic speaker verification (ASV) solutions; (2) application-agnostic speech deepfake detectors.
- Join us **either** as data providers (phase 1) **or** as challenge participants (phase 2). **Data providers** use a source database and a protocol to generate spoofed utterances. **Challenge participants** receive a database consisting of bona fide and spoofed data and an experimental protocol. They apply their spoofing/deepfake detection 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: we aim to collect spoofing/deepfake attack data from external data contributors who will be provided with a spoofing data generation protocol and access to surrogate automatic speaker verification and spoofing countermeasure sub-systems. Spoofing/deepfake attacks are expected to be more adversarial than in previous editions of ASVspoof and to fool both automatic speaker verification (ASV) and countermeasure (CM) surrogate sub-systems.
  - Phase 2: the traditional challenge is extended to include the optimisation of both ASV and CM sub-systems, with the option to focus only upon the latter through the use of a default ASV sub-system. It also incorporates a Spoofing-aware Speaker Verification (SASV) task to facilitate the design of single-classifier, integrated CM-ASV solutions. ASVspoof 5 also contains both fixed and flexible training conditions.
- ASVspoof 5 involves logical access and speech deepfake tasks and use of an entirely new source database (Multilingual Librispeech and, optionally, Common Voice) and new experimental protocols for the training, development and evaluation of CM/ASV and SASV solutions.

*Document Version 0.1 (September 28, 2023). 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 focuses upon a common LA and DF task. 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 will be conducted in two phases:

- **Phase 1** is the task of data contributors and entails the generation and collection of spoofed utterances which succeed in compromising the reliability of automatic speaker verification (ASV) and spoofing countermeasure (CM) sub-systems. A database collected through Phase 1 will be used by challenge participants in Phase 2.
- **Phase 2** is the task of challenge participants and has two tracks. The first track encompasses the usual approach comprising the preparation of separate CM and ASV subsystems. Participants will be able to use ASV scores generated using a system provided by the organisers or, alternatively, they may design and use their own ASV system. The second track is geared towards the evaluation of spoofing-aware speaker verification (SASV) systems [6], namely single-classifier, integrated solutions. For both tracks, participants are invited to process a database containing an undisclosed mix of bona fide and spoofed utterances and to submit scores to an online ASVspoof 5 leaderboard.

ASVspoof 5 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. An overview of ASVspoof 5 phases is illustrated in Figure 1. This document provides a technical description of the ASVspoof 5 challenge. The focus is currently upon Phase 1 and the task of data contributors. It will evolve in the coming months to include a description of the forthcoming ASVspoof 5 database, the training, development and evaluation partitions, experimental protocols, metrics, evaluation rules, submission procedures, and the Phase 2 schedule.

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;

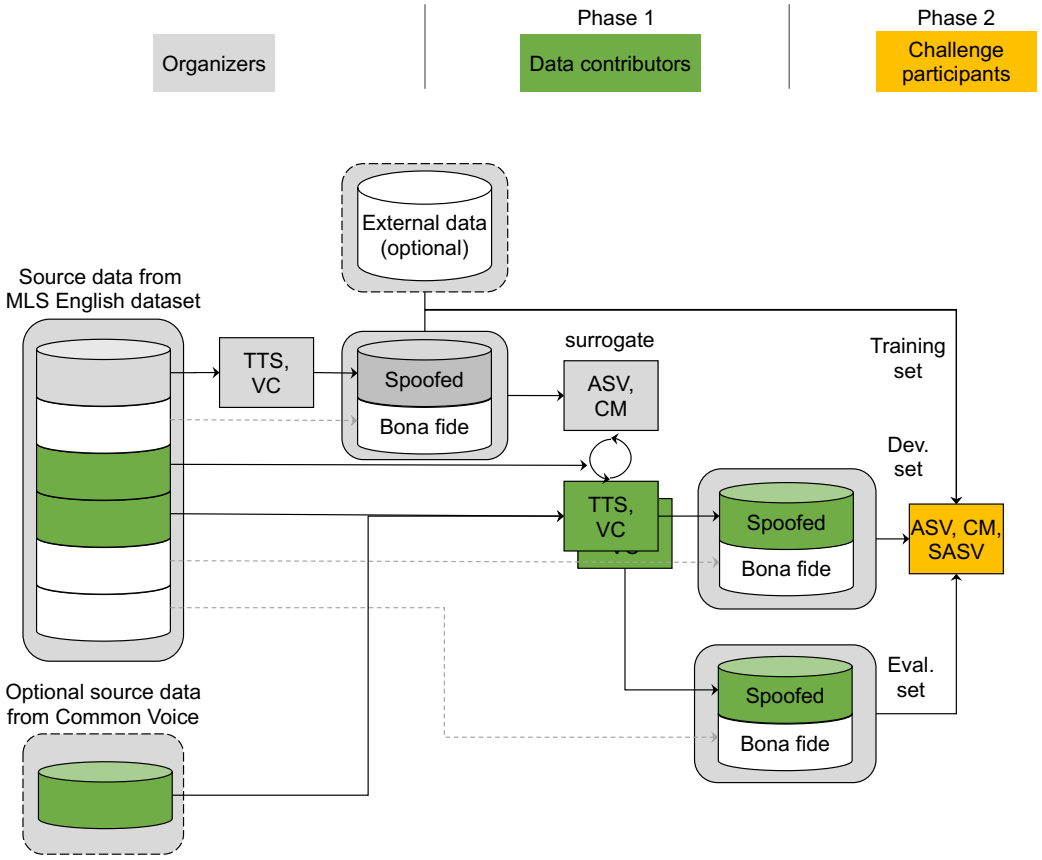

Figure 1: An overview of ASVspoof 5 phases reflecting the roles of organisers (gray-coloured components), data contributors (green) and challenge participants (yellow).

- 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.

3 Scenario and 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 [7]. 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 [8] for the training of speaker encoders.

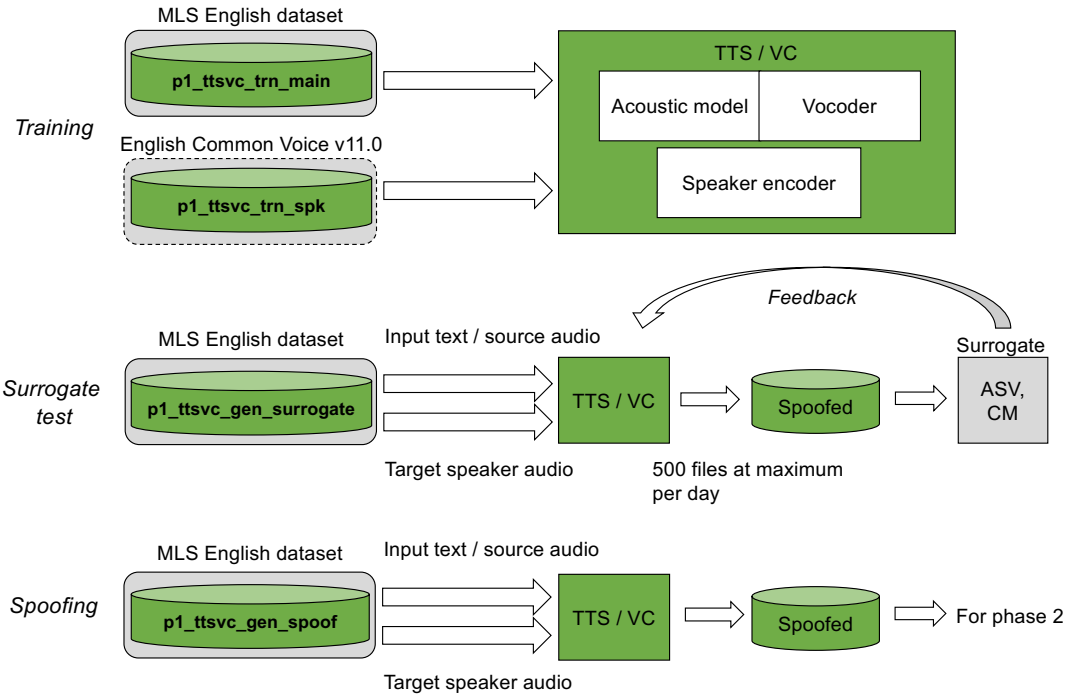

Figure 2: Steps of phase 1.

4 Phased data collection and evaluation

For all prior ASVspoof challenge editions held between 2015 and 2021, the organisers verified that spoofing attack data were successful in manipulating an ASV sub-system. Only modest attention was paid to verify that the same attacks were successful in fooling CM sub-systems, e.g. the better performing CM solutions reported in previous challenge editions. Stronger, more adversarial attacks can be designed to overcome both ASV and CM sub-systems, and will likely provide a stronger test of reliability.

To support the exploration of a stronger attack model as well as to continue the policy of collecting data from external data contributors, ASVspoof 5 is organised into the two phases outlined in Section 1. They are described further in the following. More elaborate details for Phase 2 will follow in future updates to this document.

4.1 Phase 1: Spoofed data collection

The organisers will work with groups of external data contributors during Phase 1 to collect a substantial quantity of spoofed data. This data will form the ASVspoof 5 database to be used by participants in Phase 2. New to ASVspoof 5 are *surrogate* ASV and CM systems. Provided by the organisers, surrogate systems can be used by data contributors during Phase 1 for the design, adaptation or optimisation of spoofing attacks. Data contributors will be provided with a set of protocols for the *training* of spoofing attack algorithms, their evaluation using *surrogate testing*, and a full *spoofing* protocol for data generation and submission. The process to be followed by data contributors is shown in Figure 2 and described below while the specific protocols and source data are summarised in Table 1.

Table 1: Phase 1 protocols. Use of protocols in gray color is optional and can be provided to data contributors upon request. Asterisk (*) is a placeholder whose value is either female or male.

	Protocol file name	Source data	Usage
Training	`p1_ttsvc_trn_main.lst`	MLS	training TTS / VC
	`p1_ttsvc_trn_spk.lst`	Common Voice	(optionally) training speaker encoder
Surrogate test	`p1_ttsvc_surrogate.tsv`	MLS	generating spoofed data for surrogate models
	`p1_asv_*_surrogate.trn`	MLS	surrogate ASV enrollment data
	`p1_asv_*_surrogate.trl`	MLS	surrogate ASV evaluation data
	`p1_cm_surrogate.trl`	MLS	surrogate CM evaluation data
Spoofing	`p1_ttsvc_gen_spoof.tsv`	MLS	generating spoofed data for phase 2

1. **Source database download**: data contributors should download the source databases which contain the audio files to be processed as well as corresponding text transcriptions:
   - MLS database [7], English, https://www.openslr.org/94 (mls_english.tar.gz, flac) and, optionally,
   - Common Voice Corpus 11.0, English [8] https://commonvoice.mozilla.org/en/datasets, if needed for the training of speaker encoders.
2. **Phase 1 Training:** data contributors should use MLS English data in `p1_ttsvc_trn_main.lst` for the training of TTS and VC models. Common Voice data in `p1_ttsvc_trn_spk.lst` can be used optionally for the training of speaker encoders [9].
3. **Phase 1 Surrogate Test:** data contributors should then use their TTS and VC models to generate spoofed data by following the surrogate test protocol `p1_ttsvc_surrogate.tsv`. The protocol specifies source utterances for VC, text input for TTS,[1] and a list of target speaker utterances, which can be used for TTS or VC model adaptation or for the extraction of speaker embeddings for zero-shot generation. Data contributors can optimise attack algorithms using an online platform (see details in Section 7) to which a modest quantity of spoofed data (up to 500 files per upload, one upload per day) can be submitted for evaluation. The selection of 500 files is left to data contributors. The platform will return a set of scores for each submission: surrogate CM scores; surrogate ASV scores; predicted mean opinion scores (MOS) [10]. Per-utterance scores are provided for submitted spoofed utterances. Comparable reference scores for bona fide utterances (CM) and target trials (ASV) will also be provided. MOS scores are provided to facilitate the optimisation of TTS and VC in terms of perceptual quality. Scores are generated using a set of protocols: surrogate CM evaluation (`p1_cm_surrogate.trl`); surrogate ASV enrollment (`p1_asv_female_surrogate.trn`, `p1_asv_male_surrogate.trn`); surrogate ASV evaluation (`p1_asv_female_surrogate.trl`, `p1_asv_male_surrogate.trl`). Data contributors do not need to use these protocols themselves. They are used by the automatic online evaluation platform to compute scores, but they can also be provided to data contributors upon request. Last, while each data contributor may provide utterances generated with different attack algorithms (e.g. one TTS algorithm and one VC algorithm, or two different TTS algorithms), the submission quota will be the same for each data contributor, no matter what the number of spoofing algorithms they provide.

[1] The protocol hence lists the ID of the MLS utterance, the text transcription of which is used as TTS input.

4. **Phase 1 Spoofing**: once an attack algorithm is fixed, data providers should then use the full spoofing protocol `p1_ttsvc_gen_spoof.tsv` to generate a complete set of spoofed utterances. These data should then be uploaded to a URL provided by the organisers.

4.2 Phase 2: Evaluation

Phase 2 is the traditional ASVspoof evaluation with which past participants will be familiar. Notwithstanding the differences outlined above, Phase 2 shares similarities with both ASVspoof 2021 and SASV 2022 challenges. Participants will be provided with two evaluation protocols, with the first defining a progress set, a subset of the full evaluation set. Participants will be able to submit to the evaluation platform a limited number of scores for the progress set prior to the final evaluation submission deadline for which scores will be required for the full evaluation set. Further details will be announced in future updates to this document.

5 Surrogate CM/ASV

Surrogate CM and ASV solutions are provided to data contributors as black-box references to assist in the design, adaptation or optimisation of attack algorithms. Surrogates have been trained using a modest quantity of bonafide and spoofed data, with the latter having been provided by selected external data providers and generated with a held-out subset of MLS data. Data contributors who wish to optimise their attack algorithms in a white-box setting or who wish to exceed the quota for online evaluation may request access to open-sourced CM and ASV surrogates.

Surrogate CMs include AASIST [11], RawNet2 [12], LCNNs with LFCC and other front-end features similar to the ASVspoof 2021 baselines, and CMs based on Wav2vec2.0 (XLSR-53) [13] (to be confirmed).

The surrogate ASV system is based on deep neural speaker embeddings followed by a probabilistic linear discriminant analysis (PLDA) based scoring back-end. The neural speaker embedding extractor is trained using the VoxCeleb2 database. The scoring backend is trained on the same dataset, and then adapted to the data condition using a subset of the MLS training partition.

6 Guidelines for Phase 1 data contributors

The following provides some guidelines as to what is expected from data contributors.

- Data contributors in Phase 1 cannot subsequently participate as *regular* challenge participants in Phase 2; insider knowledge would amount to an unfair advantage. Data contributors may nonetheless still submit scores in Phase 2, but their results will not be ranked alongside those of regular participants.

- The organisers will need a description of each attack algorithm and will circulate guidelines for the preparation of minimal attack algorithm descriptions. While we do not wish to burden data contributors unnecessarily, attack descriptions will be critical to data selection and the design of the evaluation database for Phase 2. They will also form the basis of attack descriptions to be included in a database description journal article, similar to that for the ASVspoof 2019 database [14], to which selected data contributors will be invited to contribute as co-authors. The selection will be subject to technical suitability/relevance (attack potential and detectability).

- Data contributors commit to using *only* data specified in the ASVspoof 5 Phase 1 protocol to design and optimise attack algorithms. The use of *any* other external *speech* dataset is strictly forbidden; this includes, but is not limited to, the use of any other public or in-house corpora, found speech or spoofed speech samples, externally trained models, feature vectors, speaker encoders (or other statistical descriptors) extracted from external data, or externally trained speech activity detectors.
- The use of external *non-speech* resources (e.g. noise samples and impulse responses from databases such as MUSAN), speech codecs and audio compression tools for training or data augmentation *is* permitted as long as their use is declared.
- Data contributors are permitted to partition the ASVspoof 5 Phase 1 training data as they wish (e.g. to create a speaker-disjoint development subset, or to derive early stopping criteria etc.)

7 Evaluation Platform

The ASVspoof 5 evaluation platform will be used for both Phase 1 and Phase 2 and is implemented using CodaLab [15],[2] an open-source web platform for the organisation and hosting of competitions and challenges in data science. The following describes use of the platform for Phase 1 only. Its use for Phase 2 will be described in future updates to this document.

Phase 1 data contributors will need to use CodaLab to upload spoofed data for surrogate testing. Uploads in the form of a single, zipped archive should be made to the CodaLab web platform.[3] Zipped archives should be prepared according to the instructions described below.

- All audio files should have a **.flac** extension and should not exceed a duration of **20 seconds**.
- All files should have a **16 kHz** sample rate with **16 bits** per sample and **PCM** encoding. The encoding can be verified using the `soxi` utility:

```
soxi file.flac
>
  Input File     : 'file.flac'
  Channels       : 1
  Sample Rate    : 16000
  Precision      : 16-bit
  Duration       : 00:00:05.32 = 85121 samples
  File Size      : 84.3k
  Bit Rate       : 127k
  Sample Encoding: 16-bit FLAC
```

- The maximum number of files per submission is **500**.
- The size of the archive should not exceed **300 MB** (platform limitation).
- The archive should be compiled without a parent directory structure, e.g. using a command equivalent to:

[2] https://codalab.org
[3] URLs for data uploads will be communicated to data contributors by email.

```
zip -r submission.zip -j submission  # w/o parent directory
```

where the submission folder is in the local directory and contains only audio files to be added to the archive.

After successful submission, data contributors will be able to download a `scores.tsv` file corresponding to the submission. It contains ASV, CM, and MOS scores in the following format:

```
filename                  asv_1      .    cm_1       .    mos
10108-9714_000388.flac    0.04002    .    1.16853    .    3.66776
10108-9714_000094.flac    0.07474    .    -0.1583    .    3.48259
10108-9714_000062.flac    0.03587    .    -0.9105    .    3.39509
...
```

where `asv` and `cm` fields indicate evaluation scores for the indicated audio files produced by surrogate ASV and CM systems and where `mos` scores are predicted mean opinion scores [10] (see Sections 4.1 and 5 for details).

8 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, available at https://www.isca-speech.org/iscaweb/index.php/about-isca?id=279, applies to all research publications and reports originating from the ASVspoof initiative and challenge series.

9 Registration

Phase 1 data contributors and Phase 2 challenge participants are required to register their interest. The following describes the actions necessary to register as data contributors. The registration procedure for challenge participants will be added in future updates to this document.

We invite expressions of interest from teams and individuals wishing to contribute to the Phase 1 data collection effort. We welcome diverse contributions including spoofed data generated using state-of-the-art algorithms as well as legacy algorithms, in addition to attacks of varying strength

(varying success in manipulating surrogate CM and ASV systems). To initiate discussions, interested parties should first check the guidelines described in Section 6 and then contact the organisers by email at organisers@lists.asvspoof.org with brief details of the algorithms envisaged for spoof/deepfake data generation. Access to protocols, data and the online evaluation platform will be communicated subsequently to registered data contributors.

10 Schedule

A tentative schedule is as follows:

Challenge

- Initial release of eval plan: July 1, 2023
- Phase 1
  - registration opens: July 1, 2023
  - training and development data available: July 1, 2023
  - TTS/VC adaptation and input data available: July 1, 2023
  - surrogate ASV/CM available: July 15, 2023
  - Phase 1 CodaLab platform opens: July 15 to October 15, 2023
  - submit TTS/VC spoofed data: October 22, 2023
- Phase 2
  - TBC: beginning last quarter, 2023

Workshop

- TBC: second quarter, 2024

11 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.

12 Acknowledgements

The ASVspoof 5 consortium expresses its gratitude to the following individuals who generated spoofed data for the training of CM and ASV surrogates: Myeonghun Jeong, Seoul National University, South Korea; Soumi Maiti, Carnegie Mellon University, USA; Ge Zhu, University of Rochester, USA.

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*, 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, Annual Conf. of the Int. Speech Comm. Assoc.*, 2017, pp. 2–6.

[3] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech 2019*, 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] J.-w. Jung, H. Tak, H.-j. Shim, *et al.*, "SASV 2022: The first spoofing-aware speaker verification challenge," in *Proc. Interspeech*, 2022.

[7] 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.

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

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[10] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8442–8446. DOI: 10.1109/ICASSP43922.2022.9746395.

[11] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. ICASSP*, IEEE, 2022, pp. 6367–6371.

[12] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[13] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430. DOI: 10.21437/Interspeech.2021-329.

[14] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101 114, Nov. 2020, ISSN: 08852308. DOI: 10.1016/j.csl.2020.101114.

[15] 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.