End-to-end framework for spoof-aware speaker verification

Woo Hyun Kang, Jahangir Alam, Abderrahim Fathan
Computer Research Institute of Montreal (CRIM)
woohyun.kang, jahangir.alam, abderrahim.fathan@crim.ca

Abstract
In this paper, we propose a novel end-to-end framework for training a spoof-aware speaker verification (SASV) system. To match the SASV scenario, where the test samples may be either spoof or genuine, we propose a novel contrastive objective and a modified mixup regularization strategy. The proposed end-to-end system and other SASV systems were evaluated on the ASVSpoof2019 LA evaluation set according to the SASV 2022 challenge rules. Our results demonstrate that the proposed framework can learn complementary information to the conventional embedding fusion-based SASV systems. Using the proposed system in conjunction with the conventional embedding fusion systems has achieved a relative improvement of 61.07% in terms of SASV-EER compared to the best performing baseline result provided by the challenge organizers.

Index Terms: spoof-aware, speaker verification, SASV 2022 challenge, ASVSpoof2019, logical access

1. Introduction
In the past decade, attributed to the widespread deployment of smart devices, automatic speaker verification (ASV) has become an essential technology for user authentication. As the speech synthesis and voice conversion techniques have improved dramatically in recent years, spoof detection, also known as anti-spoofing, has become a key technology for protecting the ASV systems from spoof attacks generated by speech synthesis and voice conversion algorithms.

Conventionally, most research focused on training a stand-alone spoof countermeasure (CM) system, which is trained with no consideration of the ASV system. Recent trends in stand-alone voice anti-spoofing is to employ deep learning architectures in an end-to-end fashion on the top of raw signal or hand-crafted features to discriminate between bonafide and spoof speech signals [1–9]. Although these stand-alone spoof detection systems have shown promising performance in terms of the CM task, not much exploration was done in terms of the joint optimization of the ASV and CM systems. Moreover, since these stand-alone CM systems are mostly evaluated in tandem with a fixed or no ASV system [10, 11], their performance is not guaranteed when deployed to a new ASV system. Therefore, it is important to jointly optimize the CM and ASV submodules to achieve reliable performance in a realistic scenario.

The spoof-aware speaker verification (SASV) challenge [12] aims to tackle this issue by providing a standard benchmark for evaluating ASV systems on a scenario where a spoof attack is present. More specifically, the main objective of the SASV challenge is to develop a speaker verification system robust to both zero-effort impostor access attempts and spoof attacks. Thus unlike the previous benchmarks for CM systems (e.g., ASVSpoof [10, 11]) which mainly concerns the CM performance, the SASV challenge measures the ASV and CM performance jointly. The challenge organizers suggested two possible approaches for solving this problem: ensembled ASV and CM system, and integrated single system. The ensembled ASV and CM method is a straightforward approach for achieving the objective, and experimental results reported by the challenge organizers show that it can yield promising performance. However, since the pre-training of the ASV and CM are carried out independently, its performance in a SASV scenario may be limited. On the other hand, an integrated single system method can potentially overcome this limitation as it optimizes the system with both ASV and CM objectives. Nonetheless, training with multiple objectives is known to require careful planning for optimal performance [13]. Especially when training the integrated system for two disjoint tasks (e.g., ASV, CM), the system may suffer from negative transfer [14] and task imbalance [15] problems, which can degrade the overall performance.

In light of this, we propose an end-to-end approach for training an integrated system for SASV with a single objective function. More specifically, we formulated a contrastive loss function that can reflect both the CM and ASV tasks in a SASV scenario. Furthermore, in order to compensate for the low number of unique speakers within the SASV training set, we have proposed a spoof-aware mixup regularization strategy. To evaluate the performance of the proposed framework, we conducted a set of experiments using the ASVSpoof2019 logical access (LA) dataset, which is the core training and evaluation set for the SASV 2022 Challenge. Our experimental results showed that the proposed method can provide promising performance and a significant improvement was observed when used in conjunction with the ensemble ASV and CM system.

2. Background

2.1. SASV Challenge 2022
The spoof-aware speaker verification (SASV) challenge provides a common and standard benchmark for evaluating speaker verification systems on a spoof attack scenario. More specifically, the SASV challenge aims to develop a new solution for speaker verification robust to both zero-effort impostor access attempts and spoofing attacks.

For training the systems, the SASV challenge allows using the VoxCeleb2 [16], ASVSpoof2019 train and development sets [10]. The evaluation is done on the ASVSpoof 2019 LA ASV protocol [10], where three equal error rates are measured:

- SASV-EER: the primary metric for the challenge. For computing the SASV-EER, bonafide utterances from the target speaker are considered as positive and the rest are considered as negative.
- SPF-EER: the CM metric. For computing the SPF-EER, bonafide utterances from the target speaker are considered as positive and all spoof utterances are considered as negative.
- SV-EER: the ASV metric. For computing the SV-EER, bonafide utterances from the target speaker are consid-
2.2. Angular prototypical objective

The angular prototypical loss function \([20, 21]\) is a contrastive objective function widely adopted to train end-to-end ASV systems. Given a batch of prototype embedding vectors \(\omega^1\) and query embeddings \(\omega^2\), where \(\omega^i\) indicates the embedding extracted from the \(k^{th}\) utterance of the \(i^{th}\) speaker, the angular prototypical function is defined as follows:

\[
L_{AP} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\cos(\omega^i, \omega^2)\right)}{\sum_{j=1}^{N} \exp(\cos(\omega^i, \omega^j))},
\]

where \(\cos\) represents the cosine similarity operation. Equation 1 can be interpreted as the cross-entropy loss which aims to maximize the similarity between the embeddings extracted from the same speaker, while minimizing the similarity between different speaker embeddings.

2.3. Mixup regularization strategy

The mixup is a data-driven augmentation strategy for improving the generalization of the learned representation \([22, 23]\). For arbitrary objective function \(L_{pair}(x, y)\), where \(x\) is the input sample and \(y\) is the corresponding label, given two data instances \((x_i, y_i)\) and \((x_j, y_j)\), the mixup loss is defined as follows:

\[
L_{mix}((x, y), (x_j, y_j)) = L_{pair}(\lambda x_i + (1 - \lambda)x_j, \lambda y_i + (1 - \lambda)y_j),
\]

where \(\lambda\) is a mixing coefficient sampled from a beta distribution \(~ Beta(\alpha, \alpha)\). Essentially, the mixup generates a synthetic training sample \(\lambda x_i + (1 - \lambda)x_j\) with label \(\lambda y_i + (1 - \lambda)y_j\). Therefore, the system can benefit from the mixup strategy in terms of generalization, as the augmented training data contains samples and classes not included in the original training set.

3. End-to-end spoof-aware speaker verification system

In this section, we describe our proposed end-to-end framework for training a spoof-aware speaker verification system. Instead of training the CM and ASV system separately, given a pre-trained ECAPA-TDNN system trained for ASV, the proposed framework fine-tunes the network using a CM- and ASV-aware
Table 2: SASV-EER [%] performance comparison between the ECAPA-TDNN systems trained with and without the proposed spoof-aware mixup strategy on the ASVspoof2019 evaluation set.

<table>
<thead>
<tr>
<th>ECAPA-TDNN systems</th>
<th>SASV-EER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without spoof-aware mixup</td>
<td>7.11</td>
</tr>
<tr>
<td>With spoof-aware mixup</td>
<td>6.83</td>
</tr>
</tbody>
</table>

3.2. Spoof-aware mixup regularization strategy

It is well known that the performance of the contrastive objectives greatly varies depending on the number of unique speakers within the mini-batch [21]. This often not a crucial problem in most ASV tasks with large-scale datasets (e.g., VoxCeleb2), but in a SASV scenario, the number of speakers in the training set may be limited. Especially in the ASVspoof2019 LA dataset, which is the core training set for the SASV Challenge, only 20 speakers are included in the training set.

To circumvent this problem, we have adopted the mixup regularization strategy described in Equation 2. More specifically, unlike the standard mixup strategy which interpolates randomly sampled labels, we modify the mixup strategy to match the SASV scenario. Let us say that we have prototype utterances $x^1_c$ and query utterances $x^2_j$, where $x^1_c$ indicates the $c^{th}$ utterance from speaker $y_c$ with CM label $c_i \sim \{\text{spoof}, \text{bonafide}\}$. Given two pairs of utterances $(c^1_i, c^2_j)$ and $(c^1_i, c^2_j)$, we propose to interpolate both the prototypes and the queries as follows:

$$x^1_{\text{mixup}} = \lambda x^1_c + (1-\lambda) x^1_c, \quad \lambda x^1_c = (1-\lambda)x^1_c + \lambda x^1_c,$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ is the mixing coefficient. The speaker labels are also interpolated with the same coefficient:

$$y_{\text{mixup}} = \lambda y_i + (1-\lambda)y_j.$$

On the other hand, the mixed CM labels are not created via interpolation. Since the spoof artifacts are likely to be preserved during the linear interpolation process, we propose to consider the synthetic samples created by mixing with spoof samples as spoof:

$$c_{\text{mixup}} = \begin{cases} 
\text{bonafide} & \text{if } c_i = c_j = \text{bonafide} \\
\text{spoof} & \text{otherwise}.
\end{cases}$$

The general labeling scheme of the proposed spoof-aware mixup strategy is depicted in Figure 2.

4. Experiments

4.1. Experimental setup

As local frame-level hand-crafted features we use 60-dimensional (including the delta and double delta coefficients) linear frequency cepstral coefficients (LFCC) extracted using 25ms analysis window over a frame shift of 10ms. No offline data augmentation was performed in our experiments.

According to the SASV 2022 Challenge rule, for training and evaluating the experimented systems, the ASVspoof 2019 LA dataset was used, which provides a common framework for conducting spoofing detection research on LA attacks. In our experiments, the development set was used for validation and score normalization. For more details about the corpora, the interested readers are referred to [10].

For the baseline embedding fusion systems, not only we have used the pre-trained ECAPA-TDNN-based ASV [18] and AASIST CM [19] systems provided by the organizers, we have also trained and experimented with a hybrid neural network (HNN)-based ASV system [24, 25] and a TDNN with higher-order statistics (TDNNHOSP)-based CM [26] system. As in the baseline system, the HNN and TDNNHOSP were trained on VoxCeleb2 and ASVspoof2019 LA train set, respectively.
Table 3: The experimental results of the SASV systems on the ASVSpoof2019 Logical Access development and evaluation sets in terms of official evaluation metrics.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SASV-EER [%]</td>
<td>CM-EER [%]</td>
</tr>
<tr>
<td>Baseline ASV [12]</td>
<td>17.38</td>
<td>20.80</td>
</tr>
<tr>
<td>Baseline ASV [12]</td>
<td>4.85</td>
<td>0.13</td>
</tr>
<tr>
<td>1 ECAPA-TDNN-ASV + ASIST-CM</td>
<td>4.85</td>
<td>0.13</td>
</tr>
<tr>
<td>2 HNN-ASV + TDNNHOSP-CM</td>
<td>4.60</td>
<td>0.32</td>
</tr>
<tr>
<td>3 Proposed end-to-end SASV system</td>
<td>3.36</td>
<td>1.28</td>
</tr>
<tr>
<td>Scaled Average of #1 and #3</td>
<td>2.63</td>
<td>0.29</td>
</tr>
<tr>
<td>Scaled Average of #1 and #2</td>
<td>3.59</td>
<td>0.05</td>
</tr>
<tr>
<td>Scaled Average of #2 and #3</td>
<td>2.83</td>
<td>0.27</td>
</tr>
<tr>
<td>Scaled Average of #1, #2 and #3</td>
<td>2.43</td>
<td>0.07</td>
</tr>
</tbody>
</table>

4.2. Experimental results

4.2.1. Effect of the proposed spoof-aware contrastive speaker verification objective

In this experiment, we compare the effect of different within-bonafide mixup strategies for training the spoof detection system. As shown in Table 1, fine-tuning the ASV system with the proposed spoof-aware angular prototypical loss can greatly improve the performance. The performance was further enhanced with spoof-prototype trimming (Equation 4), which achieved a relative improvement of 12.53% in terms of SASV-EER.

4.2.2. Effect of the proposed spoof-aware mixup regularization strategy

In this experiment, we compare the performance of the end-to-end system trained with and without the proposed spoof-aware mixup strategy. As shown in Table 2, using the proposed spoof-aware mixup strategy was found to be beneficial, which achieved a relative improvement of 3.94% in terms of SASV-EER.

4.2.3. Spoof-aware speaker verification performance comparison between different systems

In this experiment, we compare the performance of different SASV systems on the ASVSpoof2019 LA development and evaluation sets. As depicted in Table 3, while the proposed end-to-end system (i.e., System #3) have shown comparable SASV-EER performance to the embedding fusion schemes (i.e., System #1, #2), they showed very different behaviors in terms of CM and ASV performances. For example, the embedding fusion schemes generally showed very low CM-EER and high ASV-EER. Meanwhile, the proposed end-to-end system yielded the opposite tendency, where the CM-EER was high and ASV-EER was much lower. From these observations, we can assume that there may exist complementary information between the embedding fusion framework and the proposed end-to-end method.

In order to exploit the complementarity among different systems, we have performed score-level fusion where each system's scores were normalized using the mean and standard deviations of the development scores, and the average of the normalized scores from different systems was computed. Score-level fusion has improved the SASV-EER performance in all cases, but it could be seen that the improvement via combining scores from different embedding fusion systems was relatively small (i.e., Scaled Average of #1 and #2). On the other hand, the score-level fusion between the proposed end-to-end system and embedding fusion systems was found to be very effective (i.e., Scaled Average of #1 and #3, Scaled Average of #2 and #3). This clearly indicates that there exists complementary information between the proposed end-to-end framework and the conventional embedding fusion scheme.

The best performance was achieved by score-level fusing the proposed end-to-end system with two different embedding fusion schemes (i.e., Scaled Average of #1, #2 and #3), which resulted a relative improvement of 61.07% in terms of SASV-EER compared to the best performing baseline system (i.e., Baseline SASV).

5. Conclusion

In this work, we proposed a novel end-to-end framework for training a spoof-aware speaker verification (SASV) system. More specifically, we introduced a spoof-aware contrastive speaker verification loss function and a modified mixup strategy for the SASV scenario.

The performance of the proposed end-to-end system and other SASV systems were evaluated on the ASVSpoof2019 LA evaluation set by following the SASV 2022 challenge rules. Experimental results depicted that the proposed end-to-end framework can learn complementary information to the conventional embedding fusion-based SASV method. Score level fusion of the proposed system and the conventional embedding fusion systems yielded a relative improvement of 61.07% in terms of SASV-EER compared to the best performing baseline result provided by the challenge organizers.

6. Acknowledgement

Authors wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2019-05381 and Ministry of Economy and innovation (MEI) of the Government of Quebec for the continued support.
7. References


