The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

A Probabilistic Fusion Framework for Spoofing Aware Speaker Verification

You Zhang, Ge Zhu, Zhiyao Duan

Audio Information Research Lab, University of Rochester, Rochester, NY, USA
you.zhang@rochester.edu

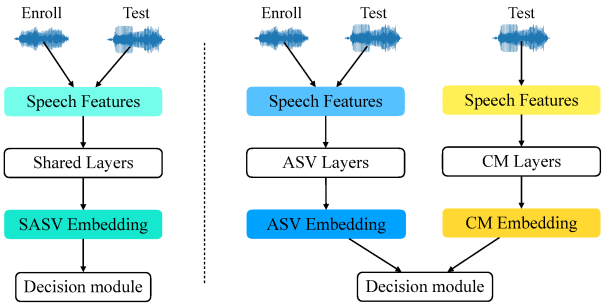
Abstract

The performance of automatic speaker verification (ASV) systems could be degraded by voice spoofing attacks. Most existing works aimed to develop standalone spoofing countermeasure (CM) systems. Relatively little work targeted at developing an integrated spoofing aware speaker verification (SASV) system. In the recent SASV challenge, the organizers encourage the development of such integration by releasing official protocols and baselines. In this paper, we build a probabilistic framework for fusing the ASV and CM subsystem scores. We further propose fusion strategies for direct inference and fine-tuning to predict the SASV score based on the framework. Surprisingly, these strategies significantly improve the SASV equal error rate (EER) from 19.31% of the baseline to 1.53% on the official evaluation trials of the SASV challenge. We verify the effectiveness of our proposed components through ablation studies and provide insights with score distribution analysis.

1. Introduction

Automatic speaker verification (ASV) aims to verify the identity of the target speaker given a test speech utterance. A typical speaker verification process involves two stages: First, a few utterances of the speaker are enrolled, then the identity information extracted from the test utterance is compared with that of the enrolled utterances for verification [1]. ASV researchers have been developing speaker embedding extraction methods [2, 3, 4] to encode speaker identity information for verification. However, it is likely that the test utterance is not human natural speech but *spoofing attacks* that try to deceive the ASV system. Spoofing attacks usually include impersonation, replay, text-to-speech, voice conversion attacks. Studies have shown that ASV systems are vulnerable to spoofing attacks [5].

In recent years, researchers have been developing spoofing countermeasure (CM) and audio deepfake detection systems to detect spoofing attacks. With the ASVspoof 2019 challenge which provides a large-scale standard dataset and evaluation metrics, the CM systems have been improved in various aspects, especially on the generalization ability [6, 7, 8] and channel robustness [9, 10, 11] for in-the-wild applications. However, all of the above works focused on the evaluation of standalone CM systems. Intuitively, an imperfect CM system would accept spoofing attacks but reject bona fide speech from the target person [12]. After all, the ultimate goal of developing a CM system is to protect the ASV system from falsely accepting spoofing attacks. However, how an improved CM system benefits the ASV system is not clear. Although the minimum t-DCF [13] used in the ASVspoof challenge [14] evaluates the reliability of CM systems to ASV systems, it is calculated on a fixed ASV system provided by the ASVspoof organizers instead of being adapted to the ASV system at hand. For better protecting the ASV system from being spoofed and maintaining its discrimination ability on speaker identity, the CM component should be jointly optimized with the ASV system. As a result, an integrated ASV and CM system is promising.



(a) Multi-task (b) Fusion

Figure 1: Illustration of two categories of methods in the literature of SASV systems. The "layers" represent different machine learning models aiming to extract embeddings such as i-vectors. The "decision module" could be (i) a layer for computing the final score on the SASV embedding, or (ii) a score fusion module that fuses ASV and CM scores.

Relatively little attention is paid to improving the integration of ASV and CM systems. As reviewed in Section 2, some work has 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 [15] has been held to further encourage the study of integrated systems of ASV and CM. In this challenge, only cases of logical access (LA) spoofing attacks, i.e., TTS and VC attacks, are taken into consideration. The test utterances of the SASV system can be categorized into three classes: *target*—bona fide speech belonging to the target person, *non-target*—bona fide speech but not belonging to the target speaker, and *spoof*—spoofing attacks.

In this work, we formulate a fusion-based SASV system under the probabilistic framework on top of the ASV and CM subsystems. We also propose a fine-tuning strategy on the integrated system for further improvement. With the proposed fusion strategies, we outperform the SASV baseline systems by a large margin. Our best performing system achieved 1.53% SASV-EER on the official evaluation trials. We also provide an ablation study and score distribution analysis for future study.

2. Literature review

In the literature, the SASV system is usually referred to as joint ASV and CM systems. There are mainly two categories of methods: (a) multi-task learning-based and (b) fusion-based.

10.21437/Odyssey.2022-11

The comparison of their general structures is illustrated in Fig. 1.

2.1. Multi-task learning-based methods

Li et al. [16] proposed a SASV system to perform a joint decision by multi-task learning. The ASV task and CM task share the same spectrum features and a few network layers. A three-stage training paradigm with pre-training, re-training, and speaker enrollment is proposed to extract a common embedding and perform classification with separate classifiers for the two sub-tasks. They further extended their work in [17] by training the common embedding with triplet loss and then using probabilistic linear discriminant analysis (PLDA) scoring for inference. Zhao et al. [18] adapt the multi-task framework with max-feature map activation and residual convolutional blocks to extract discriminative embeddings.

The training of such multi-task neural networks requires both the speaker label and the spoofing labels, so they are trained on ASVspoof datasets which have a limited number of speakers. This might lead the model to overfit the seen speakers and limit their performance in real-world applications.

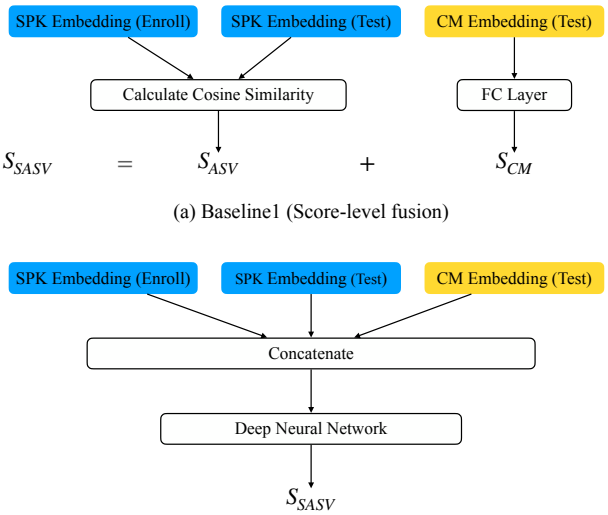
2.2. Fusion-based methods

As shown in Fig. 1(b), independent ASV and CM models extract separate embeddings to make a joint decision. The speaker (SPK) embedding aims to encode the identity information. The CM embedding is usually the output from the second last layer in the anti-spoofing network.

Some methods perform fusion in the embedding space. Sizov et al. [19] proposed a two-stage PLDA method for optimizing the joint system in the i-vector space. First, it trains a simplified PLDA model using only the embeddings of the bona fide speech. Then, it estimates a new mean vector, adds a spoofing channel subspace, and trains it using only the embeddings of the spoofed speech. Gomez et al. [20] proposed an integration framework with fully connected (FC) layers following the concatenated speaker and CM embeddings.

Some methods perform fusion in the score level. The ASV score is usually the cosine similarity between the speaker embeddings of the enrollment utterances and test utterances. The CM score is the final output of the anti-spoofing model. Sahidullah et al. [12] first studied the cascade and parallel integrations of ASV with CM to combine scores. Todisco et al. [21] proposed a Gaussian back-end fusion method that fuses the scores with log-likelihood ratio according to separately modeled Gaussian mixtures. Kanervisto et al. [22] proposed a reinforcement learning paradigm to optimize tandem detection cost function (t-DCF) by jointly training a tandem ASV and CM system. Shim et al. [23] proposed a fusion-based approach that takes the speaker embedding and CM prediction as input and weighs the ASV score, CM score, and their multiplication to make the final decision.

SASV Baseline methods. The SASV challenge [15] introduces two baselines built upon pre-trained state-of-the-art ASV and CM systems. The structure of the two methods is shown in Fig. 2. Baseline1 is a score-level fusion method that sums the scores produced by the separate systems. There is no training involved. Besides, Baseline2 is an embedding-level fusion method that trains a deep neural network based on concatenated embeddings. The pre-trained speaker and CM embeddings are fixed during training the deep neural network. This is similar to the method proposed in [20].



(a) Baseline1 (Score-level fusion)

(b) Baseline2 (Embedding-level fusion)

Figure 2: Model structure of the baseline methods from the SASV challenge. Colored boxes denote the embeddings and the bordered boxes represent the operations.

3. Method

3.1. Problem formulation

Given an enroll utterance u^e and a test utterance u^t, SASV systems need to classify u^t into $y^t \in \{0, 1\}$, where 1 represents *target* and 0 includes both *non–target* and *spoof*. In this paper, we focus on a fusion-based SASV system consisting of a pre-trained ASV subsystem and a pre-trained CM subsystem. In fusion-based SASV systems, The ASV subsystem computes speaker embeddings x^e_{ASV} for the enrollment utterance u^e and x^t_{ASV} for the test utterance u^t. The CM subsystem computes the CM embedding x^t_{CM} for u^t. We use pre-trained embedding methods for the ASV subsystem [24] and the CM subsystem [25], as they both achieve state-of-the-art discrimination abilities on their respective tasks.

As it is a binary classification problem, we use the posterior probability that the test utterance belongs to the positive class (i.e., the *target* class), conditioned on the speaker embeddings, as the final decision score $\mathcal{S}_{SASV}$.

$$\mathcal{S}_{SASV} = P(y^t = 1 | x^e_{ASV}, x^t_{ASV}, x^t_{CM}). \quad (1)$$

For score-level fusion methods, the ASV and CM subsystems each computes a decision score. Similar to Eq. (1), such decision scores can be defined as the posterior probabilities, as $P(y^t_{ASV} = 1 | x^e_{ASV}, x^t_{ASV})$ and $P(y^t_{CM} = 1 | x^t_{CM})$, respectively. Here y^t_{ASV} and $y^t_{CM} \in \{0, 1\}$ are the underlying ground-truth labels along the ASV and CM aspects, respectively. In other words, $y^t_{ASV} = 1$ and $y^t_{ASV} = 0$ indicate that the test utterance is target and non-target, respectively. $y^t_{CM} = 1$ and $y^t_{CM} = 0$ indicate that the test utterance is bona fide and spoof, respectively.

It is noted that these definitions of scores using posterior probabilities are different from those in the baseline methods in Figure 2. There $\mathcal{S}_{ASV}$ is defined as the cosine similarity between the enrollment embedding and the test embedding, and $\mathcal{S}_{CM}$ is defined as the output of an FC layer. Both value ranges are not between 0 and 1. In the following, we will propose ways to revise the scores in Figure 2(a) to fit into the proposed probabilistic framework.

3.2. Probabilistic framework

We propose a probabilistic framework based on product rule (PR) inspired by [26]. By definition, $y^t = 1$, i.e., the test utterance is *target*, if and only if $y^t_{ASV} = 1$ and $y^t_{CM} = 1$. Therefore, assuming conditional independence between y^t_{ASV} and y^t_{CM} on the speaker embeddings, we have

$$\begin{aligned} &P(y^t = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM}) \\ =&P(y^t_{ASV} = 1, y^t_{CM} = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM}) \\ =&P(y^t_{ASV} = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM})P(y^t_{CM} = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM}) \\ =&P(y^t_{ASV} = 1|x^e_{ASV}, x^t_{ASV})P(y^t_{CM} = 1|x^t_{CM}). \end{aligned} \tag{2}$$

The last equation follows from the fact that y^t_{ASV} is independent from x^t_{CM} and that y^t_{CM} is independent from x^e_{ASV} and x^t_{ASV}, as we use pre-trained ASV and CM subsystems. If however, such subsystems are fine tuned during the SASV task, as in Section 3.3.2, this independence will not be valid anymore.

3.3. Proposed strategies

3.3.1. Direct inference strategy

We adopt the same model structure as the base of the Baseline1 method, shown in Fig. 2 (a). The ASV subsystem outputs the cosine similarity between the speaker embedding x^e_{ASV} and x^t_{ASV}. The CM system outputs the CM score $\mathcal{S}_{CM}$ from an FC layer. As both the ASV and CM subsystems are pre-trained and there is no fine tuning in any part of the entire system, this is a direct inference strategy.

As mentioned above, both the ASV score and the CM score do not fit to the proposed probabilistic framework. Therefore, we propose ways to modify their value range to $[0, 1]$. The CM subsystem was pre-trained with a softmax binary classification loss, so the output score $\mathcal{S}_{CM}$ after a sigmoid function $\sigma(x)$ would naturally fit to the range of $[0, 1]$, therefore, we define

$$P(y^t_{CM} = 1|x^t_{CM}) = \sigma(\mathcal{S}_{CM}). \tag{3}$$

For the ASV score, we need some function f to monotonically map the cosine similarity score to a value between 0 and 1:

$$P(y^t_{ASV} = 1|x^e_{ASV}, x^t_{ASV}) = f(\mathcal{S}_{ASV}), \tag{4}$$

where f can be a hand-crafted function or some data-driven mapping. Combining Eq.(1)-(4), the final decision score for SASV is represented as:

$$\mathcal{S}_{SASV} = \sigma(\mathcal{S}_{CM}) \times f(\mathcal{S}_{ASV}). \tag{5}$$

By varying the function f, we propose three systems using the direct inference strategy. A straightforward method is through a linear mapping $f(s) = (s + 1)/2$. We refer to this system as `PR-L-I`, where `L` stands for "linear" and `I` is short for "inference". For non-linear mapping, we choose the sigmoid function and denote the system as `PR-S-I`, where `S` means "sigmoid". A potential advantage of a sigmoid function over the linear mapping is that it expands the data range around 0, the more ambiguous region for decisions. It is noted that neither the linear or sigmoid mapping can result in probabilities that follow the true posterior distribution, therefore, we introduce a third mapping that is trained on the bona fide trials of the development set for $\mathcal{S}_{ASV}$. To be specific, we sample target and non-target trials and train a calibration function with logistic regression [27], where the target class is map to 1 and the non-target class is mapped to 0. This can be viewed as a data-driven score calibrator. This system using the data-driven calibrated ASV score is represented as `PR-C-I`. It is expected that when the test utterance is drawn from the same distribution of the trials used to train the calibrator, the ASV subsystem performance would be improved. This hypothesis is verified in our experiments in Table 3.

3.3.2. Fine-tuning strategy

When the ASV and CM subsystems are fine tuned on the SASV task, then the conditional independence assumption in the last equality of Eq. (2) no longer holds. Instead, we can have an alternative derivation of the posterior probability:

$$\begin{aligned} &P(y^t = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM}) \\ =&P(y^t_{ASV} = 1, y^t_{CM} = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM}) \\ =&P(y^t_{ASV} = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM})P(y^t_{CM} = 1|y^t_{ASV}, x^e_{ASV}, x^t_{ASV}, x^t_{CM}) \\ =&P(y^t_{ASV} = 1|x^e_{ASV}, x^t_{ASV})P(y^t_{CM} = 1|y^t_{ASV}, x^t_{CM}). \end{aligned} \tag{6}$$

The second equality is based on the chain rule and it treats y^t_{ASV} as a condition. It can be interpreted as that the prediction of the CM subsystem depends on that of the ASV subsystem. This dependency can be realized through fine-tuning the CM subsystem conditioned on the ASV system's output score. To do so, we fine-tune the FC layer of the CM subsystem while keeping the ASV score fixed in Figure 2(a). Instead of fitting $\mathcal{S}_{CM}$ with CM labels, our model directly optimizes the joint score. The training is based on the ground-truth label of whether the test utterance belongs to the *target* class. In other words, the *spoof* and *non-target* utterances share the same negative labels. The final decision score $\mathcal{S}_{SASV}$ is calculated with Eq. (5).

We fine-tune the system with a prior-weighted binary cross-entropy loss for $\mathcal{S}_{SASV}$. The ASV embedding network is pre-trained and fixed, hence the ASV score $\mathcal{S}_{ASV}$ is fixed. Only the FC Layer on top of the CM embedding network is trained and the CM score $\mathcal{S}_{CM}$ is adjusted. During back-propagation, thanks to the multiplication, the gradient of the CM score with respect to the parameters in the FC layer is weighted based on the scaled ASV scores. The gradient receives a larger weight for larger $\mathcal{S}_{ASV}$, which corresponds to utterances that are more similar to the target speaker. This helps the model to pay more attention to such more difficult samples, manifesting an idea of *speaker-aware anti-spoofing*.

In fine tuning strategy, we choose f as the linear or the sigmoid function, denoted as `PR-L-F` and `PR-S-F` respectively. `L` and `S` represent the two mapping functions as in Section 3.3.1, while `F` is short for "fine-tuning". We discard the calibration method to prevent over-fitting on the trials dataset.

4. Experimental setup

4.1. Dataset

ASVspoof 2019 LA [28] is a standard dataset designed for the LA sub-challenge of ASVspoof 2019. It consists of bona fide speech and a variety of TTS and VC spoofing attacks. The bona fide speech is collected from the VCTK corpus [29], 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 set are totally different from those in the Train and Dev sets. The non-overlap is designed to encourage the generalization ability to unseen attacks for CM systems. Details are shown in Table 1.

Table 1: Summary of the ASVspoof 2019 LA dataset.

Partition	#speakers	Bona fide	Spoofing attacks	
		#utterances	#utterances	Attacks type
Train	20	2,580	22,800	A01 - A06
Dev	20	2,548	22,296	A01 - A06
Eval	67	7,355	63,882	A07 - A19

For the SASV challenge, the organizers provided official development and evaluation protocols listing the *target*, *non-target*, and *spoof* trials based on the ASVspoof 2019 LA dataset. For each test trial, there are multiple corresponding enrollment utterances to register the target speaker.

4.2. 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 the three classes. For all of the EERs mentioned above, only the target class is considered positive samples.

Table 2: Three kinds of EERs for evaluation (Adapted from [15]). "+" denotes the positive class and "-" denotes the negative class. A blank entry denotes classes not used in the metric. SASV-EER is the primary metric for the SASV challenge.

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

4.3. Implementation details

Our implementation is based on PyTorch [1]. The pre-trained embeddings are provided by the SASV organizers. They are extracted with already-trained state-of-the-art ASV and CM systems. The ASV system is an ECAPA-TDNN [24] model trained on the VoxCeleb2 dataset [30]. The CM system is an AASIST [25] model trained on ASVspoof 2019 LA training set [28]. For a speech utterance, the speaker embedding has a dimension of 192 and the CM embedding is a 160-dim vector.

For the Baseline2 model structure, the DNN is composed of four FC layers, each with the number of output dimensions as 256, 128, 64, 2, respectively. Each intermediate layer is followed by a leaky ReLU activation function. For inference, we use the official trials provided by the SASV challenge organizers as described in Section 4.1. The calibrator in `PR-C-I` is trained on the bona fide utterances of the development trials.

During training `PR-L-F` and `PR-S-F`, we randomly select pairs of utterances from the training set. For the binary cross-entropy loss, we set the prior probability for a target trial as 0.1. We train our systems using Adam optimizer with an initial learning rate of 0.0003. The batch size is set to 1024. We train the model for 200 epochs and select the best epoch according to the SASV-EER on the development set. The model in the best epoch is used for final evaluation.

5. Results

5.1. Comparison with separate systems and baselines

To demonstrate the effectiveness of our proposed strategies, we compare our methods with the individual systems and baseline methods in the SASV challenge[2]. The performance comparison is shown in Table 3.

The individual systems perform well on their own tasks but have much worse performance on the other task. The ECAPA model achieves the lowest SV-EER but a high value in SPF-EER. This verifies that the state-of-the-art speaker verification system is vulnerable to spoofing attacks. Quite a number of spoofed trials can deceive the ASV system and degrade the SASV performance. The AASIST system has the lowest SPF-EER but close to 50% SV-EER. This is reasonable since all bona fide speech, no matter target or non-target, are considered positive samples in training CM systems. The well-trained CM system is not expected to have discrimination ability for ASV.

Both baseline methods surpass the separate systems in terms of SASV-EER, showing the superiority of an ensemble solution for the SASV problem. Baseline1, a score-level fusion-based method, has the same SPF-EER performance as the single CM system but degrades the ASV performance compared to the ECAPA model. This suggests that the non-calibrated scores might degrade the performance on sub-tasks. Baseline2, the embedding level fusion-based model, has much better performance on all three metrics overall with only the SPF-EER degraded a little on the evaluation set.

All of our proposed systems show a significant improvement over the baseline methods in terms of SASV-EER. They also achieve universally good performance over all three metrics. Both the SV-EER and SPF-EER are close to the performance of the best separate model. This shows the effectiveness of our product rule (PR)-based probabilistic framework with our proposed direct inference strategy and fine-tuning method. Our `PR-S-F` system achieves the best performance on the evaluation trials.

5.2. Comparison among the proposed strategies

Comparing our proposed systems with direct inference strategy (i.e., with `-I`) and systems with fine-tuning strategy (i.e., with `-F`), the latter generally achieve better performance. This suggests the effectiveness of the joint optimization by slacking the conditional independence of ASV and CM subsystems.

Among all the systems with direct inference strategy, we can compare the impact of different choices for the mapping function f applied to the ASV cosine similarity score. The linear mapping achieves better SV-EER and SASV-EER compared

[1] Our work is reproducible with code available at `https://github.com/yzyouzhang/SASV_PR`.

[2] Note that the baseline results we report have differences from those reported in [15]. Based on our implementation, we achieved close results for ECAPA-TDNN and Baseline1, but better results for Baseline2.

Table 3: Comparison of our proposed methods with separate systems and SASV challenge baselines.

Systems	SV-EER↓		SPF-EER↓		SASV-EER↓	
	Dev	Eval	Dev	Eval	Dev	Eval
ECAPA-TDNN	1.86	1.64	20.28	30.75	17.31	23.84
AASIST	46.01	49.24	0.07	0.67	15.86	24.38
Baseline1	32.89	35.33	0.07	0.67	13.06	19.31
Baseline2	7.94	9.29	0.07	0.80	3.10	5.23
PR-L-I (Ours)	2.13	2.14	0.11	0.86	1.21	1.68
PR-S-I (Ours)	2.43	2.57	0.07	0.78	1.34	1.94
PR-C-I (Ours)	1.95	1.64	0.97	2.94	**1.08**	2.70
PR-L-F (Ours)	2.02	1.92	0.07	0.80	1.10	**1.54**
PR-S-F (Ours)	2.02	1.94	0.07	0.80	1.10	**1.53**

to the sigmoid mapping, this might be attributed to the non-linearity of the sigmoid function that distorts the ASV score distribution. The calibrated ASV score achieves the best performance on the development trials in terms of SASV-EER, and the SV-EER is the closest to ECAPA-TDNN, suggesting that the calibration on ASV scores is effective for SASV. However, the calibration degrades the SASV-EER performance and the SPF-EER performance on the evaluation trials prominently. Note that the spoof trials in the development and evaluation trials are generated with different attack algorithms. The performance degradation verifies our hypothesis that the calibration would cause the joint system to overfit the distribution of the trials that the calibrator is trained on hence cannot generalize well to unseen attacks.

Among the two systems with our fine-tuning strategy, both of them achieve top similar performance in all three metrics. This suggests that joint optimization is effective and robust to both linear and sigmoid mapping functions. Although the score mapping functions affect the performance in the direct inference strategy, they do not make much difference in the fine-tuning strategy, thanks to the FC layer re-trained on SASV labels.

5.3. Ablation study on Baseline1

Since our model structure is based on Baseline1, we perform an ablation study to recover the components back to the counterparts in Baseline1 and observe the performance degradation. The results are shown in Table 4. The performance degradation from PR-S-F to PR-S-I verifies the effectiveness of our proposed joint optimization by fine-tuning. Both PR-S-I and Baseline1 are direct inference methods. Comparing Eq. (5) and the formula in Fig. 2 (a), changes on the computation of the SASV score in our proposed approach compared to Baseline1 are: 1) applying sigmoid score mapping on both ASV score and CM score, 2) using multiplication rather than addition.

If we change the multiplication back to summation, i.e., $\mathcal{S}_{SASV} = \sigma(\mathcal{S}_{CM}) + \sigma(\mathcal{S}_{ASV})$, the performance degrades to 2.45% SASV-EER, which is still a relatively good performance. The degradation indicates the superiority of our proposed probabilistic fusion framework with the product rule.

If we only remove the score mapping but keep the multiplication, i.e., $\mathcal{S}_{SASV} = \mathcal{S}_{CM} \times \mathcal{S}_{ASV}$, the performance degrades to 2.89% SASV-EER, which is also an acceptable performance.

When we restore both components back to the Baseline1 method, then the SASV-EER performance degrades significantly. This suggests that both components in our proposed PR-S-I make an effective contribution. What exactly causes the dramatic degradation from PR-S-I to Baseline1? Our hypothesis is that the scores output from the ASV and CM subsystems of Baseline1 are in different ranges, and the summation of the scores makes one subsystem dominates the other. Looking at the Table 3 again, it is the CM system that dominates. Applying score mapping, with multiplication or summation, also addresses this issue. Replacing summation with multiplication, with or without score mapping, addresses this issue, as the difference between the score ranges is just a constant scalar of the final decision score. This explains why both revised methods in Table 4 do not degrade too much from PR-S-I.

Table 4: Results of ablation study from our proposed best performing system PR-S-F to Baseline1.

Systems	SASV-EER	
	Dev	Eval
PR-S-F (Ours)	1.10	1.53
PR-S-I (Ours)	1.34	1.94
Restore multiplication to sum (Baseline1 + score mapping)	1.69	2.45
Remove score mapping (Baseline1 + score multiplication)	2.16	2.89
Restore both (Baseline1)	13.06	19.31

In the next section, we will verify this hypothesis by investigating the scores output from the two subsystems of Baseline1, as well as the revised scores after applying score mapping.

6. Score distribution analysis

Fig. 3 shows the score distribution of the systems we compared in Table 3. We plot the histogram of score distributions on both the official development and evaluation trials.

Fig. 3 (a) and (b) first plot score distributions of the ASV subsystem (ECAPA-TDNN) and the CM subsystem (AASIST). They demonstrate good discriminative abilities on their individual tasks, but fails to differentiate classes defined in the other task. For example, ECAPA-TDNN well distinguishes *target* and *non-target*, but the distribution of *spoof* expands a wide range, overlapping with both the *target* and *non-target* classes. This shows that the ASV system is vulnerable to spoofing attacks. It is interesting to see that the scores of spoofing attacks

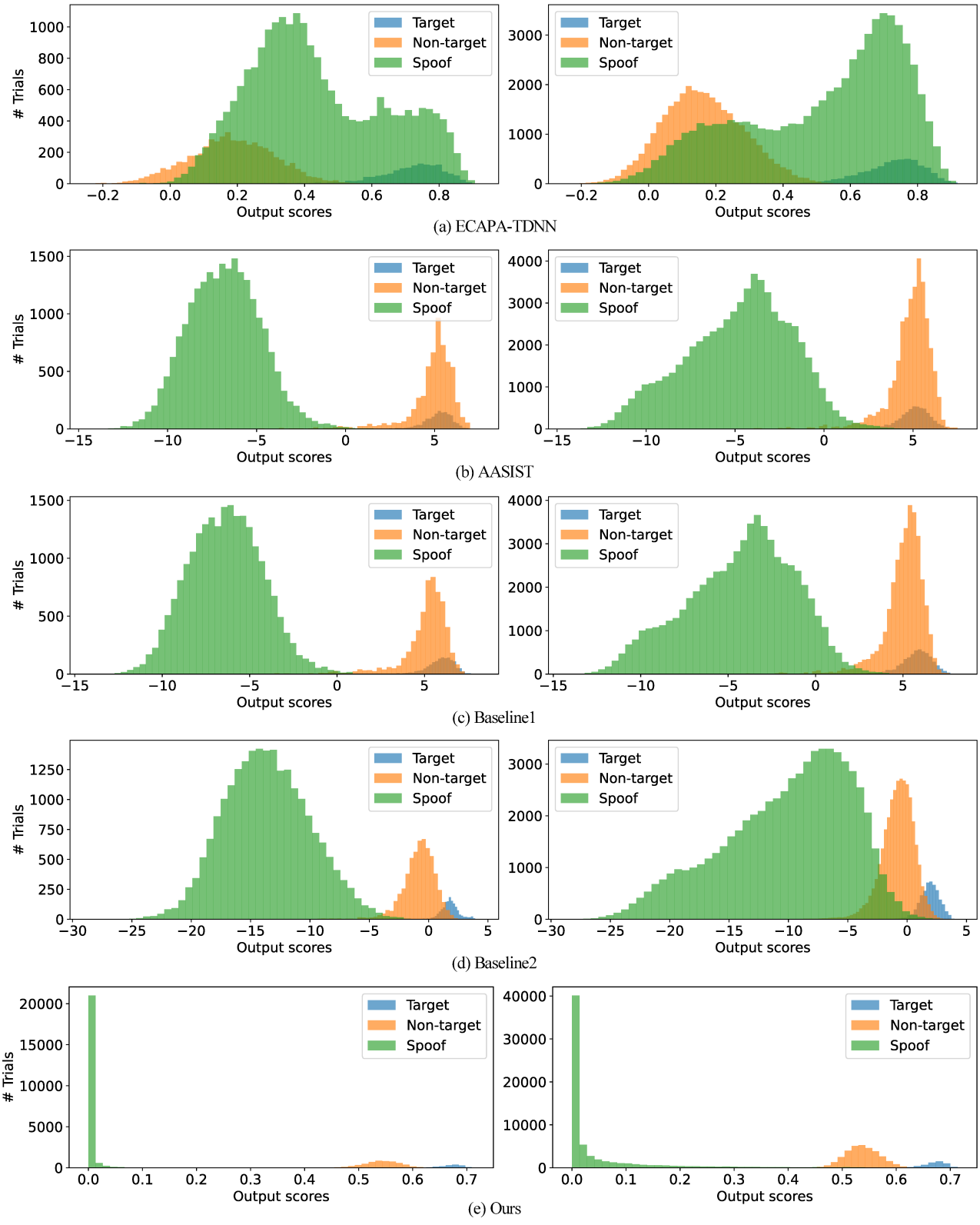


Figure 3: Comparison among score distributions of (a) the ASV subsystem (ECAPA-TDNN), (b) the CM subsystem (AASIST), (c) Baseline1, (d) Baseline2, and (e) our proposed best-performing method `PR-S-F`. The left column is the performance on the development set and the right column is on the evaluation set. Different colors correspond to the three label classes: target, non-target, and spoof.

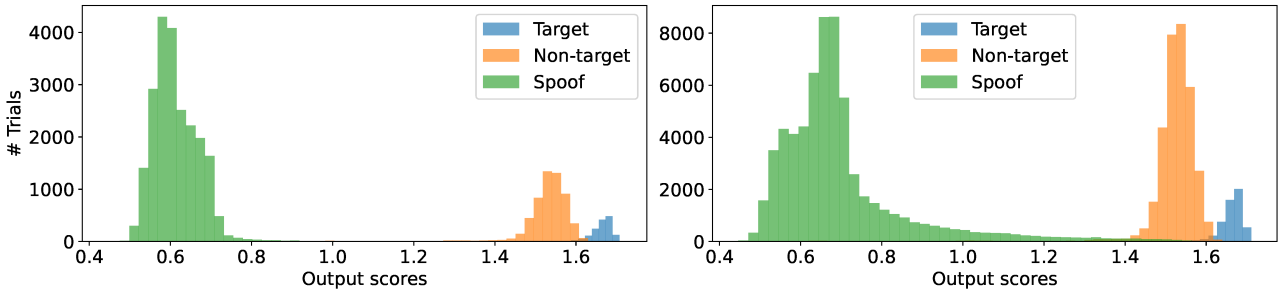


Figure 4: Score distributions of applying score mapping on Baseline1 system.

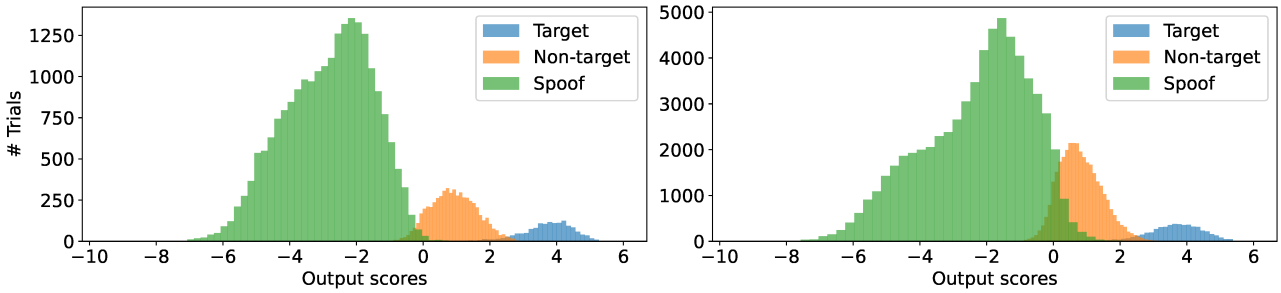


Figure 5: Score distributions of applying score multiplication on Baseline1 system.

on the evaluation set (right column) are closer to those of the target class. This might suggest that the spoofing attacks in the evaluation set are more challenging to the whole system.

Similarly, for AASIST in Fig. 3 (b), the spoof class score is well-separated from the target and non-target classes. However, the target and the non-target classes are highly overlapped since they are both bona fide speech. The CM system only has the ability to discriminate spoofing attacks from bona fide speech.

For Baseline1 in Fig. 3 (c), the distribution is similar to that in (b), the difference is that the non-target cluster and the target cluster are deviated by some distance. Recall that Baseline1 takes the sum of the independent scores output by ECAPA-TDNN and AASIST. Comparing (a), (b), and (c), we can infer that the CM system dominates the score. From the score ranges shown in (a) and (b), the absolute values of the CM scores are larger than those of the ASV scores. This verifies our reasoning for why Baseline1 degrades from our proposed `PR-S-I` so much in the previous section.

For the Baseline2 system in Fig. 3 (d), the distribution shows that the three classes are more separated than previous systems. This suggests that the embedding-level fusion maintains a good discrimination ability for the target class.

From the ablation study in Section 5.3, we find that with simple score mapping and score multiplication, the resulting system is able to achieve a significant improvement over the score-sum baselines. To better understand the mechanisms behind each operation, we plot the histogram of the SASV score distribution with $\mathcal{S}_{SASV} = \sigma(\mathcal{S}_{CM}) + \sigma(\mathcal{S}_{ASV})$ and $\mathcal{S}_{SASV} = \mathcal{S}_{CM} \times \mathcal{S}_{ASV}$ in Fig. 4 and Fig. 5 respectively. From Fig. 4, we can observe that the scores are in the range of (0, 2) and the three classes are well separated, indicating the effectiveness of score scaling, where both individual scores are mapped to the same range. Similarly, Fig. 5 shows scores from the distinct three classes clearly, but not as well separated as the previous scaling method.

7. Conclusion

In this paper, we proposed effective fusion-based methods for spoofing aware speaker verification (SASV). Specifically, we introduced a probabilistic framework with the product rule and a fine-tuning strategy to a score-sum fusion baseline structure. We demonstrated promising performance with a SASV-EER at 1.53%, a significant improvement from the previous EER of 19.31%. Our ablation study verified the effectiveness of our proposed strategies and we investigated the SASV decision score distributions of various systems.

8. Acknowledgment

This work is supported by National Science Foundation grant No. 1741472, New York State Center of Excellence in Data Science award, and funding from Voice Biometrics Group. You Zhang thanks the synergistic activities provided by the NRT program on AR/VR funded by NSF grant DGE-1922591.

The authors would like to thank Xinhui Chen for delivering a literature review presentation on *Joint Speaker Verification and Spoofing Countermeasure Systems* during her master's study at University of Rochester.

The authors would like to thank the organizers of the SASV 2022 challenge for providing the pre-trained embeddings.

9. References

[1] Asmaa El Hannani, Dijana Petrovska-Delacrétaz, Benoît Fauve, Aurélien Mayoue, John Mason, Jean-François Bonastre, and Gérard Chollet, *Text-independent Speaker Verification*, pp. 167–211, Springer, London, 2009.

[2] Youzhi Tu and Man-Wai Mak, "Mutual information enhanced training for speaker embedding," in *Proc. Interspeech*, 2021, pp. 91–95.

[3] Ge Zhu, Fei Jiang, and Zhiyao Duan, "Y-vector: Multi-scale waveform encoder for speaker embedding," in *Proc. Interspeech*, 2021, pp. 96–100.

[4] Hongning Zhu, Kong Aik Lee, and Haizhou Li, "Serialized multi-layer multi-head attention for neural speaker embedding," in *Proc. Interspeech*, 2021, pp. 106–110.

[5] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[6] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey*, 2020, pp. 132–137.

[7] You Zhang, Fei Jiang, and Zhiyao Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[8] Xingliang Cheng, Mingxing Xu, and Thomas Fang Zheng, "Cross-database replay detection in terminal-dependent speaker verification," in *Proc. Interspeech*, 2021, pp. 4274–4278.

[9] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan, "UR channel-robust synthetic speech detection system for ASVspoof 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 75–82.

[10] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan, "CRIM's system description for the ASVSpoof2021 challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 100–106.

[11] Hongji Wang, Heinrich Dinkel, Shuai Wang, Yanmin Qian, and Kai Yu, "Cross-domain replay spoofing attack detection using domain adversarial training.," in *Proc. Interspeech*, 2019, pp. 2938–2942.

[12] Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Hong Yu, Tomi Kinnunen, Nicholas Evans, and Zheng-Hua Tan, "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on asvspoof 2015," in *Proc. Interspeech*, 2016, pp. 1700–1704.

[13] Tomi Kinnunen, Kong Aik Lee, and Héctor Delgado *et al.*, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey*, 2018, pp. 312–319.

[14] Andreas Nautsch, Xin Wang, and Nicholas Evans *et al.*, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[15] Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Hong-Goo Kang, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen, "SASV challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.

[16] Jiakang Li, Meng Sun, and Xiongwei Zhang, "Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2019, pp. 1517–1522.

[17] Jiakang Li, Meng Sun, Xiongwei Zhang, and Yimin Wang, "Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss," *IEEE Access*, vol. 8, pp. 7907–7915, 2020.

[18] Yuanjun Zhao, Roberto Togneri, and Victor Sreeram, "Multi-task learning-based spoofing-robust automatic speaker verification system," *Circuits, Systems, and Signal Processing*, pp. 1–22, 2022.

[19] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 10, no. 4, pp. 821–832, 2015.

[20] Alejandro Gomez-Alanis, Jose A Gonzalez-Lopez, S Pavankumar Dubagunta, Antonio M Peinado, and Mathew Magimai Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 16, pp. 1579–1593, 2020.

[21] Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Proc. Interspeech*, 2018.

[22] Anssi Kanervisto, Ville Hautamäki, Tomi Kinnunen, and Junichi Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 477–488, 2021.

[23] Hye-jin Shim, Jee-weon Jung, Ju-ho Kim, and Ha-jin Yu, "Integrated replay spoofing-aware text-independent speaker verification," *Applied Sciences*, vol. 10, no. 18, pp. 6292, 2020.

[24] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[25] Jee-weon Jung, Hee-Soo Heo, and Hemlata Tak *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[26] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 3, pp. 226–239, 1998.

[27] Niko Brummer, "Focal-ii: Toolkit for calibration of multiclass recognition scores," `https://sites.google.com/site/nikobrummer/focalmulticlass`, August, 2006.

[28] Xin Wang, Junichi Yamagishi, and Massimiliano Todisco *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[29] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[30] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.