VTCC System for SASV Challenge 2022

Bao Thang Ta[1], *Tung Lam Nguyen*[1], *Van Hai Do*[1]

[1]Viettel Cyberspace Center, Hanoi, Vietnam

tabaothang97@gmail.com, lamfm95@gmail.com, haidovan@gmail.com

Abstract

Spoofing attacks conducted via logical methods such as voice synthesis, and voice conversion could significantly degrade the performance of a speaker verification system. In recent years, there have been many efforts to develop an integrated speaker verification system which is robust against spoofing attacks. However, there are only a few efforts to design a single model capable of rejecting both utterances spoken by different speakers as well as spoofing utterances. In this work, we proposed a multi-task Conformer with Statistical Pooling model for speaker verification and voice spoofing detection. Our system achieved a SASV EER of 1.8% with the ASVspoof 2019 LA evaluation set.

Index Terms: speaker verification, speaker verification, spoofing speaker detection

1. Introduction

X-Vector [1] and its variants [2, 3] have been state-of-the-art methods for the automatic speaker verification (ASV) problem. These models adopted Time Delay Neural Network (TDNN) layers to capture short and long context features of speakers. Furthermore, to adapt non-fixed length inputs, X-Vector models implemented a pooling layer to aggregate frame-level embeddings into a fixed-length embedding. A statistical pooling is implemented to calculate frame-level features' mean and standard deviation. However, this method assigns equal weight to all frames, ignoring the importance of some special frames. Therefore, some recent studies have added an attention layer to pooling layers. For example, Okabe [4] weighted on the mean and standard deviation of the signal frames. These weights are learned by an attention mechanism. Zhu [4] introduced a pooling mechanism based on self-attention and multiple attention heads. However, a common limitation of these methods is that the weights of the signal frames are scalar. As a result, the elements in each frame have the same weight when calculating the mean and standard deviation, leading to missing essential features. Recently Desplanques [5] proposed an ECAPA-TDNN model using 1D Res2Net blocks with residual connections to improve the previous X-Vector model.

In addition, various methods have been proposed to tackle spoofing speaker detection. Jung [6] proposed a Gaussian Mixture Model (GMM) using six different hand-crafted features. Volkova [7] proposed Light Convolution Neural Network (Light CNN) architecture for detecting different types spoofing attack. Tak [8] proposed an end-to-end anti-spoofing system based on Rawnet2 model. More recently, Jung [9] designed a graph attention network using both spectral and temporal features for this task.

However, there are little-to-no studies to build a multi-task model for both speaker verification and anti-spoofing. Such spoofing aware speaker verification (SASV) system could avoid sub-optimal results, simplify the training process, and further reduce computation cost. Besides, in recent researches, Conformer [10] model has proven remarkably good for automatic speech recognition systems, thanks to its ability to capture local context progressively via a local receptive field layer by layer. Instead of using convolution and self-attention mechanism individually, Conformer combines them together to learn both position-wise local features and content-based global interactions in each utterance. For these reasons, it's worth of investigating its effectiveness in speaker verification and anti-spoofing systems.

In this work, we proposes using Conformer block [10] instead of TDNN layer in the X-Vector model for spoofing aware speaker verification system. The rest of this work is organized as follows. Section 2 introduces our system architecture. Section 3 presents achieved results on the ASVspoof 2019 LA dataset.

2. Proposed System

Our model is presented in Figure 1. We use a statistical pooling layer to aggregate frame-level embeddings into a fixed-length embedding. We use two loss functions. The first one (SV) aims to help model learn speaker embeddings.

$$Loss_{SV} = CrossEntropyLoss(spk\_utts, spk\_ids) \tag{1}$$

The second loss (SASV) aims to help the model distinguish among target utterances and nontarget or spoofing utterances.

$$Loss_{SASV} = CrossEntropyLoss(utt_i, utt_j) \tag{2}$$

The training loss of this model is as follows:

$$Loss = Loss_{SV} + \alpha \times Loss_{SASV}, \tag{3}$$

where α value is a hyperparameter. The obtained embedding before the final fully connected layer is concatenated with two embeddings from pre-trained speaker model ECAPA-TDNN [11] and anti-spoofing model AASIST [11]. The speaker's embedding is averaged over all enrollment utterances.

3. Experimental results

We only use the ASVspoof 2019 LA dataset [12] for training the proposed model. The total parameters of our model is about 15M parameters.

The obtained EER results on the development and evaluation protocol are presented in Table 1 where SV-EER, SPF-EER are SASV-EER are the equal error rate of each model on speaker verification, spoofing detection, and spoofing aware speaker verification tasks, respectively. Our proposal achieved 70% relative SASV improvement over baselines in the Challenge 2022 on the evaluation set.

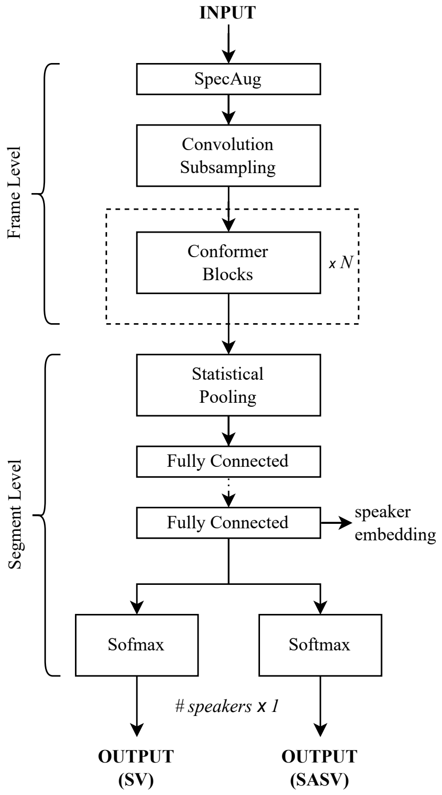

Figure 1: *Multi-task conformer-based speaker embedding network.*

Table 1: *Compare EER (%) of our model with other baselines in the SASV Challenge 2022[11]*

Model	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
ECAPA-TDNN	1.88	1.63	20.03	30.75	17.38	23.83
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
Our model	2.91	2.42	0.07	0.97	**1.82**	**1.86**

4. Conclusion

This paper proposed a multi-task Conformer-based speaker embedding network for join optimizing speaker verification and spoofing detection tasks. The obtained results are very promising and competitive compared to previous works.

5. References

[1] 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.

[2] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” *arXiv preprint arXiv:1803.10963*, 2018.

[3] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, “Speaker recognition for multi-speaker conversations using x-vectors,” in *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[4] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, “Self-attentive speaker embeddings for text-independent speaker verification.” in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[5] 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.

[6] B. Chettri and B. L. Sturm, “A deeper look at gaussian mixture model based anti-spoofing systems,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5159–5163.

[7] M. Volkova, T. Andzhukaev, G. Lavrentyeva, S. Novoselov, and A. Kozlov, “Light cnn architecture enhancement for different types spoofing attack detection,” in *International Conference on Speech and Computer*. Springer, 2019, pp. 520–529.

[8] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with rawnet2,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[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 preprint arXiv:2110.01200*, 2021.

[10] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[11] 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.

[12] 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.