Baseline Systems for the First Spoofing-Aware Speaker Verification Challenge: Score and Embedding Fusion

Hye-jin Shim[*,1], *Hemlata Tak*[*,2], *Xuechen Liu*[3,8], *Hee-Soo Heo*[4], *Jee-weon Jung*[4], *Joon Son Chung*[5], *Soo-Whan Chung*[4], *Ha-Jin Yu*[1], *Bong-Jin Lee*[4], *Massimiliano Todisco*[2], *Héctor Delgado*[6], *Kong Aik Lee*[7], *Md Sahidullah*[8], *Tomi Kinnunen*[3], *Nicholas Evans*[2]

[1]School of Computer Science, University of Seoul, South Korea
[2]EURECOM, Sophia Antipolis, France, [3]University of Eastern Finland, Finland
[4]Naver Corporation, South Korea, [5]KAIST, South Korea
[6]Nuance Communications, Spain, [7]A*STAR, Singapore, [8]Inria, France

sasv.challenge@gmail.com

Abstract

Deep learning has brought impressive progress in the study of both automatic speaker verification (ASV) and spoofing countermeasures (CM). Although solutions are mutually dependent, they have typically evolved as standalone sub-systems whereby CM solutions are usually designed for a fixed ASV system. The work reported in this paper aims to gauge the improvements in reliability that can be gained from their closer integration. Results derived using the popular ASVspoof2019 dataset indicate that the equal error rate (EER) of a state-of-the-art ASV system degrades from 1.63% to 23.83% when the evaluation protocol is extended with spoofed trials. However, even the straightforward integration of ASV and CM systems in the form of score-sum and deep neural network-based fusion strategies reduce the EER to 1.71% and 6.37%, respectively. The new Spoofing-Aware Speaker Verification (SASV) challenge has been formed to encourage greater attention to the integration of ASV and CM systems as well as to provide a means to benchmark different solutions.

Keywords: automatic speaker verification, anti-spoofing, spoofing-aware speaker verification, spoofing countermeasures.

1. Introduction

Recent years have seen rapid progress in automatic speaker verification (ASV) [1–3]. Even for unconstrained *in the wild* scenarios, the latest systems deliver low equal error rates (EERs) that are close to those for well-constrained conditions [2, 4, 5]. However, there is evidence that these improvements might not offer protection against *spoofing attacks* – the presentation of utterances specially crafted to deceive the ASV system.

Solutions to protect ASV systems from such attacks take the form of countermeasures (CMs), typically separate sub-systems designed to detect manipulated or synthetic utterances [6]. The threat of spoofing attacks has intensified in recent times due to the rapid advances in other speech technologies which can be used to generate spoofed utterances. They include: speech-to-speech voice conversion (VC); text-to-speech (TTS) speech synthesis; replay attacks. Since ASV systems are increasingly deployed in security-critical operations as a part of a biometric authentication system, vulnerabilities to spoofing attacks are unacceptable.

In response to the threat, the ASVspoof initiative has held biennial challenges to promote the development of research in spoofing detection [6]. Two different use case scenarios have been defined, namely physical access (PA) and logical access (LA). The work in this paper relates to the latter, typically telephony applications and robustness to TTS and VC spoofing attacks. When assessed using the ASVspoof 2019 LA evaluation set, today's leading CM systems deliver EERs of less than 2% [7–17].

While the EER metric was adopted in almost all early assessments of standalone CM performance, the ASVspoof community has now transitioned to the minimum tandem detection cost function (min t-DCF) [18] as the primary metric. The t-DCF reflects the impact of spoofing and countermeasures upon a typically-fixed ASV system. Even with this strategy, CMs are often designed in standalone fashion, independently from ASV. Until now, and with only few notable exceptions [19–26], very little work has investigated the benefit of jointly optimised, or integrated CM+ASV solutions.

The **Spoofing-Aware Automatic Speaker Verification** (SASV) challenge[1], a special session at INTERSPEECH 2022, aims to promote greater research in this direction and extends the traditional ASV scenario to

[*] These authors contributed equally to this work.

[1] https://sasv-challenge.github.io

10.21437/Odyssey.2022-46

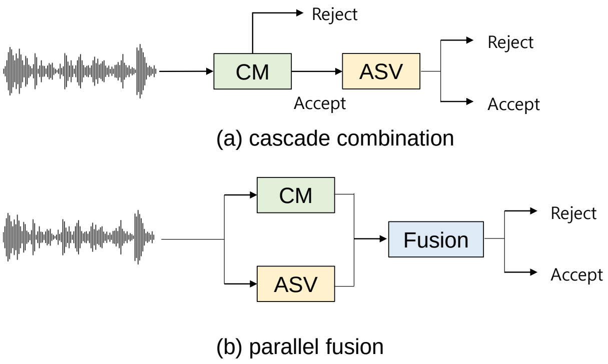

Figure 1: Back-end fusion of CM and ASV sub-systems. (a): cascaded combination, a form of decision level fusion. (b): parallel fusion which can operate at either decision, score, or embedding levels. When parallel fusion operates at the decision level, it is the same as the cascaded combination.

consider spoofing attacks. The first SASV challenge [27] utilises existing ASVspoof databases with metrics extended to support the evaluation of integrated CM+ASV solutions. Ultimately, SASV aims to strengthen the foundations between research in spoofing detection and ASV.

New contributions reported in this paper include: (i) baseline SASV solutions to integrated CM+ASV leveraging state-of-the-art sub-systems; (ii) metrics designed specifically for the SASV task; (iii) experimental results and detailed, per-attack analyses.

2. Related work

The majority of previous related work focuses upon the development of independent CM and ASV solutions. Comparatively very little work has explored their integration. The literature can be divided into two strands: (i) back-end fusion using independent ASV and CM subsystems and (ii) single SASV systems.

The earlier works in the first strand investigated decision-level cascade or parallel combinations [19, 20, 28]. As shown in Fig. 1-(a), the cascaded combination typically involves the use of a CM as a gate prior to ASV so that the latter treats only input utterances labelled by the CM as bona fide. It can be regarded as a form of decision level fusion. For parallel combinations, as illustrated in Fig. 1-(b), every input utterance is treated by both sub-systems before the outputs are fused. Fusion can be performed at decision, score or embedding levels. Given decision level fusion and identical CM and ASV thresholds, cascade and parallel solutions give identical results.

Beyond straightforward score fusion, [20] reports a Gaussian back-end fusion strategy with different front-ends for CM and ASV sub-systems. The Gaussian back-end fusion method is used to model ASV and CM scores as two-dimensional vectors from which single scores are derived. The Gaussian back-end is shown to outperform both cascade and straightforward score fusion parallel combination strategies by a large margin.

The second strand of single SASV systems has been also explored [21, 22, 24, 25]. An approach to the joint training of CM and ASV systems using multi-task learning (MTL) [29] is reported in [21, 22]. However, the framework requires DNN training towards the speakers in the enrolment set and cannot incorporate new speakers, making the framework somewhat inflexible. The first single SASV system adaptable towards unlimited speakers is reported in [24]. However, the single SASV system is outperformed by a back-end DNN fusion approach (the former strand) by a large margin.

While some of the above referenced works were performed with standard databases, none used common protocols to assess ensemble or integrated CM+ASV approaches. The absence of benchmarking frameworks means that results for different approaches cannot be compared meaningfully and hinders the development of integrated systems. The SASV challenge has been designed to address these issues, to establish such a common benchmarking framework and to promote progress in the integration and joint optimisation of CM+ASV solutions. The remainder of this paper introduces our baseline systems, protocol, metrics, and results.

3. Embedding extraction

Our DNN-based baseline system, depicted in Fig. 2 and introduced in Section 4, is inspired by the solution in [24] and is based upon an ensemble of three embeddings. The first and second are speaker (ASV) embeddings extracted from enrolment and test utterances respectively. The third is a spoofing (CM) embedding extracted only from the test utterance. Described in this section are the backbone models used for their embedding extraction.

3.1. Speaker embedding

Driven by the availability of massive datasets, e.g. VoxCeleb [30], and competitive challenges, the performance of ASV systems has improved substantially in recent years [1, 2]. The majority of today's best-performing ASV systems utilise some form of speaker embedding in a latent space in which linear classifiers can be applied (e.g., cosine similarity, probabilistic LDA). We use the ECAPA-TDNN[2] speaker embedding extractor, one of the most popular models in the recent ASV literature [2]. It consists of a Res2net backbone architecture [31] with squeeze-excitation (SE) modules [32]. The model oper-

[2] https://github.com/TaoRuijie/ECAPATDNN

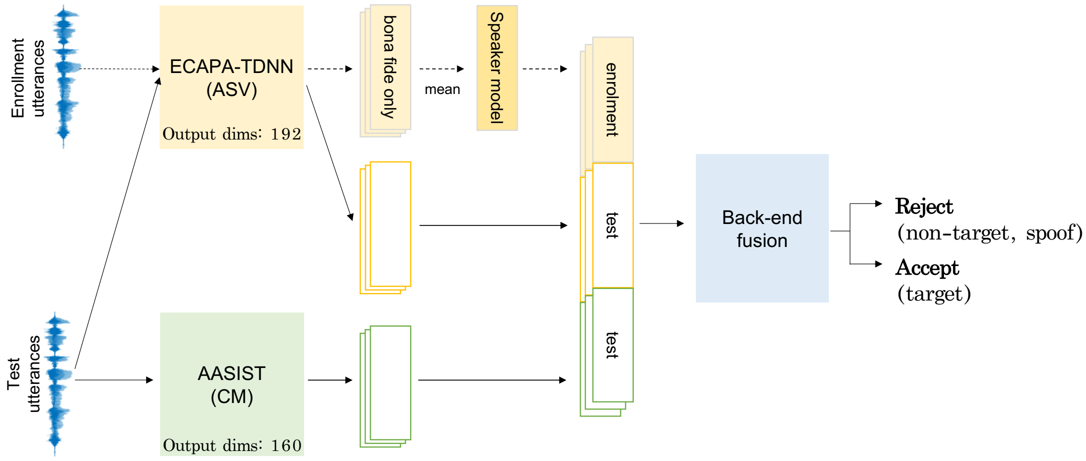

Figure 2: Illustration of the back-end DNN fusion. Three embeddings are fed to the DNN; only a speaker embedding is extracted from enrolment utterance and both speaker and spoofing embeddings are extracted from test utterance. In the training phase, 'mean' is removed because only one enrolment utterance is involved. In both the development and evaluation phases, there exist multiple enrolment utterances, hence, embeddings are averaged element-wise before being fed to the DNN.

ates upon cepstral acoustic features and uses three SE-Res2net blocks where three-block outputs are all concatenated. Concatenated frame-level embeddings are then aggregated into a single utterance-level embedding leveraging an attentive statistical pooling (ASP) layer. The ASP layer is a variant of the original reported in [33] which is also dependent on channel and context statistics among frame-level embeddings. 192-dimensional speaker embeddings are obtained by applying an affine transform with a fully-connected layer to the ASP layer output. The model is trained using an additive angular margin softmax (AAM-softmax) objective function [34]. Further details are available in [2].

3.2. Spoofing embedding

Progress in spoofing detection has been led by the ASVspoof initiative[3] and associated challenge series [35] which provide benchmarking using common datasets, protocols, and metrics. The state-of-the-art methods apply end-to-end (E2E) DNNs with diverse architectures and strategies such as graph neural networks and graph attention networks (GATs) [9, 16].

We use the E2E AASIST[4] spoofing detection model which delivers state-of-the-art performance for the ASVspoof 2019 LA database [16]. It operates directly upon raw waveform inputs using a variant of the RawNet2 encoder [36] to generate three-dimensional feature maps (channel, spectral, and temporal). Two different views, (channel, spectral) and (channel, temporal), are then composed by applying an element-wise maximum operation to spectral and temporal axes. Those two views are further processed using graph modules that consist of GATs and a graph pooling layer. Specially designed heterogeneous graph attention layers and max graph operations are then used to integrate the two views thereby combining temporal and spectral cues. A two-class prediction output is finally generated using a readout operation comprising a hidden fully connected (FC) layer. 160 dimensional spoofing embeddings are extracted immediately prior to the FC output layer. Further details are available in [16].

4. SASV

Described in this section are three different approaches to combine CM and ASV sub-systems. They involve: i) score-sum fusion; ii) non-linear embedding fusion using DNNs; and iii) cascaded combination. Each of the three strategies is described in the following.

4.1. B1: Score-sum fusion

The simplest strategy to system combination involves a straightforward score-sum. score-sum fusion is used widely and needs neither training nor fine-tuning. Our approach is based upon the ASV cosine similarity score (derived from enrolment and test utterances) and the CM output score (derived from the test utterance). ASV scores are calculated using cosine similarity with the range of -1 to 1. CM scores are the softmax non-linearity outputs with the range of 0 to 1. The score-sum back-end fusion serves as baseline1 (B1) for the SASV 2022 Challenge.[5]

[3] https://www.asvspoof.org/

[4] https://github.com/clovaai/aasist

[5] Performance is different to that stated in the evaluation plan [27] on account of the softmax applied to the CM output which improves

Table 1: Description of EERs. The system involves enrolment utterance(s) and a test utterance. The 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	+	-	-

4.2. B2: DNN back-end fusion

The DNN-based fusion strategy is illustrated in Fig. 2. It operates upon the set of three embeddings described in Section 3: a pair of speaker (ASV) embeddings extracted from enrolment and test utterances; a spoofing (CM) embedding extracted from the test utterance alone. As illustrated in Fig. 2, ASV embeddings and the CM embedding are combined using back-end DNN fusion. The DNN back-end fusion serves as baseline2 (B2) for the SASV 2022 Challenge. The DNN model contains three fully connected layers with leaky ReLU non-linear activation functions. The last layer consists of two neurons which correspond to two classes: (i) target, (ii) non-target / spoof. Element-wise average speaker embeddings are used in the case of multiple enrolment utterances.

4.3. Cascaded combination

Though not an SASV baseline, a cascaded combination is also reported here for reference and is the same approach as illustrated in Fig. 1-(a). In practice, we utilise separate CM and ASV thresholds set using development data to make false acceptance rates and false rejection rates equal for both sub-systems. Separate thresholds, optimised for the CM and ASV systems using the development protocol, are then applied without modification to evaluation data. The gate applied in the cascaded approach effectively combines CM and ASV systems at the decision level. While the CM produces a score, the ASV system produces scores only for trials labelled by the CM as bona fide (the gated decision). Only decisions, not scores, are made consistently for each trial and, as a result, there is no straightforward way to estimate the EER. Thus, to estimate performance in the case of cascaded combination, we resort to error counting and estimation of the half-total error rate (HTER). Given that we use the EER for the two baseline systems, but the HTER for the cascaded combination, we stress that results are not comparable.

5. Databases and protocols

We describe the databases used in this work: (i) the VoxCeleb2 database [30] used for training the ASV system; (ii) the ASVspoof 2019 LA database [37] used for training the CM system and for SASV assessment.

5.1. VoxCeleb2

The ECAPA-TDNN model used to extract speaker embeddings was trained using the development partition of the VoxCeleb2 database. The dataset was collected by crawling online videos of celebrities' interviews. Its development partition includes data collected from 5,994 speakers of which 61% are male and 39% are female. Network inputs are 80-dimensional mel filterbank acoustic features. The network was trained following the recipe described in [38], where data augmentation is based on use of the room impulse response (RIR) database [39] and additive noise recordings contained in the MUSAN database [40].

5.2. ASVspoof 2019

CM experiments were performed following the standard ASVspoof 2019 LA CM protocol described in [37]. It consists of disjoint train, development, and evaluation partitions. Each partition contains both bona fide and spoofed utterances where the latter are generated using 19 VC and TTS algorithms (6 for the train and development sets, 13 for the evaluation set). SASV evaluation is performed using the ASVspoof 2019 LA ASV protocol. The ASV protocol is not used by ASVspoof participants and is, instead, used only by the ASVspoof organisers to estimate ASV performance and tandem CM+ASV performance using the min t-DCF metric. For the SASV challenge, the ASV protocol is used by participants for experimentation involving three different trials:

1. **target** bona fide trials uttered by the same speaker as the enrolment utterance(s);
2. **non-target** (zero-effort impostor) bona fide trials uttered by a different speaker as the enrolment utterance(s);
3. **spoofed** trials which are synthesised or converted to spoof the voice of the speaker in the enrolment utterance(s).

Both development and evaluation protocols are provided with the freely available ASVspoof 2019 LA dataset[6] or with the open-source SASV baseline implementations.[7]

- development protocol: `ASVspoof2019.LA.asv.dev.gi.trl.txt`;
- evaluation protocol: `ASVspoof2019.LA.asv.eval.gi.trl.txt`.

results notably.

[6] https://datashare.ed.ac.uk/handle/10283/3336

[7] https://github.com/sasv-challenge/SASVC2022_Baseline

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 are shown for a conventional ASV system (ECAPA-TDNN) and the two baseline solutions. B1 and B2 are baseline systems used for the SASV challenge

	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
ECAPA-TDNN	1.88	1.63	20.30	30.75	17.38	23.83
B1[5]: Score-sum	1.99	1.66	0.23	1.76	1.01	1.71
B2: DNN fusion	12.87	11.48	0.13	0.78	4.85	6.37

Table 3: HTERs (%) of the cascade solution for the SASV 2022 development and evaluation partitions. We use the thresholds that give equal false acceptance and false rejection rates for each of the ASV and CM systems (i.e., the ones that are used to measure EER).

	SV-HTER		SPF-HTER		SASV-HTER	
	Dev	Eval	Dev	Eval	Dev	Eval
Cascade	1.90	1.60	0.99	1.40	1.18	1.47

6. Metrics

We use the classical EER (SASV-EER) as the primary metric. In keeping with the metrics used in [19, 20], the SASV-EER does not distinguish between non-target/zero-effort impostor and spoofed access attempts. Additional insights into performance can be gained from comparisons between the SASV-EER and: (i) more traditional estimates of speaker verification performance (SV-EER) estimated from a set of target and non-target bona fide trials; (ii) estimates of performance when non-target trials are replaced with spoofed trials (SPF-EER).

Table 1 illustrates the trials types and ground-truth labels used to measure each of the three different EERs. As shown, all three EERs are estimates of ASV performance, with both the SV-EER and SPF-EER being estimated using different subsets of the full set of trials that are used for estimating the SASV-EER.

7. Experiments

We describe specific implementation details relating to embedding extraction, DNN-based fusion and hardware, followed by the presentation of experimental results.

7.1. Implementation details

Speaker and spoofing embeddings – We used open source implementations for both the ECAPA-TDNN[2] and AASIST[4] models described in Sections 3.1 and 3.2, respectively. We used pre-trained weight parameters making embedding extraction fully reproducible.

DNN-based back-end fusion, B2: We used a simple multi-layer perceptron for the back-end fusion with three hidden layers comprising 256, 128 and 64 nodes respectively, without regularisation (e.g., dropout, batch normalisation, and weight decay). The DNN model for back-end fusion is trained on the ASVspoof 2019 LA train partition. The Adam optimiser was applied, and the learning rate was scheduled with warm starts between 0.1 to 0.001 [41]. The output node indicates whether the utterance is a target or not (non-target or spoof). The fusion model is also publicly available.[7]

Hardware specification – All experiments reported in this paper were preformed using a single Nvidia 3090 GPU. The provided scripts can also be run on GPUs with less memory, e.g. an Nvidia 1080ti GPU.

7.2. Results

Results are presented in Table 2 which shows all three EERs for both development and evaluation partitions for the ECAPA-TDNN system and the two SASV baselines. Results for the cascaded ensemble system in terms of HTERs are shown in Table 3.

ECAPA-TDNN – Results for the ECAPA-TDNN are illustrated in the first row of Table 2. In view of the domain mismatch between ASVspoof data and the VoxCeleb2 data used for ASV training, the system performs reliably, with SV-EERs of 1.88% and 1.63% for the development and evaluation protocols respectively. The SPF-EERs of 20.30% and 30.75% indicate that the system is vulnerable to spoofing attacks. The SASV-EERs, at 17.38% and 23.83%, are also high, confirming that the standalone ASV system provides little robustness to spoofing attacks.

Score-sum fusion, B1 – The second row of Table 2 shows results using score-sum fusion described in Section 4.1. SV-EER results are inline with results for the ECAPA-TDNN system. SPF-EER results, however, are substantially reduced. Together, these results indicate the potential of the B1 baseline to improve robustness to both

Table 4: Breakdown of SPF-EER (%) and their pooled (P) EER for all 13 different spoofing attacks in the ASVspoof 2019 LA evaluation set, measured using SASV protocol without non-target trials.

System	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	P
ECAPA-TDNN	32.66	18.80	2.20	50.61	47.08	39.56	11.62	35.39	36.54	60.71	1.85	2.38	4.77	30.75
B1[5]:Score-sum	2.05	0.69	0.07	7.19	0.32	4.42	0.07	0.08	1.75	1.18	0.73	1.18	0.63	1.76
B2:DNN fusion	0.50	0.34	0.00	1.28	0.20	1.20	0.16	0.10	0.55	0.85	0.77	1.87	0.42	0.78

non-target trials and spoofing attacks, without impacts upon target trials. This finding is confirmed by SASV-EERs of 1.01% and 1.71% for the development and evaluation protocols. The SASV-EER of 1.71% remains 5% higher relative to the ECAPA-TDNN SV-EER of 1.63% implying that, while results are encouraging, impacts of spoofing remain.

DNN fusion, B2 – The last row of Table 2 shows results for DNN-based back-end embedding-level fusion. With speaker and spoofing embeddings projected to a new representation space via an affine transformation, we expected DNN-based embedding-level fusion to outperform score-level fusion. Relative to B1, the SPF-EER of 1.76% is reduced by more than 50% for the evaluation protocol. However, the SV-EER increases from 1.66% for B1 to 11.48%, meaning that, with a vanilla multi-layer perception, the discriminative power of the speaker embedding is degraded in the joint representation space. This explains the increase in the SASV-EER, to 6.37%.

Cascade ensemble – Table 3 shows results in terms of the HTER for cascaded CM and ASV systems as described in Section 4.3. Lower error rates below the SV-HTER are observed for both the SPF-HTER and the SASV-HTER which drops to 1.47%. While error rates for the cascaded ensemble are lower than those for the two SASV baseline systems, results for the cascaded system are HTER estimates, not EER estimates. Therefore, they are not comparable and should not be interpreted as meaning that the cascaded ensemble is the best approach. They might be interpreted, instead, as scope to improve SASV-EER results for the B1 and B2 baselines.

Breakdown of SPF-EER – Table 4 shows a breakdown of SPF-EER results for each of the 13 different spoofing attacks, with pooled results to the right. B2 outperforms B1 for the majority of attacks. B1 outperforms B2 for A13, A14, A17, and A18 attacks, but the difference is almost negligible. For B2, the SPF-EER is under 1% for all attacks except three. For A10 and A12, B1 SPF-EERs are 461% and 268% higher relative to B2 results. While B1 gives a lower SV-EER, B2 better harnesses the protection of the CM in deflecting spoofing attacks which hence leads to a lower SPF-EER. It is the relative weakness in terms of the SV-EER that results in B2 having a higher SASV-EER.

Further analysis. The trends shown in Table 2 are unexpected; we expected better results for the DNN-based back-end. Using current ASV and CM sub-systems, simple fusions that operate at embedding or score levels give the best performance. However, we remain convinced in the merit of DNN-based embedding-level fusion approaches and predict that further research can reduce error rates considerably. B2 has better potential to harness the *synergy* between CM and ASV sub-systems; this is not the case for B1. With advanced architectures, regularisation, and training strategies that better exploit the synergy, ASV and CM embedding fusion should give better performance and might even deliver SASV-EERs below the SV-EER. We expect to observe such improvements in the forthcoming SASV 2022 Challenge. We argue that, with better potential for joint optimisation and hence better performance, future work should focus on the development of *single, integrated* SASV solutions. Their development is the ultimate goal of the SASV challenge.

8. Conclusions

We found that, despite rapid advances in ASV, the state-of-the-art ECAPA-TDNN system remains vulnerable to spoofing attacks, hence, motivating either (i) its combination with a standalone CM system or (ii) the development of integrated spoofing aware ASV (SASV) solutions. We explored different back-end integration (fusion) techniques, a straightforward score-sum fusion and a more sophisticated DNN-based approach. Surprisingly, the simple score-sum ensemble outperforms the DNN-based approach. This result may imply that simple back-end fusions which operate upon ASV and CM scores may be a sufficient solution. Nonetheless, DNN-based back-end solutions and single integrated approaches have greater potential for joint-optimisation to better exploit the synergy between CM and ASV solutions. We hope to encourage this work through the SASV 2022 Challenge. In the end, there is only a *single*, common task - reliable ASV – a task that might best be solved with a *single* SASV system.

9. Acknowledgements

The contributions of EURECOM authors is supported by the ExTENSoR project funded by the French Agence Nationale de la Recherche (ANR).

10. References

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018.

[2] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. INTERSPEECH*, 2020.

[3] Z. Bai and X.-L Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, 2021.

[4] Daniel Garcia-Romero, Greg Sell, and Alan Mccree, "Magneto: X-vector magnitude estimation network plus offset for improved speaker recognition," in *Proc. Speaker Odyssey Workshop*, 2020.

[5] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification," in *Proc. INTERSPEECH*, 2021.

[6] A. Nautsch, X. Wang, et al., "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.

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

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

[9] Tak, Hemlata and Jung, Jee-weon and Patino, Jose and Todisco, Massimiliano and Kamble, Madhu and Massimiliano and Evans, Nicholas, "End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection," in *Proc. ASVspoof 2021 Workshop 2021*, 2021.

[10] G. Hua, A. Beng jin teoh, and H. Zhang, "Towards End-to-End Synthetic Speech Detection," *IEEE Signal Processing Letters*, 2021.

[11] 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.

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

[13] Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang, "The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System," in *Proc. INTERSPEECH*, 2021.

[14] 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.

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

[16] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," in *Proc. ICASSP (to appear)*, 2022.

[17] Xin Wang and Junichi Yamagishi, "A Practical Guide to Logical Access Voice Presentation Attack Detection," *arXiv preprint arXiv:2201.03321*, 2022.

[18] 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.

[19] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Hong Yu, Tomi Kinnunen, Nicholas Evans, and Zheng-Hua Tan, "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015," in *Proc. INTERSPEECH*, 2016.

[20] Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion," in *Proc. INTERSPEECH*, 2018.

[21] Jiakang Li, Meng Sun, and Xiongwei Zhang, "Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection," in *APSIPA*, 2019.

[22] 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*, 2020.

[23] 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, 2015.

[24] Hye-jin Shim, Jee-weon Jung, Ju-ho Kim, and Hajin Yu, "Integrated replay spoofing-aware text-independent speaker verification," *Applied Sciences*, 2020.

[25] 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, 2021.

[26] A. Kanervisto, V. Hautamäki, and others Kinnunen, "Optimizing Tandem Speaker Verification and Anti-Spoofing Systems," *IEEE/ACM TASLP*, vol. 30, pp. 477–488, 2021.

[27] 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, "SASV Challenge 2022: A Spoofing Aware Speaker Verification Challenge Evaluation Plan," *arXiv preprint arXiv:2201.10283*, 2022.

[28] Elie Khoury, Tomi Kinnunen, Aleksandr Sizov, Zhizheng Wu, and Sébastien Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Proc. INTERSPEECH*, 2014.

[29] Rich Caruana, "Multitask learning," *Machine learning*, 1997.

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

[31] 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.

[32] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018.

[33] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. INTERSPEECH*, 2018.

[34] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019.

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

[36] Jung, Jee-weon and Kim, Seung-bin and Shim, Hye-jin and Kim, Ju-ho and Yu, Ha-Jin, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms," in *Proc. INTERSPEECH*, 2020.

[37] X. Wang, J. Yamagishi, et al., "ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech," *Computer Speech & Language (CSL)*, vol. 64, 2020, 101114.

[38] Rohan Kumar Das, Ruijie Tao, and Haizhou Li, "HLT-NUS SUBMISSION FOR 2020 NIST Conversational Telephone Speech SRE," *arXiv preprint arXiv:2111.06671*, 2021.

[39] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017.

[40] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[41] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2017.