Multi-task based fusion system for SASV 2022

Xingjia Xie[1], *Yiming Zhi*[2], *Haodong Zhou*[1], *Lin Li*[1], *Qingyang Hong*[2]

[1]School of Electronic Science and Engineering, Xiamen University, China
[2]School of Informatics, Xiamen University, China
{lilin,qyhong}@xmu.edu.cn

Abstract

Traditional automatic speaker verification (ASV) system will be greatly affected by spoofing attacks. A series of ASVspoof challenges is held every two years and dedicated to anti-spoofing work. This paper describes the system from XMUSPEECH for Spoofing Aware Speaker Verification Challenge 2022, which is a new competition related to the ASVspoof challenge. We propose a fusion strategy based on score-level fusion. For the task, we evaluate our system on ASVspoof 2019 LA development set and evaluation set which greatly improves the performance compared with the Baseline2. Our best submission obtained 1.155% SASV-EER on the evaluation set, while the performance on the development set is 0.723% SASV-EER.

Index Terms: SASV, ASVspoof, Speaker verification, Fusion-based system, BOSARIS

1. Introduction

Automatic speaker verifification (ASV) systems has improved dramatically in recent decades, this type of systems aim to verify the identity of the target speakers given a test speech utterances. For example, ECAPA-TDNN [1] models for speaker verifification systems achieves state-of-the-art (SOTA) performance in this field. However, it could be greatly affected by spoofing attacks, which contain synthesized, converted and replayed speech. According to the SASV official evaluation plan [2], when the SV-EER (speaker recognition equal error rate) of ECAPA-TDNN encounters a data set with a large number of unknown spoof speech such as the 2019 ASVspoof LA evaluation set [3], the system performance drops sharply from 1.63% to 23.83%. Therefore, it is necessary to study a good-performance anti-spoofing countermeasures system as a safety door for ASV systems. At present, many systems relies on the high-resolution features and robust models, and they produce ideal effects [4].

Automatic speaker verifification (ASV) is a biometric authentication task to determine the identification of a speaker from his or her audios. There are a number of significant technologies to constantly improve the performance of ASV systems, such as joint factor analysis (JFA) [5], i-vector based frameworks [6], end-to-end (E2E) [7] and deep embedding frameworks [8]. However, ASV systems have been encountering the spoofing attacks from the advanced speech synthesis algorithms and high fidelity replay devices. Currently, there are four known spoofed attacks, including accent mimic, text-to-speech (TTS), voice conversion (VC) and replay attack [9][10]. And in the speech community enables, many researchers have been developing the anti-spoofing countermeasure (CM) systems from two directions, including feature engineering and binary classifier [11]. The feature engineering focuses on researching high time-frequency resolution acoustic features to capture the hidden and unnatural signs processing, namely the spoof cues. On the other hand, effective classifiers have been developed to accurately discriminate the bonafide and spoof speech.

The rest of this paper is organized as follows. In Section 2, we briefly introduced the relevant information of ECAPA-TDNN, AASIST, as well as our multi-task system. Section 3 includes our experiments and results. Finally, Section 4 concludes the paper and indicates the future work.

2. System Description

For the challenge, inspired by the system from the team of Y Zhang et al [12], we adopt the similar system settings and the same dataset without any extra data other than score-level fusion method. The initial system sums the scores produced by the separate systems, we replace addition with multiplication and make a great improvment on the evaluation set and development set for SASV-EER. Based on this, we use the BOSARIS toolkit [13] to fuse multiple system's scores. Surprisingly, the three EERs calculated by the fused scores can be improved to some extent, compared with the original independent systems.

Baseline1 simply sums the score from ASV system and CM system to generates the final SASV score, as formula (1) shows. There is great numerical variation between the scores of these two systems in backbone network. Thus, we have a try to replace addition with multiplication, and generates the final SASV score, as formula (2) shows. The final score make a great improvment, as shown in Table 3 ID 6.

$$S_{\text{fusion}} = S_{cm} + S_{SV} \quad (1)$$

$$S_{\text{fusion}} = S_{cm} * S_{SV} \quad (2)$$

First of all, when it comes to the ECAPA-TDNN network, it achieved great success in speaker verifification in these years. This model use 1024 feature channels to expand network scale. In the SE-Block of the bottleneck, 256 is set as the size. The front-end feature extractor is followed by an attentive statistics pooling layer [14] that calculates the mean and standard deviations of the fifinal frame-level features. We adopt this SOTA ASV model as our ASV system to extract 192 dimensions vector for enroll utterances and test utterances.

Secondly, AASIST [4] is also used as the backbone network in the Baseline1 and Baseline2. AASIST has a graph attention network and a RawNet2 based encoder. The system uses the original waveform as the input information to learn the meaningful high-dimensional spectral time-domain feature map, and then extracts the graphic nodes of the feature map in time-domain and frequency-domain separately. Using the stack nodes that learn information from all nodes, the final output CM embedding is realized by connecting the average

and maximum values of each node. And we adopt this advanced CM model as our CM system to extract 160 dimensions vector for test utterances.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{c} e^{W_j^T x_i + b_j}} \quad (3)$$

$$L_{multi-task} = L_{spk} + L_{spoof} \quad (4)$$

As for our multi-task system, the frame level information in the front network layer is shared, however, when the network processing segment level information, two branches are divided. The left branch learns speaker classification and the right branch learns bonafide and spoof speech classification. The loss functions on both sides are softmax loss, as shown in formula (3). The loss function on the left is L_{spk}, and the loss function on the right is L_{spoof} , as shown in formula (4). We adopt the strategy of multi-task learning, and use the sum of spk-loss and spoof-loss as the joint loss function to optimize the network parameters. Besides, more configuration information of multi-task system is shown in Table 1.

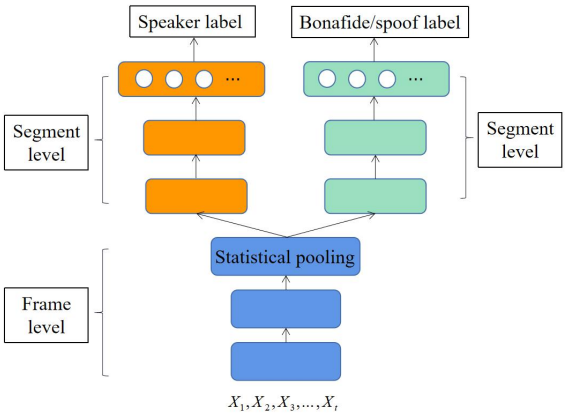

Figure 1: Multi-task system structure.

Table 1: Specific configuration of multi-task system.

Layer number	Layer name	Delay parameters	Frame number	Node number
1	Frame1	{t-2,t+2}	5	512
2	Frame2	{t-2,t,t+2}	9	512
3	Frame3	{t-3,t,t+3}	15	512
4	Frame4	{t}	15	512
5	Frame5	{t}	15	1500
6	Statistical pooling	[0,T)	T	3000
7	Segment6	{0}	T	512
8	Segment7	{0}	T	512
9	Softmax	{0}	T	Speaker number
7	Segment6	{0}	T	512
8	Segment7	{0}	T	512
9	Softmax	{0}	T	2

Similar to the score combination method in SASV Baseline 1, we implemented a triple-system score-level fusion model, which effectively combines the classifification scores of ASV and CM systems and decreases SASV-EER to 1.155% on the 2019 ASVspoof LA evaluation set.

3. Experiment

All datasets we used for training and validation are ASVspoof2019 [3] LA train partition, ASVspoof2019 LA development partition, and VoxCeleb 2 [15] as requested by the organizers.The ASVspoof2019 LA database consists of 12,483 bonafide and 108,978 spoof audios. The number of total speakers is less than 100, and the database contain 6 known spoofing attacks in the train and development partitions and 11 unseen spoofing attacks in the evaluation partition [3], therefore, it was generally only used for speech anti-spoofing. The VoxCeleb 2 database was widely used for ASV training, which contains far more audios and speakers than ASVspoof2019 LA database.

SASV-EER represents the equal error rate between target samples, nontarget samples and spoof samples, which is set as the main measure. SPF-EER and SV-EER were used as secondary indicators. SPF-EER measure only consider target samples and spoof samples, and SV-EER measure only considers target samples and nontarget samples, the specific difference is shown in Table 2.

Table 2: Description of EERs.

	Target	Nontarget	Spoof
SASV-EER	+	-	-
SV-EER	+	-	
SPF-EER	+		-

As for our multi-task system, we adopt the strategy of multi-task learning, and use the sum of spk-loss and spoof-loss as the joint loss function to optimize the network parameters. We extracted 80 dimensions for fbank feature, and set the training cycles for 21 epochs, batch_size is 512, learn_rate set to 0.001, use Adam optimizer, and use a single A40 GPU, weight_ Decay is 0.3.

The fusion of systems which include ID 7 and 8 could obtain 0.56% SASV-EER and 0.88% SV-EER on the development set obtained 1.16% SASV-EER, which achieved the best result in the Table 3.

BOSARIS toolkit plays a important role for our system. This toolkit provide score calibration based multi-system, and the parametric solution usually performs better on independent test data. Before fusion, our multi-task system only obtained 11.38% SASV-EER on the evaluation set, and 11.60% SASV-EER on the development set. But after the fusion of systems

include ID 5, 7 and 8 in Table 3, the fusion score obtained 1.16% SASV-EER on the evaluation set, and 0.72% SASV-EER on the development set, which is better than any of them. Before joining our system, the dual system fusion can not get a better score. Therefore, we can infer that our multi-task system can provide complementary information with the other two systems, so we can get a better score result after fusion.

Table 3: The results of *SASV Challenge.*

ID	Model	DEV			EVAL		
		SASV-EER(%)	**SV-EER(%)**	**SPF-EER(%)**	**SASV-EER(%)**	**SV-EER(%)**	**SPF-EER(%)**
1	ECAPA-TDNN [16]	17.38	1.88	20.30	23.83	1.63	30.75
2	Baseline1 [2]	13.07	32.88	**0.06**	19.31	35.32	**0.67**
3	Baseline2 [2]	4.85	12.87	0.13	6.37	11.48	0.78
4	Our Baseline2	4.78	12.80	0.10	6.33	11.32	0.80
5	Multi_task	11.60	9.761	12.19	11.38	8.36	12.41
6	Baseline1 with multiplication for score-level fusion [12]	2.16	4.18	0.20	2.89	4.28	0.89
7	pr_s_f [12]	1.09	2.02	0.07	1.53	1.96	0.80
8	Baseline1_s_i [12]	1.69	2.56	0.07	2.45	3.09	0.76
9	Fusion of ID 7 and 8	**0.56**	**0.88**	0.07	1.84	2.42	0.93
10	Fusion of ID 5, 7 and 8	0.72	1.39	0.07	**1.16**	**1.49**	0.77

4. Conclusions

It can be seen from the score part in the Table 3 that the simple summation method performs poorly due to the different score distribution of different systems. This problem can be effectively solved by standardizing and multiplying fractions using Sigmoid functions. It is surprising that a simple strategy can improve the SASV performance a lot. We proposed a simple but effective fusion-based method for spoofing aware speaker verification (SASV). The result suggests that the multiplication for score-level fusion has a better discrimination ability. In the future, we will further explore the differences between the various fusion strategies and improve our multi-task system.

5. References

[1] 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.

[2] J Jung, H Tak, and H Shim, "SASV Challenge 2022: A Spoofing Aware Speaker Verification Challenge Evaluation Plan," arXiv preprint arXiv:2201.10283, 2022.

[3] X Wang, J Yamagishi, and M Todisco, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language," 2020, 64: 101114.

[4] J Jung, H HS, and H Tak, "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," arXiv preprint arXiv:2110.01200 (2021).

[5] P Kenny, "Joint factor analysis of speaker and session variability:Theory and algorithms," CRIM, Montreal,(Report) CRIM-06/08-13, vol. 14, pp. 28-29, 2005.

[6] 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.

[7] D Snyder, P Ghahremani, D Povey, D Garcia-Romero, Y Carmiel, and S Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016, pp. 165-170.

[8] E Variani, X Lei, E McDermott, I L Moreno, and J GonzalezDominguez, "Deep neural networks for small footprint textdependent speaker verification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4052-4056.

[9] Z Wu, T Kinnunen, N Evans, J Yamagishi, C Hanilci, M Sahidullah, and A Sizov, "Asvspoof 2015: the fifirst automatic speaker verification spoofifing 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 spoofifing 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] Y Zhang, G Zhu, and Z Duan, "A New Fusion Strategy for Spoofing Aware Speaker Verification," arXiv preprint arXiv:2202.05253, 2022.

[13] N Brümmer and E De Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," arXiv preprint arXiv:1304.2865 (2013).

[14] K Okabe, T Koshinaka, and K Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," Proc. Interspeech, 2018.

[15] A Nagrani, J S Chung, and A Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.

[16] RK Das, R Tao, and H Li, "HLT-NUS SUBMISSION FOR 2020 NIST Conversational Telephone Speech SRE," arXiv preprint arXiv:2111.06671 (2021).