SASV Challenge 2022: A Spoofing Aware Speaker Verification Challenge Evaluation Plan

Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee,
Soo-Whan Chung, Hong-Goo Kang, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen

19 January 2022[*]

1 Introduction

The performance of automatic speaker verification (ASV) systems has improved dramatically in recent decades [1–4]. Even for some *in the wild* scenarios, state-of-the-art systems can deliver equal error rates (EERs) of less than 1% [5, 6]. However, these impressive results are typically derived without the consideration of spoofing attacks, namely specially crafted utterances generated by adversaries in order to deceive the ASV system and to provoke false accepts. Even state-of-the-art ASV systems can be vulnerable to spoofing attacks generated using speech synthesis / text-to-speech (TTS), voice conversion (VC) or replay attacks. Some such attacks can degrade ASV reliability considerably [7].

Led by the ASVspoof initiative and corresponding challenge series, countermeasure (CM) systems have hence been developed in order to help detect and deflect spoofing attacks [8–10]. In the case of a logical access, telephony scenario involving only TTS and VC attacks, the best performing spoofing CM systems can deliver EERs of less than 2% [11–21].

This measure of performance only reflects that of the CM, however, whereas it is the reliability of the ASV system which is of primary importance. This can remain poor, even when it operates in tandem with a strong CM [22]. While the minimum tandem detection cost function (t-DCF) [23] reflects the impact of spoofing attacks and CMs upon the ASV system, the ASVspoof challenge series focuses on the development of CMs for a **fixed** ASV system with a pre-determined operating point. We argue that better performance can be delivered when CM and ASV subsystems are both optimised. Herein lies the difference between ASVspoof and the new **S**poofing-**A**ware **S**peaker **V**erification (SASV) challenge. SASV extends the focus of ASVspoof upon CMs to the consideration of integrated systems where both CM *and* ASV subsystems are optimised to improve reliability.

2 Challenge objectives

The goal of the new SASV challenge is hence to further improve robustness to both zero-effort impostor access attempts and spoofing attacks by providing a framework to support the optimisation of CM and ASV systems operating in tandem and, ultimately, facilitate the development of single integrated systems. With only relatively little previous work in this direction [24–28], the objectives of the first challenge are to:

- bridge the gap between the study of ASV and CM systems, and corresponding research communities;
- extend the ASV scenario to take spoofing attacks into account;

[*] version 0.1.

- promote the development of ensemble models towards integrated SASV solutions which operate upon speaker and anti-spoofing embeddings;
- encourage the development of single models which have the capacity to reject both utterances spoken by different speakers as well as spoofed utterances.

3 SASV solutions

SASV solutions can take the form of two different processing pipelines.

3.1 Ensemble solutions based upon separate ASV and CM systems

Ensemble SASV solutions are assumed to comprise pre-trained ASV and CM subsystems. Different ensemble techniques can be used to combine embeddings/scores produced by the ASV subsystem with embeddings/scores produced by the CM subsystem.

Potential solutions include, e.g.:

- score-sum ensembles using cosine similarity scores generated from speaker embeddings produced by a pre-trained ASV subsystem and the scores produced by a pre-trained CM subsystem;
- ensemble models which operate upon three different embeddings, namely a pair of speaker embeddings extracted from enrolment and test utterances and a CM embedding.

3.2 Integrated single system solutions

SASV solutions can also take the form of an integrated, single system.

Potential solutions include, e.g.:

- deep neural networks (DNNs) trained in multi-task fashion using a pair of output layers, namely one for speaker identification and another for spoofing detection;
- end-to-end systems with additional objective functions applied to intermediate, hidden layers.

4 Metrics

SASV performance will be assessed using the classical EER (SASV-EER) as the primary metric. Identical to the metrics used in [25,29], the SASV-EER does not distinguish between different speaker (zero-effort, non-target, or impostor) access attempts and spoofed access attempts. Additional insights into SASV performance can be gained from comparisons to more traditional estimates of speaker verification performance (SV-EER) estimated from a set of target and non-target trials, in addition to performance when the same system is subjected to spoofing attacks (SPF-EER) whereby non-target trials are replaced with spoofed trials. All three EER estimates reflect ASV performance, with both SV-EER and SPF-EER being estimated using different subsets of the full set of trials (i.e., protocol) used for estimating the SASV-EER. All SASV metrics are hence different to the EER metric used for ASVspoof challenges. The latter is estimated using a CM protocol, not an ASV protocol; the SPF-EER is measured when an SASV system processes pairs of enrolment and test utterances whereas the EER in the case of ASVspoof challenges is measured when a standalone CM system processes single utterances. Table 1 illustrates the ground-truth labels and trial subsets used to measure each of the three different EERs to be used for the SASV challenge.

Table 1: Description of EERs. The system involves enrolment utterance(s) and a test utterance. Enrolment utterance(s) is bona-fide (i.e. genuine) and test utterance belongs to either of the three types.

	Target	Non-target	Spoof
SV-EER	+	-	
SPF-EER	+		-
SASV-EER	+	-	-

5 Protocols

Participants will be provided with two protocols:

- Development protocol:
  `ASVspoof2019_LA_asv_protocol/ASVspoof2019.LA.asv.dev.gi.trl.txt`
- Evaluation protocol:
  `ASVspoof2019_LA_asv_protocol/ASVspoof2019.LA.asv.eval.gi.trl.txt`

Both protocols, which list target, non-target and spoofed trials, can be downloaded from the SASV 2022 GitHub repository at https://github.com/sasv-challenge/SASVC2022_Baseline or from ASVspoof challenge resources. The first is to be used for the development of SASV solutions. The second is to be used only for final performance evaluation. Both protocols are identical to those used for the ASVspoof 2019 LA challenge, albeit by the organisers for ASV experimentation instead of by participants for CM experimentation. This is hence the first time that the two protocols have been used by the participants of any common challenge. One thing worth noting about the protocols are that multiple enrolment utterances exist for each trial; this would be more familiar to researchers in the ASV community where speaker embeddings from each enrolment utterance is averaged to compose the final enrolment speaker embedding.

6 Baselines

We provide two baseline systems, one for each solution strategy described in Section 3.1. Each system is based upon the same pre-trained ASV and CM subsystems described further below. Reproducible software for each subsystem and baseline are also available from the SASV 2022 GitHub repository from which participants can download packages for:

- the extraction of speaker embeddings and CM embeddings using corresponding pre-trained subsystems;
- the estimation of SASV-EER, SV-EER, and SPF-EER metrics;
- Baseline1 and Baseline2 SASV solutions described in Section 6.4.

6.1 ASV subsystem

We adopt the ECAPA-TDNN [5] pre-trained ASV subsystem using the VoxCeleb2 dataset [30][1]. The system leverages several recent advances in deep learning, achieves state-of-the-art performance for the VoxCeleb1-O (VoxCeleb1 test set) protocol [31] and is widely adopted in the community. The architecture is based upon Res2Net with a squeeze-excitation module [32,33]. Participants are referred to [5] for full details.

[1] We use the implementation available at https://github.com/TaoRuijie/ECAPATDNN.

Results for the ECAPA-TDNN are illustrated in the first row of Table 2. The SV-EER of 0.83% demonstrates that non-target trials are reliably rejected without causing target trials to be rejected too, even though there is domain mismatch between the ASVspoof data and the VoxCeleb2 data used for ASV training. The SPF-EER of 29.3% confirms that a conventional ASV system is vulnerable to spoofing attacks, a result confirmed by the SASV-EER of 22.4%.

6.2 CM subsystem

The baseline CM subsystem is the AASIST model described in [20][2]. ASVspoof2019 LA train partition is used for training the system [7]. It is based upon an integrated spectro-temporal graph attention network which operates directly on raw-waveform inputs and is used for the extraction of CM embeddings. Participants are referred to [20] for further details.

Table 2: The three different EERs (%) for the SASV 2022 development and evaluation partitions. SASV-EER for all baselines are calculated using the entire protocol that includes trials used to measure the SV-EER (target vs. non-target) and those used to measure the SPF-EER (target vs. spoof). Results shown for a conventional ASV system (ECAPA-TDNN) and the two baseline solutions.

	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
ECAPA-TDNN	1.24	0.83	19.04	29.32	16.21	22.38
Baseline1 (score-sum)	14.89	35.1	6.94	0.50	2.09	19.15
Baseline2 (back-end ensemble model)	14.38	16.01	0.01	1.23	5.41	8.75

6.3 Baseline1: score-sum ensemble

Baseline1 involves a simple sum of the scores produced by the ASV and CM subsystems. Thus, no data is used for this baseline as it does not involve any training nor fine-tuning. Results are shown in the second row of Table 2. The Baseline1 SASV-EER of 19.2% corresponds to a relative reduction of 10% compared to the ECAPA-TDNN solution (22.4%). However, while Baseline1 improves performance when assessment includes spoofed trials, performance in the case of a typical ASV assessment scenario is poor; the SV-EER of the ECAPA-TDNN of 0.8% increases to 35.1%. This degradation is caused by summing of non-calibrated scores, each derived using different techniques (e.g., the cosine similarity for ASV scores but the DNN softmax output for CM scores). The back-end model ensemble strategy of Baseline2 is proposed as a potentially better solution.

6.4 Baseline2: back-end model ensemble

Baseline2 involves the fusion of three embeddings: one extracted from an ASV enrolment utterance using the ECAPA-TDNN system; a second extracted in identical fashion from a test utterance; a third extracted from the same test utterance using the AASIST spoofing CM. The model is a vanilla multi-layer perceptron with three hidden layers, trained using the ASVspoof2019 LA train partition.

[2] We used the implementation available at `https://github.com/clovaai/aasist` for CM embedding extraction.

7 Training datasets

Participants are permitted to use the following datasets:

- ASVspoof 2019 LA train partition [7];
- ASVspoof 2019 LA development partition [7];
- VoxCeleb 2 [30].

While participants can utilise the above data as they wish, it is stressed that ASVspoof 2019 LA train and development partitions were originally intended for the training and development of spoofing CMs. Since ground-truth speaker labels are available for the ASVspoof 2019 LA database, it can also be used for the training and development of ASV systems. The VoxCeleb 2 database was designed for ASV experimentation; it does not contain spoofed data. The use of data from the ASVspoof 2019 LA evaluation partition for training or development purposes is strictly prohibited.

The ASVspoof 2019 LA dataset can be downloaded at https://datashare.ed.ac.uk/handle/10283/3336. The VoxCeleb2 dataset can be downloaded at https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html. The use of any additional datasets is permitted, but only if they do not contain recording of speech (e.g., Musan [34] and the Room Impulse Response and Noise Database [35].

8 Registration process

Participants are required to register by completing and submitting the registration form available at:

- https://forms.gle/htoVnog34kvs3as56

9 System descriptions

Each participant/team is requested to submit a brief system description formatted according to the INTERSPEECH 2022 paper template together with their challenge submission. Instructions for challenge scores/results submission will follow in due course. Submitted system descriptions are expected to report the SASV-EER for both development and evaluation protocols as well as the corresponding SV-EER and SPF-EER. System descriptions will not be peer-reviewed and there is no page limit. After the challenge is finished, rankings accompanied by submitted system descriptions will be made publicly available at the challenge webpage:

- https://sasv-challenge.github.io

Participants may choose anonymous team names and also to anonymise their system description.

10 Paper submission

We aim to present SASV 2022 challenge results at a Special Session at INTERSPEECH 2022[3] to which participants are invited to submit their contributions.

[3] https://interspeech2022.org

11 Important Dates

- January 19, 2022: Release of evaluation plan
- March 10, 2022: Results submission
- March 14, 2022: Release of participant ranks
- March 21, 2022: INTERSPEECH Paper submission deadline
- March 28, 2022: INTERSPEECH Paper update deadline
- June 13, 2022: INTERSPEECH Author notification
- September 18-22, 2022: SASV challenge special session at INTERSPEECH

References

[1] William M Campbell, Douglas E Sturim, and Douglas A Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[3] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[5] B. Desplanques, J. Thienpondt, et al., "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech2020*, 2020.

[6] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," *arXiv preprint arXiv:2104.02370*, 2021.

[7] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A Lee, et al., "ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech," *Computer Speech & Language (CSL)*, vol. 64, 2020, 101114.

[8] Z. Wu, T. Kinnunen, et al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[9] M. Todisco, X. Wang, et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. INTERSPEECH*, 2019, pp. 1008–1012.

[10] J. Yamagishi, X. Wang, et al., "ASVspoof2021: accelerating progress in spoofed and deep fake speech detection," in *Proc. ASVspoof 2021 Workshop*, 2021.

[11] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *Proc. INTERSPEECH*, 2019.

[12] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Speaker Odyssey Workshop*, 2020, pp. 132–137.

[13] Hemlata Tak, Jeeweon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. ASVspoof workshop*, 2021.

[14] G. Hua, A. Beng jin teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, 2021.

[15] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," in *Proc. INTERSPEECH*, 2021.

[16] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. INTERSPEECH*, 2021.

[17] Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. Interspeech*, 2021.

[18] A. Luo, E. Li, Y. Liu, X. Kang, and Z J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc.ICASSP*, 2021.

[19] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "Raw differentiable architecture search for speech deepfake and spoofing detection," in *Proc. ASVspoof workshop*, 2021.

[20] J.-w. Jung, H.-S. Heo, et al., "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," *arXiv preprint arXiv:2110.01200*, 2021.

[21] Xin Wang and Junichi Yamagishi, "A practical guide to logical access voice presentation attack detection," *arXiv preprint arXiv:2201.03321*, 2022.

[22] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, vol. 3, 2021.

[23] T. Kinnunen, H. Delgado, et al., "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio Speech and Language Processing (TASLP)*, vol. 28, 2020.

[24] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[25] M. Todisco, H. Delgado, et al., "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," *Proc. Interspeech 2018*, pp. 77–81, 2018.

[26] Jiakang Li, Meng Sun, Xiongwei Zhang, and Yimin Wang, "Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss," *IEEE Access*, vol. 8, pp. 7907–7915, 2020.

[27] Hye-jin Shim, Jee-weon Jung, Ju-ho Kim, and Ha-jin Yu, "Integrated replay spoofing-aware text-independent speaker verification," *Applied Sciences*, vol. 10, no. 18, pp. 6292, 2020.

[28] A. Gomez-Alanis, J. A Gonzalez-Lopez, et al., "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[29] M. Sahidullah, H. Delgado, et al., "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on asvspoof 2015," *Proc. INTERSPEECH*, pp. 1700–1704, 2016.

[30] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[31] A. Nagrani, J. S Chung, et al., "Voxceleb: a large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017.

[32] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[33] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[34] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[35] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.