DeepASV submission for the SASV Challenge 2022

Woo Hyun Kang, Jahangir Alam, Abderrahim Fathan

Computer Research Institute of Montreal (CRIM)

woohyun.kang,jahangir.alam,abderrahim.fathan@crim.ca

Abstract

In this paper, we provide description of our submitted systems to the spoof-aware speaker verification (SASV) Challenge 2022. The main objective of this challenge is to develop a speaker verification system robust to both zero-effort imposter access attempts and spoofing attacks. In order to achieve this, we have trained two different ensembled models and an end-to-end spoof-aware speaker verification system and fused their scores to obtain the final speaker verification score. The submitted system resulted in spoof-aware speaker verification equal error rate of 2.48% on the evaluation set.

Index Terms: speaker recognition, voice spoof detection, spoof-aware speaker verification, higher-order statistics pooling, hybrid neural network, SASV 2022 Challenge

1. SASV Challenge 2022

The spoof-aware speaker verification (SASV) challenge provides a standard benchmark for evaluating speaker verification systems on a spoof attack scenario. More specifically, the SASV challenge aims to develop a new solution for speaker verification robust to both zero-effort imposter access attempts and spoofing attacks.

For training the systems, the SASV challenge allows the following datasets:

- VoxCeleb 2 [1],
- ASVSpoof 2019 LA train set [2],
- ASVSpoof 2019 LA development set [2].

The evaluation is done on the ASVSpoof 2019 LA automatic speaker verification (ASV) protocol [2], where three equal error rates are measured:

- SASV-EER: the primary metric for the challenge. For computing the SASV-EER, bonafide utterances from the target speaker are considered as positive and the rest are considered as negative.
- SPF-EER: the spoof countermeasure metric. For computing the SPF-EER, bonafide utterances from the target speaker are considered as positive and all spoof utterances are considered as negative.
- SV-EER: the ASV countermeasure metric. For computing the SV-EER, bonafide utterances from the target speaker are considered as positive and all non-target bonafide utterances are considered as negative.

More details about the SASV challenge can be found in the evaluation plan [3].

2. System description

In our submission, we have constructed 3 different spoof-aware speaker verification systems and fused their scores together to obtain the final spoof-aware speaker verification score.

2.1. Embedding fusion systems

In this section, we described the embedding fusion-based spoof-aware speaker verification systems we have trained. The embedding fusion system consists of 3 leakyReLU [4] hidden layers with 256, 128, 64 nodes respectively. The last hidden layer is followed by a 2 noded output layer, where each node represents the positive and negative SASV scores respectively. The embedding fusion system takes the embeddings extracted from pre-trained anti-spoof counter measure system and speaker verification system as input. More specifically, the following embeddings are fed into the system:

- Enrollment embedding: the speaker model embedding extracted from a pre-trained speaker verification system,
- Query embedding: the embedding of a test utterance extracted from a pre-trained speaker verification system,
- Countermeasure embedding: the embedding of a test utterance extracted from a pre-trained anti-spoofing countermeasure system.

The embedding fusion system is trained using the binary cross-entropy loss function.

In our submission, we have trained two different embedding fusion systems:

- Baseline: the baseline embedding fusion configuration provided by the SASV challenge orgatizers, where the fusion system was trained on top of embeddings extracted from a pre-trained ECAPA-TDNN [5] ASV system and a AASIST [6] countermeasure system [3].
- HNN-ASV/HOSP-CM: Embedding fusion system trained on top of embeddings extracted from a pre-trained hybrid neural network (HNN) [7, 8] ASV system and a TDNN countermeasure system with higher order statistics (HOSP) [9].

In both embedding fusion systems, the ASV network was pre-trained on the VoxCeleb2 dataset and the countermeasure system was pre-trained on the ASVSpoof2019 LA train set.

2.2. End-to-end spoof-aware speaker verification system

In addition to the embedding fusion systems, we have also trained an end-to-end spoof-aware speaker verification system (E2E-SASV). The end-to-end system is based on the ECAPA-TDNN architecture [5], which takes the LFCC feature as input. The ECAPA-TDNN system was pre-trained on the VoxCeleb2 dataset using AAM-Softmax [10], and finetuned on the ASVSpoof2019 LA train set using a modified angular prototypical loss function [11], where the pairs with only spoof utterances are excluded from the training process. Moreover, we have applied a modified mixup regularization strategy during training, where the synthetic samples created from interpolating bonafide and spoof samples are considered as spoof samples. After fine-tuning, cosine similarity between the enrollment and test embeddings are computed.

	Dev			Eval		
System	SASV-EER [%]	SPF-EER [%]	SV-EER [%]	SASV-EER [%]	SPF-EER [%]	SV-EER [%]
ECAPA-TDNN baseline [3]	17.38	20.30	**1.88**	28.83	30.75	**1.63**
Score-fusion baseline [3]	13.07	**0.06**	32.88	19.31	0.67	35.32
Embedding-fusion baseline [3]	4.85	0.13	12.87	6.37	0.78	11.48
Embedding-fusion baseline (our)	4.85	0.13	12.80	6.40	0.76	11.55
Our submission	**2.43**	0.07	4.99	**2.48**	**0.65**	3.80

Table 1: *The development and evaluation set EERs of the baseline systems and our submission.*

2.3. Score-level fusion

To produce the final scores for submission, we have fused the scores generated from the Baseline, HNN-ASV/HOSP-CM, and E2E-SASV systems. For the score-level fusion, we have first normalized the individual scores using their respective development set scores via Gaussian normalization. After the normalization, the three scores were averaged and used as the final decision score.

3. Result

Table 1 shows the primary SASV-EER metric along with the SPF- and SV-EERs of our submitted system and the baselines on the development and evaluation set. From the results, our submitted system outperformed the baseline systems, achieving a relative improvement of 61.07% in terms of EER compared to the best performing baseline (i.e., Embedding-fusion baseline).

4. Conclusion

In this report, we described our submitted system on the SASV 2022 Challenge. In this challenge, we experimented with a score-level fusion framework of two different embedding fusion systems and an end-to-end spoof-aware speaker verification system. Our submitted system outperformed all the baseline systems, achieving a relative improvement of 61.07% in terms of EER compared to the best performing baseline.

5. Acknowledgment

The authors wish to acknowledge funding from the Government of Canada's New Frontiers in Research Fund (NFRF) through grant NFRFR-2021-00338. Authors also wish to acknowledge Ministry of Economy and Innovation (MEI) of the Government of Quebec for the continued support.

6. References

[1] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[2] ASVspoof consortium, *ASVspoof 2019:Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan*, 2019 (accessed May 13, 2020).

[3] Jee-won 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," https://sasv-challenge.github.io/, 2022, [Online; accessed 11-March-2022].

[4] A.L. Maas et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013.

[5] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[6] 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*, 2022.

[7] Jahangir Alam, Abderrahim Fathan, and Woo Hyun Kang, "Text-independent speaker verification employing CNN-LSTM-TDNN hybrid networks," in *Speech and Computer*, Alexey Karpov and Rodmonga Potapova, Eds., Cham, 2021, pp. 1–13, Springer International Publishing.

[8] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan, "Hybrid network with multi-level global-local statistics pooling for robust text-independent speaker recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1116–1123.

[9] Jahangir Alam, Abderrahim Fathan, and Woo Hyun Kang, "End-to-end voice spoofing detection employing time delay neural networks and higher order statistics," in *Speech and Computer*, Alexey Karpov and Rodmonga Potapova, Eds., Cham, 2021, pp. 14–25, Springer International Publishing.

[10] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[11] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee-Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," *Interspeech 2020*, Oct 2020.