EVA-ASCA: Enhancing Voice Anti-Spoofing through Attention-based Similarity Weights and Contrastive Negative Attractors

Nghi Tran, Bima Prihasto, Phuong Thi Le, Thao Tran
National Central University, Taiwan, ROC

Chun-Shien Lu
Academia Sinica, Taiwan, ROC
lcs@iis.sinica.edu.tw

Jia-Ching Wang
National Central University

***Abstract*—Voice spoofing attacks pose an escalating security concern within the contemporary digital landscape. Attackers employ techniques such as voice conversion (VC) and text-to-speech (TTS) to generate a synthetic speech that replicates the victim's voice, to illicitly access sensitive data. Detection of these attacks hinges on identifying anomalies in audio transmission resulting from these deceptive activities. Anomalies arise from encoding and transmission conditions that are not commonly encountered, particularly in situations such as local authentication or telephony. To address this issue, our study presents a strategy featuring pivotal enhancement: Attention-based Similarity Weights and Contrastive Negative Attractors. This technique clusters authentic speeches around multiple speaker attractors together in a high-dimensional embedding space, effectively thwarting spoofing attacks across all attractors. Experimental results substantiate the superiority of our system, yielding a substantial 1.09% improvement in the equal error rate (EER) when compared to existing solutions on the ASVspoof 2019 LA evaluation dataset.**

***Index Terms*—Automatic Speaker Verification, Security, Spoofing Attacks, Voice Anti-Spoofing**

I. Introduction

Automatic speaker verification (ASV) systems have been widely employed for voice-based authentication in various applications, such as banking, security, and law enforcement [1], [2]. One of the most challenging types of spoofing attacks against ASV systems is logical access (LA) attacks that use synthetic speeches [3]. LA attacks can be launched with various levels of knowledge and resources, making them a serious threat to ASV security. Therefore, developing effective spoofing countermeasures (CM) systems against LA attacks is crucial for ensuring the security of ASV systems.

Recent researches in speech anti-spoofing have focused on exploring various techniques for extracting embeddings from speech signals. Traditionally, hand-crafted features such as linear frequency cepstral coefficients (LFCC) [4] and constant-Q cepstral coefficients (CQCC) [5], [6] have been widely used for this purpose. However, these methods have limitations in terms of capturing the complex and subtle variations in speech signals that are indicative of spoofing attacks. To overcome these limitations, there has been a tendency towards developing end-to-end models that use raw waveforms as input [7], [8]. These models leverage deep neural network architectures to learn discriminative features directly from the speech signal and have shown remarkable performance, achieving state-of-the-art results on benchmark datasets such as ASVspoof 2019.

Another line of researches that has been explored investigates the training strategies such as data augmentation [9], [10] and multi-task learning [11]. The main challenge of these methods is the capacity to generalize to unseen attacks, specifically spoofing attacks that were not utilized in generating training data [3]. To address this generalization problem, Zhang *et al.* [12] introduced a novel approach, called one-class learning, to learn a decision boundary that separates genuine speeches from spoofed speeches rather than trying to classify different types of spoofing attacks. Nevertheless, compacting these clusters into a single one may cause the misclassification of spoofing attacks [13]. In this context, it is important to explore alternative clustering strategies that can improve the accuracy and robustness of speaker verification systems.

Recently, Ding *et al.* [14] proposed a promising method, called Speaker Attractor Multi-Center One-Class Learning (SAMO), to address speaker diversity and generalization ability in one-class learning. However, the SAMO approach faces challenges in handling diverse speech patterns and characteristics. Its reliance on one-class learning limits adaptability to various speaker traits and new voices, potentially causing misclassifications. Additionally, closely resembling spoofing attacks may yield false positives. Addressing these issues is crucial for refining and enhancing SAMO, ensuring its robustness and reliability in real-world scenarios.

We propose a novel method, "Enhancing Voice AntiSpoofing through Attention-based Similarity Weights and Contrastive Negative Attractors" (EVA-ASCA), to address the limitations mentioned above. This method aims to enhance speech spoofing systems by integrating adaptive speaker attractors, which can better handle the presence of multiple speakers, and contrastive negative samples, which mitigate the risk of misclassification due to similar voice characteristics, lead to improved classification accuracy. Our method has the potential to advance the field of anti-spoofing and contribute to the development of more accurate and reliable systems for speaker verification and authentication.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00106

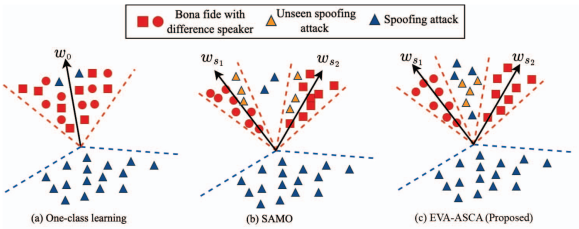

Fig. 1. Comparison between previous studies and our proposed method.

II. Methodology

A. Overview: EVA-ASCA

Our proposed method (EVA-ASCA) aims to enhance speech spoofing systems by utilizing attention-based similarity weights and contrastive negative attractors. The attention contrastive loss function module is employed to optimize the classification model in that it is supported by attention-based similarity weights that emphasize the most discriminative features for accurate speaker verification. In addition, the contrastive negative attractor contributes to robustness by accentuating the importance of separating distinct speakers within the embedding space.

Fig. 1 illustrates a comparison between our proposed method and previous approaches. Our method (Fig. 1(c)) is able to effectively separate unseen spoofing attacks from the bona fide group, whereas the one-class learning approach (Fig. 1(a)) and SAMO (Fig. 1(b)) are insufficient in addressing variations in both bona fide and spoofing instances. Here, w_0 refers to the optimization direction of target class embeddings, while w_{s_1} and w_{s_2} represent the average embeddings of the enrolled utterances of corresponding speakers.

Algorithm 1 describes the pseudo code of module - “Attention-based Similarity Weights and Contrastive Negative Attractor Loss Computation”. The details will be elaborated in the following.

B. Attention-based Similarity Weights

Our anti-spoofing system introduces the novel “Attention-based Similarity Weights” mechanism, enhancing the speaker attractor’s function. This attention mechanism, applied as similarity weights, consolidates bona fide speeches in the high-dimensional embedding space. Varying weights assigned to different features effectively deflect spoofing attacks, reinforcing Automatic Speaker Verification (ASV) against advanced synthesis technologies. The dynamic attention mechanism adjusts weights based on processed speech features, enabling adaptive learning to new speaker characteristics and bolstering resilience against diverse attacks. Furthermore, attention-based similarity weights play a pivotal role in clustering genuine speech around multiple attractors, a notable improvement over prior methods lacking speaker diversity consideration.

In the speaker verification tasks, not all features in the feature space have equal importance for discriminating between different speakers. The attention mechanism aims to assign different levels of importance or weights to these features based on their significance. The detailed steps are as follows.

Algorithm 1 Attention-based Similarity Weights and Contrastive Negative Attractor Loss Computation

```
Input: $x, spk, enroll$
Output: $final\_scores, contrastive\_loss$
$w \leftarrow$ initialize the center weights
$scores \leftarrow x@w.T$ (1)
if $attractor = 1$ then
    $att\_weights \leftarrow$ zeros
    for $idx$ in $spk$ do
        if $spk[idx]$ is in $enroll$ then
            $att\_weights[idx] \leftarrow$ Similarity($x[idx]$, $enroll[spk[idx]]$) (2)
        end if
    end for
    $att\_weights \leftarrow$ Softmax($att\_weights$) (3)
    $final\_scores \leftarrow att\_weights \times scores$ (4)
    $neg\_indices \leftarrow$ Random selection from $w$ (5)
    $neg\_scores \leftarrow x@w[neg\_indices]^T$ (6)
    $contrastive\_loss \leftarrow$ BCE_With_Logits($scores$, $neg\_scores$) (7)
end if
return $final\_scores, contrastive\_loss$
```

1) **Initialization**: Establish the initial score value and the attention weights of individual sample n_i as:

$$\begin{aligned} \text{scores}[i] &= x_i@w.T \\ \text{att\_weights}[i] &= 0, \end{aligned} \tag{1}$$

where x_i is the feature vector of sample n_i and w is the center weights.

2) **Weight Calculation**: For each sample n_i, if its corresponding speaker (‘spk[idx]’) is present in the enrollment list (‘enroll’), a similarity measure is computed between the sample and enrolled speaker’s feature as:

$$s_i = \frac{x_i}{\|x_i\|_2} \cdot \frac{\text{enroll[spk[idx]]}}{\|\text{enroll[spk[idx]]}\|_2}, \tag{2}$$

which will act as the attention weight for that sample.

3) **Normalization**: The attention weights across all samples are then normalized using the Softmax function as:

$$\text{att\_weights}[i] = \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}}, \tag{3}$$

where N is the total number of samples. This ensures that the weights sum to one and emphasizes the weights of more relevant features.

4) **Final Score Computation**: The final scores are the product of the attention weights and original scores:

$$\text{final\_scores}[i] = \text{att\_weights}[i] \times \text{scores}[i]. \tag{4}$$

This gives more importance to the features deemed more relevant by the attention mechanism.

C. Contrastive Negative Attractors

The second enhancement of our anti-spoofing system, Contrastive Negative Attractors, strengthens the classification function against disruptive samples in the high-dimensional embedding space. This feature acts as a counterbalance to spoofing attacks, offering a novel perspective on handling speaker diversity. Our approach with negative attractors considers diverse speakers, tailoring its classification for improved performance. Delving deeper, Contrastive Negative Attractors maintain classification integrity by creating stark contrasts in the embedding space, facilitating quick and accurate classification of spoofed speeches. Working in tandem with attention-based similarity weights, they form a dual-layered defense against spoofing attacks.

Contrastive learning aims to learn embeddings such that the positive pairs (*e.g.*, from the same speaker) come closer in the embedding space, while the negative pairs (*e.g.*, from different speakers) move apart. The details are described as follows.

1) **Negative Sampling**: For each sample in the batch, a random negative attractor is chosen from the center weights w. Specifically, the random selection for the i-th sample is denoted as:

$$w_{\text{neg},i} = w[\text{random\_index}(0, |w| - 1)], \tag{5}$$

where $|w|$ is the number of available center weights.

2) **Negative Score Computation**: A score, denoted as "neg_scores," is computed between the sample i and negative attractor $w_{\text{neg},i}$ using the dot product of their normalized vectors:

$$\text{neg\_scores}[i] = \frac{x_i}{\|x_i\|_2} \cdot \frac{w_{\text{neg},i}}{\|w_{\text{neg},i}\|_2}, \tag{6}$$

as their similarity, and is expected to be minimized for effective contrastive learning.

3) **Loss Computation**: Using binary cross-entropy as the logits function, the loss is calculated from the original scores and negative scores as:

$$\begin{gathered} L = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log(\sigma(\text{scores}[i])) \\ +(1 - y_i) \log(1 - \sigma(\text{neg\_scores}[i])), \end{gathered} \tag{7}$$

where σ is the sigmoid function and y_i denotes the ground truth labels (1 for positive pairs and 0 for negative pairs). This loss function pushes the embeddings to cluster positive pairs closely and pushes negative pairs farther apart in the embedding space.

III. Experiments

A. Experimental setup

Dataset. In this study, the performance of our proposed method was evaluated by comparing it with other novel approaches on the ASVSpoof 2019 LA evaluation datasets [15]. The LA subset of the dataset is especially noteworthy, as it includes both genuine speeches and various types of spoofing attacks, including text-to-speech (TTS) and voice conversion (VC) attacks.

TABLE I
Comparison of CM Systems.

CM System	EER (%)	min-tDCF
Baseline 1 (AASIST) [16]	1.25	0.042
Baseline 2 (SAMO) [14]	2.17	0.064
Capsule Network [17]	1.97	0.050
LCNN-DA [18]	2.76	0.077
SE-Res2Net50 [19]	2.86	0.060
RawNet2 [20]	3.50	0.090
STATNet [21]	2.45	0.062
Ours	1.09	0.033

Evaluation metric. To verify the performance of anti-spoofing, we employed the output score, which is referred to as the Countermeasure (CM) score within the context of an anti-spoofing system. We used Equal error rate (EER) and the minimum tandem detection cost function (min t-DCF), which are commonly adopted in speaker verification systems, as performance evaluation metrics.

Methods for comparsion. SAMO [14] and AASIST [16] were adopted as the baselines. To ensure a fair assessment, we compared the results under the same training conditions. Both the baselines and our model were trained for 100 epochs, utilizing a learning rate of 0.0001 with the cyclic learning rate schedule to optimize the training process for a balanced convergence rate and overall stability.

B. Experimental results

Comparison with state-of-the-art methods. Table I presents a comparative result of various Countermeasure (CM) systems based on EER and min-tDCF. Our proposed system demonstrates better performance, with the lowest EER at 1.09 and the lowest min-tDCF at 0.033. This performance surpasses those of both Baseline 1 (AASIST) [16] with an EER of 1.25 and min-tDCF of 0.042, and Baseline 2 (SAMO) [14] with an EER of 2.17 and min-tDCF of 0.064. While Capsule Network [17] performs competitively among the other systems, it still exhibits higher values for both EER and min-tDCF.

Various attention parameter adjustment. Table II shows the performance of proposed model across different attention parameter values in terms of α. A smaller α results in a more uniform attention distribution, while a larger α prioritizes samples with higher similarities. Given a sample x_i and a weight w, the attention weight A_i before softmax activation is calculated as:

$$A_i = \alpha \times \text{similarity}(x_i, w). \tag{8}$$

TABLE II
PERFORMANCE OF OUR METHOD UNDER VARIOUS ATTENTION PARAMETER α.

Attention value	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
EER (%)	1.09	1.17	1.21	1.46
min-tDCF	0.033	0.038	0.035	0.047

Upon scaling by α, the attention weights are normalized using the softmax function as:

$$\text{att\_weights}_i = \frac{\exp(A_i)}{\sum_j \exp(A_j)}. \tag{9}$$

The results illustrate the impact of varying α's on the performance of our CM System. At α = 0.01, the system achieves the lowest EER (1.09%), suggesting that a subtle attention mechanism promotes better generalization. As α increases to 0.1 and 0.5, EER rises slightly, indicating a potential trade-off between attention prominence and performance. However, at α = 1, EER spikes (1.46%), signaling overfitting. In conclusion, the parameter α significantly influences the efficiency of attention mechanism, underscoring the importance of careful tuning in that extreme values can lead to suboptimal results in the context of a specific problem and dataset.

Ablation studies. Table III compares countermeasure systems based on EER and min-tDCF with $\alpha = 1$. The baseline system, "SAMO without Enrollment," has an EER of 2.17%. "Proposed 1" enhances this with "Contrastive," reducing EER to 2.16% and significantly improving min-tDCF to 0.050. "Proposed 2," which combines "Contrastive" and "Attention Mechanism," achieves a remarkable EER of 1.46% and the best min-tDCF of 0.047. Thus, it is concluded that combining both contrastive learning and attention mechanism together demonstrate notable improvements over the baseline, highlighting its potential to enhance countermeasure.

IV. CONCLUSION

We present an enhanced strategy that incorporates attention-based similarity weights and contrastive negative attractors to address the escalating threat of voice-spoofing attacks. This approach bolsters the reliability of spoofing detection even in the face of unknown encoding and transmission conditions. Experimental results underscore the superiority of our system, showcasing a substantial improvement over existing solutions on the ASVspoof 2019 LA evaluation dataset.

V. ACKNOWLEDGEMENT

This work was supported by the National Science and Technology Council, Taiwan, ROC, under Grants NSTC 112-2221-E-001-011-MY2 and 112-2634-F-001-002-MBK.

TABLE III
COMPARISON OF CM SYSTEMS. BASELINE IS SAMO WITHOUT ENROLLMENT.

CM System	**EER (%)**	**min-tDCF**
Baseline	2.17	0.070
Proposed 1 (Contrastive)	2.16	0.050
Proposed 2 (Contrastive + Attention)	**1.46**	**0.047**

REFERENCES

[1] M. D. Aakshi Mittal, "Automatic speaker verification systems and spoof detection techniques: review and analysis," in *International Journal of Speech Technology*, pp. 105–134, IEEE, 2022.

[2] L. D. Fieke Jansen, Javier Sánchez-Monedero, "Biometric identity systems in law enforcement and the politics of (voice) recognition: The case of siip," in *Big Data Society*, vol. 8, IEEE, 2021.

[3] M. Sahidullah and et al., "Introduction to voice presentation attack detection and recent advances," in *Handbook of Biometric Anti-Spoofing, 2nd Ed.*, 2019.

[4] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Interspeech*, 2015.

[5] M. Todisco, H. Delgado, and N. W. D. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *The Speaker and Language Recognition Workshop*, 2016.

[6] M. Todisco, H. Delgado, and N. W. D. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

[7] H. Tak and et al., "End-to-end anti-spoofing with rawnet2," *ICASSP*, pp. 6369–6373, 2020.

[8] G. Hua, A. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[9] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, "A study on data augmentation in voice anti-spoofing," *Speech Commun.*, vol. 141, pp. 56–67, 2021.

[10] H. Tak, M. R. Kamble, J. Patino, M. Todisco, and N. W. D. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," *ICASSP*, pp. 6382–6386, 2021.

[11] Y. Mo and S. Wang, "Multi-task learning improves synthetic speech detection," *ICASSP*, pp. 6392–6396, 2022.

[12] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2020.

[13] Z. Ghafoori and C. Leckie, "Deep multi-sphere support vector data description," in *SDM*, 2020.

[14] S. Ding, Y. Zhang, and Z. Duan, "Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *ICASSP*, pp. 1–5, IEEE, 2023.

[15] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. W. D. Evans, M. Sahidullah, V. Vestman, T. H. Kinnunen, K.-A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, and Z. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, p. 101114, 2019.

[16] J.-w. Jung and et al., "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP*, pp. 6367–6371, IEEE, 2022.

[17] A. e. a. Luo, "A capsule network based approach for detection of audio spoofing attacks," in *ICASSP*, pp. 6359–6363, IEEE, 2021.

[18] X. Ma and et al., "Improved lightcnn with attention modules for asv spoofing detection," in *ICME*, pp. 1–6, IEEE, 2021.

[19] X. Li and et al., "Replay and synthetic speech detection with res2net architecture," in *ICASSP*, pp. 6354–6358, IEEE, 2021.

[20] H. Tak and et al., "End-to-end anti-spoofing with rawnet2," in *ICASSP*, pp. 6369–6373, IEEE, 2021.

[21] R. Ranjan, M. Vatsa, and R. Singh, "Statnet: Spectral and temporal features based multi-task network for audio spoofing detection," in *IJCB*, pp. 1–9, IEEE, 2022.