HCCL Total-Divide-Total SYSTEM FOR THE SASV Challenge 2022

Yuxiang Zhang[12], *Zhuo Li*[12], *Wenchao Wang*[12], *Pengyuan Zhang*[12]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China
[2]University of Chinese Academy of Sciences, China

{zhangyuxiang, lizhuo, wangwenchao, zhangpengyuan}@hccl.ioa.ac.cn

Abstract

This paper describes the Total-Divide-Total automic speaker verification and anti-spoofing integrated systems submitted to the Spoofing-Aware Speaker Verification (SASV) challenge. Shallow features in pretrained automatic speaker verification (ASV) systems were shown can be used for logical access spoofing speech detection. Based on the correlation between anti-spoofing and speaker verification, a SASV system with a Total-Divide-Total structure was proposed. Features were first extracted from the pretrained ASV system provided by the baseline. Then, the anti-spoofing system was trained as a branch. Finally the integrated scoring module of the two embeddings were trained based on score matrix operation. Submitted systems achieved EER of 3.07% on the development part of the Challenge dataset and 4.30% on the evaluation part. Performance improvements over the baseline system were obtained by adding only a few parameters to the baseline ASV system.
Index Terms: speaker verification, anti-spoofing, synthetic speech detection, SASV challenge

1. Introduction

As a type of biometric technology, automatic speaker verification (ASV) [1] has made significant progress in recent years. From early algorithms based on statistical machine learning [2, 3, 4] to today's deep learning algorithms [5, 6, 7], ASV systems with increasingly low equal error rates (EERs) are gradually being used in practice. However, a variety of spoofing attacks against ASV systems, including impersonation, replay, text-to-speech (TTS) and voice conversion (VC). These spoofing attacks can cause serious performance degradation of ASV systems [8].

To protect ASV systems from spoofing attacks, the biennial ASVspoof challenges were successfully held from 2015 to 2021 [9, 10, 11, 12]. Based on the datasets provided by the challenges, a large number of countermeasure (CM) systems with good performance for TTS and VC had emerged [13, 14, 15, 16]. Even in complex scenarios such as cross-channel, the CM systems could still achieve good results [17, 18, 19]. But these CM systems were built independently of ASV systems and were only aimed at spoofing attacks, not in conjunction with ASV systems. Although the challenges provided the tandem detection cost function (min t-DCF) [20] metric to simulate and evaluate the performance of CM systems in tandem with ASV systems, the metric still had significant limitations, as the ASV system was fixed. Few attempts have been made to jointly optimize ASV systems with CM systems. Therefore, it is of great significance to study the ASV system with anti-spoofing capability simultaneously.

The Spoofing-Aware Speaker Verification (SASV) challenge [21] was held as a special session of INTERSPEECH2022, which is expected to further promote research on ASV and CM integrated systems. In the SASV challenge, only logical access (LA), i.e., TTS and VC spoofing attacks, are considered. The Data for the SASV system is divided into three categories: bona-fide speech belonging to the target person, bona-fide speech of the non-target person, and spoofing attack speech. Only the bona-fide speech belonging to the target speaker was considered as positive samples.

In this paper, the HCCL Total-Divide-Total automic speaker verification and anti-spoofing integrated system was proposed. The shallow features in pretrained ASV systems were shown can be used for LA spoofing speech detection. This shallow feature is used in the subsequent dual-branch network for speaker verification (SV) and spoofing speech detection, respectively. The two embedded vectors obtained from each of the two branches were fed into an integrated scoring module to obtain the SASV score. The integrated scoring module included score matrix operation or convolution-based scoring module.

2. Method

The submitted systems were all based on the pretrained ECAPA-TDNN [7] baseline ASV system provided by organizers. The proposed Total-Divide-Total structure and score integration methods were described bellow.

2.1. Total-Divide-Total SASV structure

As shown in Figure 1, the "Total-Divide-Total" SASV structure consists of a total feature extraction network, dual branches for ASV and anti-spoofing CM embeddings extraction, and a total integrated scoring module.

The SASV challenge provided two baselines based on two subsystems and their EERs on the ASVspoof 2019 LA dataset. The ASV subsystem performed well on the SV task, achieving an EER of 0.83% on the evaluation set. Despite the poor performance of 29.32% EER on the anti-spoofing (SPF) task, it still had some discrimination ability for spoofing speech. This motivated us to try to obtain a countermeasure by training some layers in the given pre-trained ECAPA-TDNN ASV subsystem. Through using the features from the pre-trained ASV model, the same shallow features can be shared with the SV task thus reducing the number of parameters. The same feature dimension also facilitates the optimization of subsequent embedding scoring. The concatenated features obtained from the Multi-layer Feature Aggregation (MFA) were fed into the subsequent dual-branch network.

In the dual-branch structure, the SV branch was fixed and only the anti-spoofing branch was trained because of the excellent performance on the SV task of the pre-trained ASV system. The attentive statistics pooling and the first linear layer in CM branch was same as the ASV system. However, a linear layer is added after the CM embedding concatenated with the speaker embedding in an attempt to improve the performance of the CM

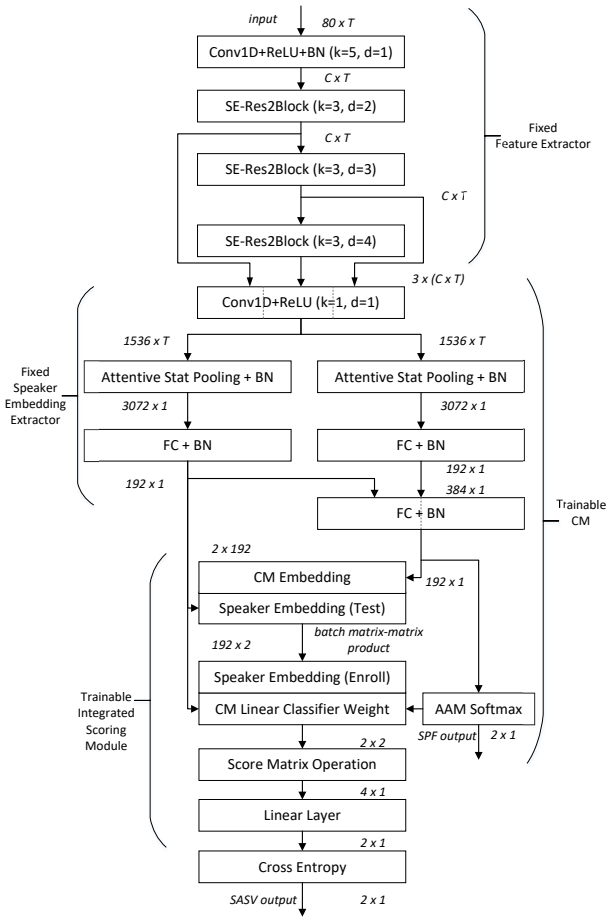

Figure 1: *Network topology of the Total-Divide-Total ECAPA-TDNN based SASV structure* .

branch by utilizing the speaker information.

After yielding the final ASV embeddings and CM embeddings separately, they were not used to scoring individually, but fed into the integrated scoring module described below.

2.2. Integrated scoring module

Two integrated scoring strategies were used in our integrated scoring module.

The first strategy is score integration through score matrix. Suppose that the test speaker embedding and the CM embedding are $\boldsymbol{E_{test}}$ and $\boldsymbol{E_{CM}}$ respectively, while the enroll speaker embedding and the weight of the last classifier in CM are $\boldsymbol{E_{en}}$ and $\boldsymbol{W}$ respectively. The final classification layer of CM is a 2×192-dimensional matrix, and the final score can be obtained by $S_1 - S_0$ when scoring, so the difference of the vectors obtained by $\boldsymbol{W} = \boldsymbol{W}_1 - \boldsymbol{W}_0$ was used as the scoring vector. Then the score matrix can be obtained from the following equation:

$$\boldsymbol{S} = \begin{bmatrix} \eta_1 & S_{SV} \\ S_{CM} & \eta_2 \end{bmatrix}_{2\times 2} = \begin{bmatrix} \boldsymbol{E_{test}} \\ \boldsymbol{E_{CM}} \end{bmatrix}_{2\times 192} \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{E_{en}} \end{bmatrix}^T_{2\times 192}$$

Since we argued that there was a correlation between ASV and spoofing speech detection, the elements on the diagonal of the score matrix were maintained. According to [22], the score matrix was transformed into a probability matrix $\boldsymbol{P}$ by the Sigmoid function:

$$\boldsymbol{P} = \sigma(\boldsymbol{S}) = \begin{bmatrix} \theta_1 & P_{SV} \\ P_{CM} & \theta_2 \end{bmatrix}$$

where σ denotes the sigmoid function.

In order to derive the probability of the bona-fide speech belonging to the target speaker, the probability matrix was then manipulated with some simple operations:

$$\begin{aligned} \boldsymbol{J} &= \boldsymbol{P} + \boldsymbol{P}^T + \boldsymbol{P} \cdot \boldsymbol{P} \\ &= \begin{bmatrix} \delta_1 + P_{SV}P_{CM} & \epsilon_1 P_{SV} + \epsilon_2 P_{CM} \\ \epsilon_1 P_{CM} + \epsilon_2 P_{SV} & \delta_2 + P_{SV}P_{CM} \end{bmatrix} \end{aligned}$$

where $\delta_1 = \theta_1^2 + 2\theta_1$, $\delta_2 = \theta_2^2 + 2\theta_2$, $\epsilon_1 = 1 + \theta_1$, $\epsilon_2 = 1 + \theta_2$

After obtaining the final probability matrix, the probability matrix was flattened into vector and the final classification was performed by a linear layer.

Another integrated scoring strategy is simpler. A matrix with shape of 5×192 was obtained by concatenating the above five vectors: $\boldsymbol{E_{test}}, \boldsymbol{E_{en}}, \boldsymbol{E_{CM}}, \boldsymbol{W}_1, \boldsymbol{W}_2$. Then, the final 192-dimentional embedding was obtained by a two-dimensional convolutional layer with a kernel of 5×1.

3. Experimental setup

3.1. Dataset and metrics

All experiments presented further were performed on the ASVspoof 2019 LA dataset [23]. Only the training set was used to train all systems. The development part was used for performance validation and weight tuning of the system fusion.

The metrics of the challenge are the classical EERs, including SASV-EER, SV-EER and SPF-EER. The SASV-EER does not distinguish between different speaker (zero-effort, non-target, or impostor) access attempts and spoofed access attempts. The SV-EER is traditional estimates of speaker verification performance estimated from a set of target and non-target trials. While the SPF-EER is similar to SV-EER except that non-target trials are replaced with spoofed trials.

3.2. Details of systems implementation

The rest of the proposed model was the same as the pre-trained ECAPA-TDNN ASV baseline system, except for the dual-branch architechture and intrgrated scoring module described above. The network leverages some of the latest advances in deep learning to achieve state-of-the-art performance on the VoxCeleb1-O [24] dataset and has been widely adopted in the community. The input features of the model were the 80-dimensional Fbank. In our implementation, the same data augmentation methods were used except the SpecAug.

During the training process, the same AAMSoftmax[25] was used to optimize the CM. The SASV loss was optimized using binary cross entropy loss with weights of $(0.1, 0.9)$.

The details of five single systems are as follows, TDT means Total-Divide-Total:

- TDT-1 and TDT-2: Two implementation of the proposed system described above. The integrated scoring strategy was based on the score matrix operation.
- TDT-A: Similar to the above system, but with learnable weights obtained by simple attention layer multiplied to the two Embeddings before calculating the score matrix.

Table 1: *Results in terms of three different EERs/% for the SASV 2022 development and evaluation subsets.*

	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
Baseline1 (score-sum)	32.88	35.32	0.06	0.67	13.07	19.31
Baseline2 (back-end ensemble model)	12.87	11.48	0.13	0.78	4.85	6.37
TDT-1	3.44	6.24	0.37	1.73	2.76	4.78
TDT-2	3.64	6.81	0.39	2.35	2.76	5.14
TDT-A	10.65	8.44	0.73	2.59	5.05	6.17
TDT-O	6.20	8.85	0.43	1.49	4.31	7.02
TDT-C	13.60	6.28	1.35	7.50	6.80	7.08
Fusion	4.25	5.62	0.20	1.21	3.07	**4.30**

- TDT-O: Similar to the TDT-1, but with the addition of a loss function to make the two weight vectors of the last classification layer of CM as orthogonal as possible.
- TDT-C: Similar to the TDT-1, but the integrated scoring strategy was the convolution-based scoring module.

The primary system submitted to the challenge is a fusion of individual systems at the score level. The fusion of subsystem scores is done with the same weights of 0.2.

4. Results and Discussion

The results for our single and fusion systems were presented in terms of SASV-EER, SV-EER and SPF-EER in Table 1.

Results obtained on the development and evaluation set of the SASV 2022 challenge dataset confirmed the efficiency of the Total-Divide-Total ECAPA-TDNN automic speaker verification and anti-spoofing integrated system for the SASV. The best single system obtained SASV-EER of 4.78%, almost 25% reduction compared to baseline2 system. The EER of the fusion system on the evaluation set is 4.30%, which is a further 10% reduction compared to the best single system.

Comparing the results of different single systems, it can be found that adding mechanisms such as attention to TDT systems can lead to performance degradation instead. Compared with the integrated scoring strategy based on score matrix operation, the scoring module based on convolution is worse in all aspects.

It also can be found that the SPF-EER performs better compared to the SV-EER, since only the spoofing speech detection and SASV were optimized. However, compared to the large gap between SV-EER and SPF-EER of the baseline systems, the two metrics of the proposed SASV system were more balanced.

5. Conclusion

This paper proposed the HCCL Total-Divide-Total automic speaker verification and anti-spoofing integrated systems submitted to the SASV 2022 challenge. Based on the superior performance of the pre-trained ASV system and the finding that the shallow features from pretrained ASV system can be used for spoofing speech detection task, it is possible to obtain a SASV system with more balanced performance by training a few parameters. An integrated scoring module based on matrix operations is also proposed, which naturally yields an integrated SASV score.

6. References

[1] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, vol. 14, no. 28-29, p. 2, 2005.

[4] 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.

[5] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[6] 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.

[7] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[8] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[9] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, 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*, Dresden, Germany, September 2015, pp. 2037–2041.

[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," in *INTERSPEECH 2017 Annual Conference of the International Speech Communication Association*, 2017, pp. 2–6.

[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," in *INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019.

[12] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed

and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[13] G. Lavrentyeva, A. Tseren, M. Volkova, A. Gorlanov, A. Kozlov, and S. Novoselov, "Stc antispoofing systems for the asvspoof2019 challenge," in *Interspeech*, 2019, pp. 1033–1037.

[14] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Proc. Interspeech 2021*, 2021, pp. 4259–4263.

[15] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8.

[16] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 22–28.

[17] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.

[18] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 89–93.

[19] Y. Zhang, W. Wang, and P. Zhang, "The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System," in *Proc. Interspeech 2021*, 2021, pp. 4279–4283.

[20] 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.

[21] 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.

[22] Y. Zhang, G. Zhu, and Z. Duan, "A new fusion strategy for spoofing aware speaker verification," *arXiv preprint arXiv:2202.05253*, 2022.

[23] 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.

[24] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.