

SA-SASV: An End-to-End Spoof-Aggregated Spoofing-Aware Speaker Verification System

Zhongwei Teng¹, Quchen Fu¹, Jules White¹, Maria E. Powell², Douglas C. Schmidt¹

¹Dept. of Computer Science, Vanderbilt University

²Dept. of Otolaryngology–Head and Neck Surgery, Vanderbilt University Medical Center

Abstract

Research in the past several years has boosted the performance of automatic speaker verification systems and countermeasure systems to deliver low Equal Error Rates (EERs) on each system. However, research on joint optimization of both systems is still limited. The SASV 2022 challenge was proposed to encourage the development of integrated spoofing aware speaker verification system (SASV) with new metrics to evaluate joint model performance. This paper proposes an ensemble-free end-to-end solution, SA-SASV, to build a SASV system with multi-task classifiers, which are optimized by multiple losses. The proposed system is evaluated on the ASVSpoof 2019 evaluation dataset and improves the performance of baseline systems from 8.76% to 4.48% in SASV-EER.

Index Terms: spoofing aware speaker verification, spoof detection

1. Introduction

Automatic speaker verification (ASV) systems have shown the ability to provide biometric authentication of users for applications that require robust reliability in changing acoustic environments, including resistance to malicious attacks [1, 2, 3, 4, 5]. However, current ASV systems are still vulnerable to spoofing attacks, such as text-to-speech(TTS) [6, 7, 8] and voice conversion(VC) [9], and can be deceived and manipulated by malicious entities using generated speech. To overcome bottlenecks in spoofing and countermeasure research for ASVs, a series of ASVSpoof challenges have been proposed since 2015 to help encourage the development of robust countermeasure (CM) systems [10, 11, 12, 13], which can complement ASV systems with an anti-spoof model. The anti-spoof gate by provides a "spoof confidence" score to help filter out spoofing attacks. Metrics on the ASVSpoof challenge are based on the minimum tandem detection cost function (tDCF) [14], which can evaluate the performance of CM systems on fixed ASV systems with pre-determined output scores. Rather than developing CM and ASV systems independently, a neglected research question is if we can develop an integrated system where CM and ASV system can be optimized together, so that a single verification score is able to determine whether an input speech sample is a target speaker, while also accounting for potential spoofing attacks.

To encourage research on integrated Spoofing-Aware Speaker Verification (SASV) systems, the SASV Challenge 2022 [15] was proposed using the ASVSpoof 2019 Logical Access Dataset with new metrics, SASV-EER. In the challenge, a single score determines if the input speech sample is the target speaker. Non-target inputs include both zero-effort and spoofed impostors. The SASV challenge provides two baseline systems by applying different fusion strategies (a. score-level fusion b. embedding-level fusion) to pre-trained ASV systems and CM

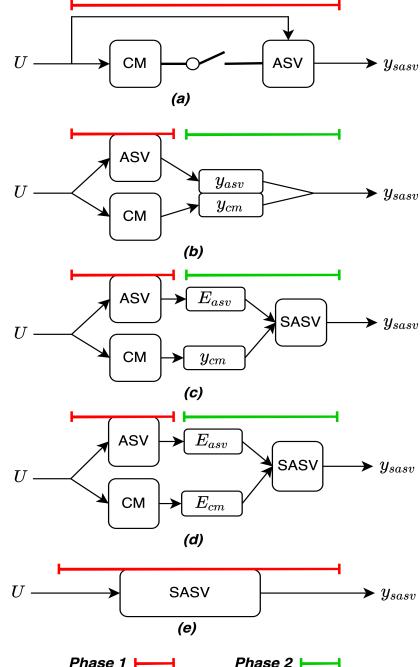


Figure 1: Feasible solutions to build integrated SASV systems. Red/green lines indicate separate training stages. (a) Cascaded ASV/CM systems. (b) Fusions of scoring prediction. (c) Fusions of scoring and feature embedding. (d) Fusions of feature embedding. (e) End-to-End SASV systems.

systems. Figure 1 shows potential solutions to the SASV problem.

In this paper, rather than fusion strategies, we propose a fully trainable end-to-end SASV system, called Spoof-Aggregated Spoofing-Aware Speaker Verification System (SA-SASV). The proposed approach combines a pre-trained ASV system with a lightweight raw waveform encoder to form the overall encoder [16]. This paper builds on our experience that encoding can be one of the most important aspects of these types of anomaly detection problems [16, 17, 18]. Multiple classifiers and spoof-source-based triplet loss are employed to further enhance model performance in generating the shared SASV feature space.

The rest of the paper is organized as follows: Section 2 introduces related research on SASV systems; Section 3 discusses the model architecture of our Spoof-Aggregated Spoofing-Aware Speaker Verification Systems; Experiment results are discussed in Section 4; and Section 5 presents concluding re-

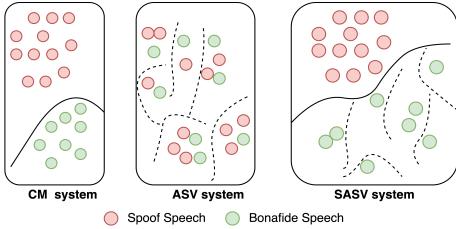


Figure 2: *Desired speech sample classification distribution of different spoof detection systems. The SASV system integrates the CM and ASV systems so that there are two types of boundaries to separate spoof/bonafide speech and target/non-target speakers.*

marks.

2. Related Work

The SASV system aims to build a single system to reject utterances from zero-effort and spoofed speech. Previous work focused on two solutions to this problem: ensemble SASV solutions and integrated single system solutions.

Ensemble SASV solutions take fixed outputs from pre-trained ASV and CM systems and apply varying fusion strategies to generate a single SASV score for both tasks. Sizov et al. [19] was the first to apply i-vectors and a PLDA back-end for joint modeling of speaker verification and spoof detection. At the score level, Todisco et al. [20] proposed a two-dimensional score modeling method to get a single score threshold for both ASV and CM systems. Shim et al. [21] discusses a back-End modular approach to train embeddings from pre-trained fixed ASV systems and spoofing predictions from CM systems to predict final SASV scores. In addition to scoring ensembles, fusions based on embeddings from different models have also been tested. For example, Gomez-Alanis et al. [22] proposed DNN-based integration methods to train three types of embeddings from ASV and CM systems jointly.

The target task of an integrated single SASV system is to build an end-to-end system that simultaneously classifies speech based on whether or not it is from the target speaker and is authentic non-spoofed speech. Zhao et al. [23] built an SR-ASV system with two classifiers to get CM scores and ASV scores from shared layers and the final decision is based on both the CM and ASV scores. Li et al. [24] applied speaker-based triplet loss to train multi-task classification networks to make a joint decision on anti-spoofing and ASV.

As a form of integrated single SASV system, our method explores the shared feature space of ASV and CM tasks. To obtain proper embeddings for speech from the multiple encoders that we employ, both hand-crafted features and raw waveforms are input into SA-SASV. We first discuss the feasibility of optimizing the SASV feature space by aggregating spoofed voice samples based on their spoofing sources. The proposed model was trained with multiple loss functions, including spoof source-based triplet loss. The final decision by our model is based on cosine similarity and CM scores of single model.

3. AS-DGASV Model Architecture

As shown in Figure 2, compared to independent CM models and ASV models, the ideal feature space learned from SASV

models should have the following characteristics: 1) Spoofed and bonafide speech should be densely aggregated, so that obvious margins can be drawn to separate them. 2) In the clusters of bonafide speech, sources from different speakers should be sparsely distributed so that models can distinguish between different speakers.

To achieve this type of optimized feature space in a SASV system, we proposed the SA-SASV model, whose decoder consists of three parts: multi-task classifiers, spoof aggregators, and spoof-source-based triplet loss, as shown in Figure 3. This fully trainable model takes both raw waveforms and hand-crafted features as input and multiple losses are used to optimize feature embedding.

3.1. The ARawNet Encoder

Previous research shows that the best-performing ASV systems [5] and CM systems [25], take hand-crafted features and raw waveforms, respectively, indicating distinctive features among each type of input that are useful for identifying speakers and spoofing attacks. However, it is difficult to simply merge existing state-of-the-art ASV and CM systems together to develop an end-to-end model, due to the resulting large model size and high computational complexity. We use our previously published ARawNet architecture [16] to help overcome this limitation. Our encoder combines a pre-trained ASV system (ECAPA-TDNN) and a lightweight raw waveform encoder to allow for simultaneous analysis of both learned features and raw wave forms.

We denote input utterance as U . An utterance's embedding, E_u , can be described as shown in Equation 1, where F_{asv} is a pre-trained ECAPA-TDNN, F_{raw} is an un-trained auxiliary raw encoder, and F_c is a concatenating encoder that handles outputs from F_{asv} and F_{raw} .

$$E_u = F_c(F_{asv}(U), F_{raw}(U)) \quad (1)$$

3.2. Multi-task Classifiers

Since end-to-end SASV systems need to determine if input speech is bonafide, and if so, if it is the target speaker, this problem can be formulated as a multi-task classification problem. Two classifiers are used to predict spoof attacks / speaker id independently, with shared feature embeddings from the encoder. The CM classifier C_{cm} receives all inputs and predicts confidence scores, indicating if the input is believed to represent a spoofing attack. A bonafide mask layer is placed before the ASV classifier, C_{asv} , so that losses produced by the ASV classifier are only from bonafide speech. Binary cross entropy(BCE) loss and AAM-softmax loss are used for the CM and ASV classifiers, as described in Equation 3 and Equation 2[26] respectively.

$$L_{asv} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i, i}) + m)}}{e^{s(\cos(\theta_{y_i, i}) + m)} + \sum_{j=1, j \neq i}^J e^{s(\cos(\theta_{j, i}))}} \quad (2)$$

$$L_{cm} = -\frac{1}{N} \sum_{i=1}^N y_i^{cm} \log C_{cm}(E) + (1 - y_i^{cm}) \log (1 - C_{cm}(E)) \quad (3)$$

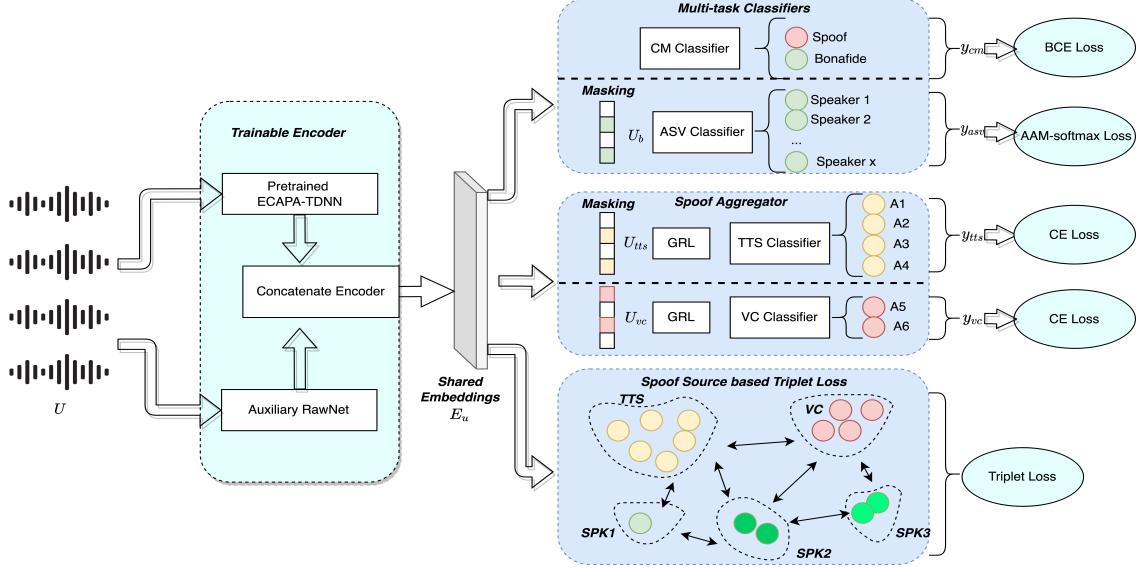


Figure 3: Model structure of the SA-SASV system. The shared embedding is fed into multiple classifiers. The feature space from the encoders optimized by combinations of various loss functions.

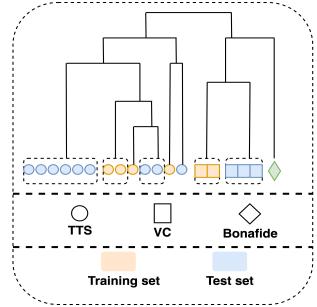


Figure 4: Results of agglomerative clustering on ASVSpoof 2019 LA Dataset.

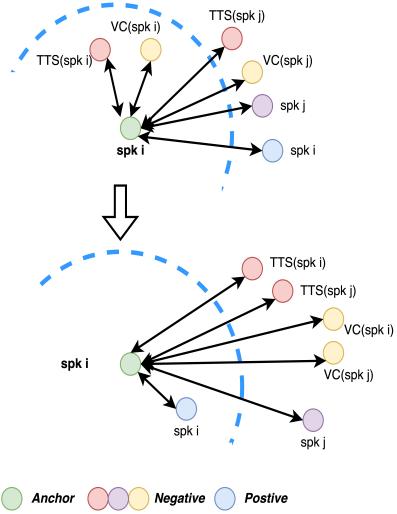


Figure 5: Training based on spoof-source based triplet loss. Positive samples (utterances with the same labels) are pulled closer and negative samples are pushed away.

3.3. Spoof Aggregator

In the SASV task, utterances, U , consists of spoof attack samples, U_s , and bonafide speech samples, U_b . As shown in Figure 2, U_s should have a relatively dense distribution in the shared feature space. However, it is challenging to aggregate the various spoofing attacks together due to their intrinsic differences in speech generation methods. This inherent difficulty in separating the two is evidenced by analyzing U_s from different sources using agglomerative clustering [27], as shown in Figure 4. The results indicate that U_{tts} , which represents produced with Text-to-Speech(TTS), and U_{vc} , which represents samples from Voice Conversion(VC), U_{vc} , tend to be closer in corresponding feature space. Thus, we conjecture that U_{tts} and U_{vc} need to be aggregated into two clusters in the feature space of SASV systems.

We use two adversarial learning layers to construct a spoof aggregator so that U_{tts} and U_{vc} aggregate separately. We labeled the U_s as $A1 \dots A6$, representing the spoof type, where $A1$ to $A4$ are from U_{tts} and $A5$ to $A6$ are from U_{vc} . Followed by a masking layer, E_{tts} and E_{vc} are sent to C_{tts} and C_{vc} , where each independently attempts to predict what spoof type U_s corresponds to. The cross entropy loss for both classifiers is shown in Equation 4

$$L_{tts} = L_{vc} = \frac{1}{N} \sum_i^N y_i^A \log C_{spoof}(E) \quad (4)$$

Since we want our embedding, E , to mix spoof attacks from the same types of generation mechanisms together, so that C_{tts} and C_{vc} fail to distinguish different spoofing attack types, a gradient reverse layer(GRL) is added before the classifiers to maximize L_{tts} and L_{vc} .

3.4. Spoof source based triplet loss

The shared feature space from SASV systems tends to be differentiated by U_{tts} , U_{vc} , and different speakers U_{spk_i} . In other words, the goal is for E with the same labels to be relatively compactly clustered and the overall cluster separated from E

Subsets	#Bonafide	Spoofed	
		#TTS	#VC
Training	2580	A1-A4	A5-A6
		15200	7600
Development	2548	A1-A4	A5-A6
		14864	7432
Evaluation	7335	A7-A16	A17-A19
		49140	14742

Table 1: Statistics of ASVSpoof 2019 LA Dataset

samples with different labels. Boundaries between the E samples with different labels should be distinct. To help achieve this outcome, rather than applying speaker-based triplet loss, we applied spoof source-based Triplet loss. Conventional triplet loss is described as Equation 5:

$$L_t = \|E^a - E^p\| - \|E^a - E^n + m\| \quad (5)$$

As shown in Figure 3, E_i is labeled as TTS , VC and SPK_i , where SPK_i indicates the i th speaker. The goal is to cluster, E_i samples, with same labels as densely as possible and scatter SPK_i to make it far away from SPK_j , TTS and VC , as shown in Figure 5. Thus, for an utterance from speaker i , U_{spk_i} , the spoof source based triplet loss is updated as shown in Equation 6

$$L_{st} = L_t(E_a, E_p, E_{tts}) + L_t(E_a, E_p, E_{vc}) + \sum_{j=0}^N L_t(E_a, E_p, E_{spk_j}) i \neq j \quad (6)$$

3.5. Overall Loss Function

As shown in Figure 3, the overall loss for AS-DGASV is determined by all of its constituent decoders, which includes five different loss functions, as shown in Equation 7.

$$L_{sasav} = L_{cm} + \lambda_1 L_{asv} + \lambda_2 L_{tts} + \lambda_3 L_{vc} + \lambda_4 L_{ts} \quad (7)$$

4. Experiment

4.1. Experiment Setting

Dataset. The SASV challenge permits the VoxCeleb2 dataset [28] and the ASVspoof 2019 LA dataset [12] for training the ASV and CM models. The VoxCeleb2 database consists of over 1 million utterances from 6,112 speakers and is designed for the ASV task, without spoofed data. The ASVspoof 2019 LA dataset, on the other hand, is prepared for the CM tasks, containing 6 types of spoof attacks in the training set and another 11 types of spoof attacks in the evaluation set, where the SASV models are tested. As shown in Table 1, the models need to generalize training attacks (A1-A6) to unseen attacks (A7-A19).

We use the VoxCeleb2 dataset to pre-train the ECAPA-TDNN and our model is fine-tuned on the ASVspoof 2019 LA dataset.

Metrics. We evaluated our model performance based on the SASV-EER, which is the primary metric in the SASV challenge. As shown in Table 3, only target speakers are labeled as positive and both non-target bonafide and spoof attacks are labeled as negative in the SASV-EER. The SV-EER and SPF-EER, are complements to SASV-EER, and reflect models' capability in different subsets of the full trials. Compared to the

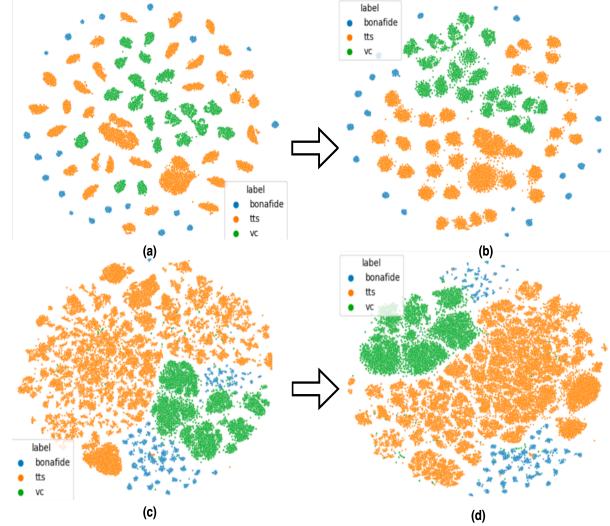


Figure 6: Visualization of the feature space in SA-SASV using t-SNE. (a) Naive multi-task classification on the development set. (b) SA-SASV on the development set. (c) Naive multi-task classification on the evaluation set. (d) SA-SASV on the evaluation set.

EER used in the ASVSpoof challenge, the SPF-EER only tests model performance in trials based on bonafide target speakers with spoofed speech.

Baseline. The SASV challenge provides two baseline models using state-of-the-art ASV and CM systems with different fusion strategies. **Baseline1** adopts a score-sum ensemble, which uses a naive sum function to integrate non-calibrated scores from the ASV and CM systems. While this method does not consider the difference between scores from different systems, scores of ASV systems are cosine similarity and scores of CM systems are from classifiers. **Baseline2** uses an extra network as a fusion strategy that takes embeddings from pre-trained ASV and CM systems to produce the final scores.

4.2. Results Discussion

4.2.1. Results

Visualization. To observe the updates of the features space produced by our encoder, we visualized utterances in the development and evaluation set using the t-SNE as shown in Figure 6. The left side shows the distribution of samples from the naive multi-task classifier without spoof-source-based triplet loss and the spoof aggregator and the right side shows the updated distribution using SA-SASV. Compared to naive the multi-task classifier, both spoof attacks from TTS and VC tend to have denser clustering and cleaner boundaries, making TTS, VC, and bona fide easier to differentiate.

4.2.2. Model Comparison with other SASV systems

We compared the characteristics and performance on the ASVSpoof 2019 LA dataset of SA-SASV with other SASV systems as shown in Table 2. Compared to other ensemble-based systems, SA-SASV takes advantage of a single training phase, intending to build a single representation in the feature space for utterances from different sources. Our SA-SASV improves SASV-EER from 6.05% (the prior best-performed INN(AUE)

Models	Inputs	Encoders	Training		Fusion	EER-SASV
			Phase1	Phase2		
SASV-Baseline1 [15]	raw waveforms, Fbanks	ECAPA-TDNN, AASIST	ASV, CM systems	\	Score	19.15
SASV-Baseline2 [15]	raw waveforms, Fbanks	ECAPA-TDNN, AASIST	ASV, CM systems	concatenated embeddings	Embeddings	8.76
Cascaded CM/ASV [22]	MFCC STFT	LC-GRNN, X-Vector	ASV, CM systems	\	\	7.67
2-stage PLDA [22, 19]	MFCC	X-Vector	PLDA	PLDA	\	28.40
Triplet TDNN [22, 24]	MFCC, CQCC	TDNN	TDNN	PLDA(CM) PLDA(ASV)	Score	8.99
INN(AUE) [22]	MFCC, STFT	LC-GRNN, B-Vector	ASV, CM systems	concatenated embeddings	Embeddings	6.05
SA-SASV	raw waveforms, Fbanks	ECAPA-TDNN ARawNet	SA-SASV	\	\	4.48

Table 2: Comparison on characteristics and performance of different SASV systems.

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

Table 3: Metrics to evaluate SASV systems.

	EER(%)		
	SASV	SV	SPF
Development	0	0	0
Evaluation	4.48	7.09	1.07

Table 4: Results on the ASVspoof 2019 Dataset

system) to 4.48%.

5. Conclusion

We proposed an end-to-end SA-SASV model, which takes both hand-crafted features and raw waveforms as input. In addition, our model is optimized with multiple loss functions to aggregate TTS, VC, and different speakers separately. Results show that the feature space of SA-SASV is better able to distinguish spoof attacks and identify speakers versus prior published approaches. Further, the SASV-EER is improved from the 6.05% produced by prior state of the art approaches to 4.48% without an multi-model ensembling strategy. Using an ensemble would likely further boost the performance and we will explore this in future work.

6. References

- [1] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using gmm supervectors for speaker verification,” *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.
- [2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. 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] G. Heigold, I. Moreno, S. Bengio, and N. 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] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. 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, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.
- [6] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [7] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [8] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, “Open source voice creation toolkit for the mary tts platform,” in *12th Annual Conference of the International Speech Communication Association-Interspeech 2011*. ISCA, 2011, pp. 3253–3256.
- [9] D. Matrouf, J.-F. Bonastre, and C. Fredouille, “Effect of speech transformation on impostor acceptance,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.
- [10] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilç, M. Sahidullah, and A. Sizov, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” in *Sixteenth annual conference of the international speech communication association*, 2015.
- [11] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” 2017.
- [12] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, “Asvspoof 2019: Future horizons in spoofed and fake audio detection,” *arXiv preprint arXiv:1904.05441*, 2019.
- [13] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, “Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection,” *arXiv preprint arXiv:2109.00537*, 2021.
- [14] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi *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.
- [15] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, “Sasv challenge 2022: A spoofing aware speaker verification challenge evaluation plan,” *arXiv preprint arXiv:2201.10283*, 2022.

- [16] Z. Teng, Q. Fu, J. White, M. Powell, and D. C. Schmidt, “Complementing handcrafted features with raw waveform using a light-weight auxiliary model,” *arXiv preprint arXiv:2109.02773*, 2021.
- [17] Y. Pan, F. Sun, Z. Teng, J. White, D. C. Schmidt, J. Staples, and L. Krause, “Detecting web attacks with end-to-end deep learning,” *Journal of Internet Services and Applications*, vol. 10, no. 1, pp. 1–22, 2019.
- [18] Q. Fu, Z. Teng, J. White, M. Powell, and D. C. Schmidt, “Fastaudio: A learnable audio front-end for spoof speech detection,” *arXiv preprint arXiv:2109.02774*, 2021.
- [19] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. 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.
- [20] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, “Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion,” in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*. ISCA, 2018.
- [21] H.-j. Shim, J.-w. Jung, J.-h. Kim, and H.-j. Yu, “Integrated replay spoofing-aware text-independent speaker verification,” *Applied Sciences*, vol. 10, no. 18, p. 6292, 2020.
- [22] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, “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, 2020.
- [23] Y. Zhao, R. Togneri, and V. Sreeram, “Multi-task learning-based spoofing-robust automatic speaker verification system,” *Circuits, Systems, and Signal Processing*, pp. 1–22, 2022.
- [24] J. Li, M. Sun, X. Zhang, and Y. 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.
- [25] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” *arXiv preprint arXiv:2110.01200*, 2021.
- [26] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, “Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.
- [27] 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 synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, p. 101114, 2020.
- [28] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.