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 Spoofing-Aware Speaker Verification (SASV) 2022 challenge was proposed to encourage the development of integrated SASV systems with new metrics to evaluate joint model performance. This paper proposes an ensemble-free end-to-end solution, known as Spoof-Aggregated-SASV (SA-SASV) to build a SASV system with multi-task classifiers, which are optimized by multiple losses and has more flexible requirements in training set. The proposed system is trained on the ASVspoof 2019 LA dataset, a spoof verification dataset with small number of bonafide speakers. Results of SASV-EER indicate that the model performance can be further improved by training in complete automatic speaker verification and countermeasure datasets.

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]. However, current ASV systems are still vulnerable to spoofing attacks, such as text-to-speech (TTS) [5, 6, 7] and voice conversion (VC) [8]. ASV systems can also 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 [9, 10, 11, 12], which can complement ASV systems with an anti-spoof model. The anti-spoof model 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 (t-DCF) [13], 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 whether 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 [14] 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 zero-effort and spoofed impostors. The SASV challenge provides two baseline systems by applying different fusion strategies (score-level fusion and embedding-level fusion) to pre-trained ASV and CM systems.

Figure 1 shows potential solutions to the SASV problem. Red/green lines indicate the following training stages: (a) Cascaded ASV/CM systems, (b) Fusions of scoring prediction, (c) Fusions of scoring and feature embedding, (d) Fusions of feature embedding, and (e) End-to-End SASV systems.

Figure 1: Feasible Solutions to Build Integrated SASV Systems.

This paper proposes a fully trainable end-to-end SASV system, called Spoof-Aggregated Spoofing Aware Speaker Verification System (SA-SASV), that combines a pre-trained ASV system with a lightweight raw waveform encoder to form the overall encoder [15]. This paper expands upon our prior experience that showed how encoding can be a key aspect of these types of anomaly detection problems [15, 16, 17]. Multiple classifiers and spoof-source-based triplet loss functions are employed to enhance model performance in generating the shared SASV feature space.

The remainder of the paper is organized as follows: Section 2 reviews related research on SASV systems; Section 3 discusses the model architecture of our SA-SASV Systems; Section 4 analyzes experiment results; and Section 5 presents concluding remarks.

2. Related Work

A 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 solu-
tions 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. [18] 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. [19] proposed a two-dimensional score modeling method to get a single score threshold for both ASV and CM systems.

Shim et al. [20] 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. [21] 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. [22] 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. [23] 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 SASV 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 from same model.

3. AS-DGASV Model Architecture

Compared to independent CM 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 and (2) in the clusters of bonafide speech sources from different speakers should be sparsely distributed so that models can distinguish between different speakers. Figure 2 shows how 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.

Figure 3: Model Structure of the SA-SASV System.

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 $F_{raw}$.

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

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 is formulated as a multi-task classification problem. Two classifiers are used to predict spoof attacks and 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.

3.3. Spoof Aggregator

In the SASV task, utterances, $U$, consist 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. It is hard, however, to aggregate the various spoofing attacks together due to their intrinsic differences in speech generation methods. This inherent difficulty in sep-
The shared feature space from SASV systems tends to be differentiated by $\text{SPK}_i$, where $\text{SPK}_i$ indicates the $i$th speaker. The goal is to cluster, $\text{SPK}_i$, samples, with same labels as densely as possible and scatter $\text{SPK}_i$, to make it far away from $\text{SPK}_j$, $\text{TTS}$ and $\text{VC}$, as shown in Figure 5. This figure shows that positive samples (utterances with the same labels) are pulled closer and negative samples are pushed away. Thus, for an utterance from speaker $i$, $U_{spk_i}$, the spoof source based triplet loss is updated as shown in Equation 3.

$$L_{st} = ||E^+ - E^-|| - ||E^+ - E^+ + m||$$

As shown in Figure 3, $E_i$ is labeled as $\text{TTS}$, $\text{VC}$ and $\text{SPK}_i$, where $\text{SPK}_i$ indicates the $i$th speaker. The goal is to cluster, $\text{SPK}_i$, samples, with same labels as densely as possible and scatter $\text{SPK}_i$, to make it far away from $\text{SPK}_j$, $\text{TTS}$ and $\text{VC}$, as shown in Figure 5. This figure shows that positive samples (utterances with the same labels) are pulled closer and negative samples are pushed away. Thus, for an utterance from speaker $i$, $U_{spk_i}$, the spoof source based triplet loss is updated as shown in Equation 3.

$$L_{st} = L_t(\text{E}_a, \text{E}_p, \text{E}_{\text{it}}) + L_t(\text{E}_a, \text{E}_p, \text{E}_{\text{vc}})$$

$$+ \sum_{j=0}^{N} L_t(\text{E}_a, \text{E}_p, \text{E}_{\text{spk}_j}) i \neq j$$

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 4.

$$L_{\text{SASV-SASV}} = L_{\text{cm}} + \lambda_1 L_{\text{AV}} + \lambda_2 L_{\text{it}} + \lambda_3 L_{\text{vc}} + \lambda_4 L_{\text{ts}}$$

4. Analysis of Experimental Results

4.1. Experiment Setting

Dataset. The SASV challenge permits the VoxCeleb2 dataset [26] and the ASVSpoof 2019 LA dataset [11] 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. 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. 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 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. Ablation Study on the Proposed Model

Configuration. An ablation study was conducted to investigate the effects of the different components on the performance of the SA-SASV system. As shown in Table 2, we evaluated our model with varying configurations of (1) just spoof source-based triplet loss, (2) just spoof aggregator, (3) and the two combined. Results indicate the absence of either component will reduce the SASV-EER of the SA-SASV model and configure-


| Models                  | Inputs            | Encoders          | Training                                             | Ensemble       | EER-SASV |
|-------------------------|-------------------|-------------------|------------------------------------------------------|----------------|----------|
| SASV-Baseline1 [14]     | raw waveforms, Fbanks | ECAPA-TDNN, AASIST | ASV, CM systems                                      | \              | Score    | 19.15   |
| SASV-Baseline2 [14]     | raw waveforms, Fbanks | ECAPA-TDNN, AASIST | ASV, CM systems                                      | \              |          |         |
| Cascaded CM/ASV [21]    | MFCC, STFT        | LC-GRNN, X-Vector | ASV, CM systems                                      | \              |          | 7.67    |
| 2-stage PLDA [21, 18]   | MFCC              | X-Vector          | PLDA, PLDA                                           | \              | 28.40    |
| Triplet TDNN [21, 23]   | MFCC, CQCC, STFT  | TDNN              | PLDA(CM) PLDA(ASV)                                   | Score          | 8.99     |
| INN(AUE) [21]           | MFCC, STFT        | LC-GRNN, B-Vector | ASV, CM systems                                      | \              |          | 6.05    |
| SA-SASV                 | raw waveforms, Fbanks | ECAPA-TDNN        | SA-SASV                                              | \              |          | 4.86    |

Table 1: Comparison on characteristics and performance of different SASV systems.

| Configuration       | SASV | SV   | SPF  |
|---------------------|------|------|------|
| ECAPA-TDNN          | 22.38| 0.83 | 29.32|
| SASV-Baseline1      | 19.15| 35.1 | 0.5  |
| SASV-Baseline2      | 8.75 | 16.01| 12.23|
| Ours                |      |      |      |
| SA-SASV w/o triplet | 5.82 | 9.14 | 2.12 |
| w/o spoof aggregator| 5.90 | 9.96 | 0.68 |
| naive multi-task classifier | 5.58 | 9.05 | 0.83 |

Table 2: Ablation study on the AS-SASV system.

The proposed model shows different generalization capabilities in SV and SPF tasks. Even though the SV-EER of the model reaches 0 in the training stage, it has limited ability to generalize the SV task to the evaluation set, which only contains unseen speakers. As a result, the overall model performance drastically decreased compared to SV-EER. We also noticed that, due to the overfitting problem, compared to SPF-EER, SV-EER in all models with different configurations tends to have unstable results. However, the SPF-EER of the model shows consistency from training to evaluation set, the best SPF-EER reaches 0.5, which is better than the baseline single CM system.

In conclusion, the model can detect unseen spoof attacks and has trouble distinguishing unknown speakers in the evaluation set. We conjecture the performance difference stems from data distribution in the training set. Only 40 speakers are contained in the training set and the ASV task usually requires a larger number of speakers to build features of human utterance, e.g., 5,994 speakers are included in the VoxCeleb2 dataset.

Although parts of our encoder are pre-trained on the VoxCeleb2 dataset, it only gave our model a feasible initializing strategy. During the training stage, the bonafide cluster in our new feature space is highly overfitted. The results of SPF-EER and SV-EER therefore show a different tendency in the training and evaluation stages. We believe it is a reasonable solution to train end-to-end SASV systems on complete ASV and CM datasets to avoid overfitting.

Visualization. To observe the updates of the features space produced by our encoder, we visualized utterances in the evaluation set using the t-SNE, as shown in Figure 6. The left side...
6. References

[1] 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.

[2] 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.

[3] 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. 3329–3333.

[4] B. Desplanques, J. Thiennout, and K. Demuynck, “Eca-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” arXiv preprint arXiv:2005.07143, 2020.

[5] 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.

[6] 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.

[7] M. Schröder, M. Charfuelan, S. Panmi, and I. Steiner, “Open source voice creation toolkit for the mary its platform,” in 12th Annual Conference of the International Speech Communication Association-Interspeech 2011. ISCA, 2011, pp. 3253–3256.

[8] 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.

[9] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilč, 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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, “Asv challenge 2022: A spoofing aware speaker verification challenge evaluation plan,” arXiv preprint arXiv:2201.10283, 2022.

[15] Z. Teng, Q. Fu, J. White, M. Powell, and D. C. Schmidt, “Complementing handcrafted features with raw waveform using a lightweight auxiliary model,” arXiv preprint arXiv:2109.02773, 2021.

[16] 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.

[17] 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.

[18] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, “Joint speaker verification and anti-spoofing in the i-vector space,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 821–832, 2015.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “Aassist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” arXiv preprint arXiv:2110.01200, 2021.

[25] 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.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” arXiv preprint arXiv:1806.05622, 2018.