The Clips System for Spoofing-Aware Speaker Verification Challenge 2022

Tingwei chen

Asp, China

chentingwei95@gmail.com

Abstract

This report describes our submission system to the Spoofing-Aware Speaker Verification Challenge 2022 (SASV 2022), the goal of SASV challenge is to develop a integrated system that can support the optimization of CounterMeasure (CM) system and Automatic Speaker Verification (ASV) system in a tandem manner. In this report, we proposed a fusion system that fixed the ASV system and constructed four Max-Feature-Map (MFM) layers to fine-tune the CM system, and the results show that our proposed fusion system can significantly improve the SASV equal error rate (SASV-EER) from 6.37% to 1.36% on the evaluation dataset and 4.85% to 0.98% on the development dataset.

Index Terms: speaker verification, anti-spoofing, audio spoofing detection, spoofing-aware speaker verification

1. Introduction

Automatic Speaker Verification (ASV) system solves the task whether two utterances are spoken by the same person. A recent shift towards neural network based speaker verification systems resulted in significantly better performance compared to the more traditional i-vector based systems [1]. The recently proposed Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) incorporates the elements of Res2Net with TDNN, ECAPA-TDNN applies attention mechanisms of Squeeze-and-Excitation (SE) to model the inter-dependencies between channels [2]. A variants of ECAPA-TDNN is called ECAPA CNN-TDNN[3], it adds a CNN-based front-end to incorporate frequency translational invariance to further strengthens ECAPA-TDNN. MFA-TDNN proposed a multi-scale frequency-channel attention (MFA) module to handle the shortcomings of TDNNs that plenty of filters are required to model the speaker characteristics occurring at some local frequency regions [4]. All these methods above only focus on the improvement of single-branch structures and neglect a multi-branch way of designing neural works. Adding parallel branches[5] can significantly enlarge the model capacity and enrich the feature space, which results in better model performance. Yu et al. [6] proposed a multi branch version of densely connected TDNN structure with a selective kernel (D-TDNN-SS) and this model achieved competitive performance in speaker verification. RepVGG as one of the multi-branch and re-parameterized models [7], has obtained great performance in VoxCeleb Speaker Recognition Challenge 2021. Although these methods mentioned above have improved the performance of SV a lot, it is likely that if the test utterance is spoofed by attacks, such as impersonation, replay, text-to-speech, voice conversion and so on, then the spoofed audio will deceive the ASV systems and degrade the ASV system performance. Studies have shown that ASV systems are vulnerable to spoofing attacks [8].

Some researchers have developed the spoofing countermeasure (CM) systems and audio deepfake detection systems to detect spoofing attacks. In [9], a new end-to-end spoofing detection system called AASIST is proposed to detect the spoofing audio, it utilizes a novel heterogeneous stacking graph attention layer to model artefacts spanning heterogeneous temporal and spectral domain with a heterogeneous attention mechanism and a stack node, which achieves the state-of-the-art performance. The goal of developing CM system is to protect the ASV systems from falsely accepting spoofing attacks. In order to better protect the ASV system from being spoofed and maintaining the discrimination ability on speaker identity, the CM component should be jointly optimized with the ASV system, so an integrated ASV and CM system is promising. Some works have proposed some frameworks to address such problem, but due to the lack of standard metrics and datasets, it is hard to benchmark the state-of-the-art spoofing aware speaker verification (SASV) system. Recently, the SASV challenge has been held to further encourage the study of integrated of ASV and CM.

In this report, we proposed a new system fusion framework for spoofing aware speaker verification system based on the pre-trained ASV system and CM system. we fixed the ASV system and construct four Max-Feature-Map (MFM) [10] layers to fine-tune the CM system, the results shows that the proposed fusion system has improved the SASV system performance a lot.

The report is organized as follows: Section 2 describes the system, including the ASV subsystem, CM subsystems and Fusion system. Section 3 describes the experiments setup and results. Finally, section 4 will give some concluding remarks.

2. System Description

The goal of the SASV challenge is to further improve robustness to both zero-effort impostor access attempts and spoofing attacks by providing a framework to support the optimization of CM and ASV systems operating in a tandem manner. It introduces two baseline built upon pre-trained state-art-of-art ASV and CM systems. The ASV system is an ECAPA-TDNN model trained on the VoxCeleb2 dataset, the CM system is an AASIST model trained on ASVspoof 2019 LA training set. Based on the ASV system and CM system, the Baseline1 involves a simple sum of the scores produced by the ASV and CM subsystems. Thus, no data is used for this baseline as it does not involve any training nor fine-tuning. The baseline2 involves the fusion of three embeddings: one extracted from an ASV enrolment utterance using the ECAPA-TDNN system; a second extracted in identical fashion from a test utterance; a third extracted from the same test utterance using the AASIST spoofing CM, the model is a vanilla multi-layer perception with three hidden layers, trained using ASVspoof 2019 LA train partition. Except the provided ASV system and CM system, we also train MFA-TDNN, ECAPA-RepVGG, DTCF-ResNet and ECAPA-SCNet to perform the ASV, and AASIST-RepVGG model to perform CM, and then system ensemble is done.

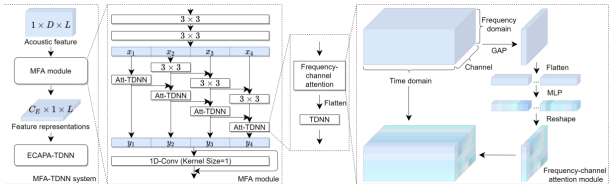

Figure 1: *the network architecture of MFA-TDNN*

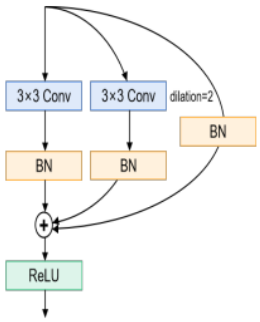

Figure 2: *the network architecture of RepVGG*

2.1. ASV subsystem

In this part, we give a brief introduction to our used ASV sub-systems.

2.1.1. MFA-TDNN

In order to handle the shortcoming of TDNNs, MFA-TDNN utilizes a multi-scale frequency-channel attention (MFA) module, which can characterize speakers at different scales through a novel dual-path design that consists of a convolutional neural network and TDNN. The network architecture of MFA-TDNN is shown in Fig.1, compared with standard Res2Net module, MFA-TDNN utilizes a Att-TDNN module to re-scale local feature responses adaptively by modeling their inter-dependencies between both channels and frequency bands, the detailed description of MFA-TDNN can be found in [4].

2.1.2. ECAPA-RepVGG

Most of state-of-the-art ASV systems focus on the single-branch structure, However, parallel branches can significantly enlarge the model capacity and enrich the feature space. Considering the great feature extraction power of RepVGG [7], so we adopt the multi-branch structure of RepVGG and substitute the Res2Net block in ECAPA-TDNN with RepVGG module, the detailed RepVGG module is shown in Fig.2.

2.1.3. DTCF-ResNet

Channel-wise attention mechanism has achieved remarked performance in ASV, but they do simple averaging on time and frequency feature maps before channel-wise attention learning and ignore the essential mutual interaction among temporal, channel as well as frequency scales. So DTCF-ResNet systems use duality temporal-channel-frequency attention to assemble the temporal and frequency information into the channel-wise attention, which can get great improvement in speaker verification performance. the detail of DTCF-ResNet can be found in [11]

2.1.4. ECAPA-SCNet

In this part, we substitute the Res2Net module with the self-calibration network (SCNet). The SCNet has the bottleneck residual block called SC-Block that integrates Res2Net and SKNet. According to the channels, it splits the input and uses normal convolution and self-calibration. In self-calibration, the split further converts into two different scale-spaces. One is the original space through convolution, the other obtains a smaller latent space as the spatial reference of the original space by average pooling (AvgPool), convolution, and Fully connection layer. Finally, the weighted output passes to another convolution and is concatenated with the normal convolution output. Over the self-calibration calculation, it can exploit different portions of convolutional filters in a heterogeneous way, integrate the different scale spaces. Thus, it allows the network to learn the weighted multi-scale features and avoid certain unrelated information. Moreover, inspired by DenseNet [12], we use the hierarchical technique to aggregate the output of different dilation rate SE-SCNet blocks.

2.1.5. ASV system fusion

In ASV system fusion, we just do the weight summation of the extracted embeddingss of DTCF-ResNet, ECAPA-SCNet, ECAPA-RepVGG and ECAPA-TDNN, the weights are determined by the performance of each model on the test dataset.

2.2. CM subsystem

The CM subsystem is aimed to detect whether the audio is spoofed, in this part, we give brief introduction of AASIST, which is the CM baseline system. We also substitute the ResNet module in AASIS Twith RepVGG module to extract speech representation.

2.2.1. AASIST

AASIST [9] is the state-art-of-art CM systems, compared with previous system, it has three contributions: (i) it proposes a heterogeneous stacking graph attention layer that can model spectral and temporal sub-graphs consists of a heterogeneous attention mechanism and a stack node to accumulate heterogeneous information, (ii) it uses a max graph operation that involves a competitive selection of artefacts, and (iii) it modifies the readout scheme. The following AASIST-RepVGG is a variant of AASIST.

2.2.2. AASIST-RepVGG

Consider the state-art-of-art performance of AASIST, so we keep the total framework of the AASIST and substitute the ResNet module with RepVGG module. Moreover, by setting the different kernel size of SincNet model, we can extract multi-scale feature from the audio, and then AASIST-RepVGG learn the speech representation from the multi-scale feature.

2.2.3. CM system fusion

As for the system fusion of CM, the way to ensemble the CM subsystem is the same as ASV system.

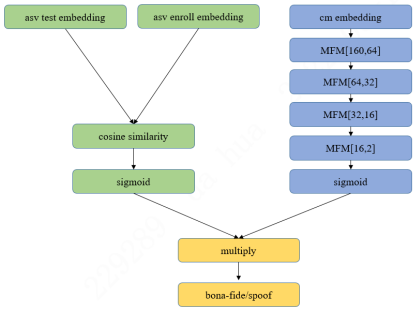

Figure 3: *the network architecture of Fusion System, the MFM means the Max-Feature-Map layer, and [a, b] indicates that the a is the input embedding dimension, and b is the output embedding dimension.*

2.3. Fusion system

The SASV challenge introduces two baselines built upon already-trained state-of-the-art and CM subsystem. Baseline1 is a score-level fusion method that sums the scores produced by separate systems. There is no training involved. Baseline2 is an embedding-level fusion method that trains a deep neural network based on the concatenated embeddings. The speaker and CM embeddings are fixed during training. This is similar to the method proposed in [13].

2.4. Proposed Fusion System

Inspired by the probabilistic fusion framework for spoofing-aware speaker verification [14], we take the same strategy as [14], in order to extend the generalization of the fusion system, we construct four Max-Feature-Map (MFM) layers to fine-tune the CM system. In the training phase of fusion system, we fixed the ASV system and CM system layer except the last layer, we add four MFM layers after the last layer of CM system. we take the AReLU funtion as the activation funtion, the detailed network information in shown Fig.3.

3. Experiments And Results

3.1. Datasets

We train the ASV subsystems on the development set of the VoxCeleb2 dataset, the VoxCeleb2 training set contains over one million utterances across 5994 different speakers, we also use the MUSAN [15] corpus and the Room Impulse Response and Noise Database [16] to augment the audio file in online training, all the ASV systems are trained by sharing the same parameters setting including learning rate, AAM-softmax and so on, as baseline system. The CM system is trained on the ASVspoof 2019 LA [17] dataset, the LA dataset consists of bona fide speech and a variety of TTS and VC spoofing attacks. The bona fide speech is collected from the VCTK corpus [18], while the speakers are separated into three subsets: training (Train), development (DEV) and evaluation (Eval), the spoofed speech in each subset is targeted to spoof the corresponding speakers, the algorithms for spoofing attacks in the evaluation sets are totally different from those in the Train and Dev sets, details are shown in Table 1. As the training of CM system, we take the same training strategy as the AASIST. The training strategy of fusion system is also the same as baseline2.

Table 1: *Summary of the ASVspoof 2019 LA dataset*

Partition	#speakers	Bona fide	Spoofing	Attacks type
Train	20	2580	22800	A01-A06
Dev	20	2548	22296	A01-A06
Eval	67	7355	63882	A07-A19

3.2. Evaluation metrics

Equal error rate (EER) is widely used for binary classification problems, especially in speaker verification and anti-spoofing. It is calculated by setting a threshold such that the miss rate is equal to the false alarm rate. The lower the EER is , the better the discriminative ability has the binary classification system.

SASV-EER is used as the primary metric to evaluate the SASV performance. The SV-EER and SPF-EER are auxiliary metrics to assess the performance of ASV and CM sub-tasks, respectively. Note that the SPF-EER is different from the common EER used in the anti-spoofing community. The difference is that the non-target class is not taken into consideration here but is regarded as the same positive class (bona fide) in the CM community. The description of EERs can be found in Table 2. The Test utterance falls into either of three classes. For all of the EERs mentioned above, only the target class is considered positive samples.

Table 2: *Description of EERs. The system involves enrollment utterance(s) and a test utterance(s). Enrollment utterance(s) is bona-fide and test utterance(s) belongs to either of the three types*

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

3.3. Results

In order to demonstrate the effectiveness of our method, we compare our method with the baseline method in the SASV challenge and [14], the performance comparison is shown in Table. 3.

As shown in the Table 3, the ASV system and CM system performs well on their own tasks but has worse performance on the other task. For example, ECAPA-TDNN has the lowest SV-EER but a high value in SPF-EER, this indicates that the spoofed audio can degrade the ASV performance. AASIST system has the lowest SPF-EER but close to 50% SV-EER, it is because that all bona-fide speech, no matter target or non-target, are considered positive samples in training CM systems. In terms of SASV-EER, baseline1 and baseline2 method surpass the separate systems, showing the superiority of an ensemble solution for the SASV system. Based on the fusion method, we experiment two systems, one is that fusion system is trained with original ECAPA-TDNN and AASIST embeddings, we denote this as **clips**, the second is $\mathbf{clips-e}$ that is trained with ensemble ASV embeddings and ensemble CM embeddings.

Table 3: *Comparison of CM systems*

CM system	min t-DCF	EER(%)
AASIST(baseline)	0.0275	0.83
AASIST-RepVGG	0.0325	1.07

Table 4: *Comparison of ASV systems*

ASV system	EER(%)
ECAPA-TDNN(baseline)	0.96%
DTCF-ResNet	1.58%
MFA-TDNN	0.98%
ECAPA-SCNet	1.21%
ECAPA-RepVGG	1.22%

In section 2, we give a brief introduction to several ASV sub-systems and CM sub-systems, we reproduce these system and show the performance in Table 3 and Table 4, respectively.

Comparing Clips system with baseline method, the Clips has better SASV-EER performance than two baseline methods and zhang's method, but the SPF-EER value is a little high than baseline and zhang'method, the reason is that the Clips system has a wrong generalization performance than others. So we take the weighted average of two CM sub-systems, we can see from the table 3, the SPF-EER on the evaluation set is reduced about 6%.

Table 5: *Comparison with SASV Challenge baselines*

System	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
ECAPA-TDNN	1.88	1.63	20.30	30.75	17.38	23.83
AASIST	46.01	49.24	0.07	0.67	15.86	24.38
baseline1	32.88	35.32	0.06	0.67	13.07	19.31
baseline2	12.87	11.48	0.13	0.78	4.85	6.37
Zhang [14]	2.09	2.07	0.07	0.75	1.14	1.58
Clips	1.96	1.77	0.06	0.81	1.01	1.43
Clips-e	**1.96**	**1.75**	**0.06**	**0.76**	**0.98**	**1.36**

4. Conclusions

In this report, we describe our submission system to the SASV challenge 2022, we fix the ASV system and construct four MFM layer to fine-tune the CM system, the proposed fusion system can reduce the SASV-EER from 6.37 % to 1.36 %, which shows the great performance of proposed fusion system.

5. Acknowledgements

This work is supported by the xxx Co. Ltd.

6. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Interspeech 2020*, Oct 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2650

[3] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxceleb speaker recognition challenge 2021 system description," 2021.

[4] T. Liu, R. K. Das, K. A. Lee, and H. Li, "Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," 2022.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *IEEE*, pp. 2818–2826, 2016.

[6] X. Chen and C. Bao, "Phoneme-unit-specific time-delay neural network for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1–1, 2021.

[7] Y. Ma, M. Zhao, Y. Ding, Y. Zheng, M. Liu, and M. Xu, "Rep works in speaker verification," 2021.

[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] 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 e-prints*, 2021.

[10] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics Security*, pp. 1–1, 2015.

[11] L. Zhang, Q. Wang, and L. Xie, "Duality temporal-channel-frequency attention enhanced speaker representation learning," 2021.

[12] G. Huang, Z. Liu, V. Laurens, and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Computer Society*, 2016.

[13] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. Magimai.-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, 2021.

[14] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," 2022.

[15] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *Computer Science*, 2015.

[16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[17] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *arXiv:1911.01601 [cs, eess]*, Jul. 2020, arXiv: 1911.01601. [Online]. Available: http://arxiv.org/abs/1911.01601

[18] C. Veaux, J. Yamagishi, and K. Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.