UR Spoofing Aware Speaker Verification System for the SASV Challenge 2022

You Zhang, Ge Zhu, Zhiyao Duan

Audio Information Research Lab, University of Rochester, Rochester, NY, USA

{you.zhang, ge.zhu, zhiyao.duan}@rochester.edu

Abstract

This paper presents UR-AIR spoofing aware speaker verification (SASV) system submission to the SASV challenge 2022. The challenge aims to encourage the integration of automatic speaker verification (ASV) and spoofing countermeasure (CM) subsystems to improve the performance of ASV systems when they are exposed to spoofing attacks. We adopt a probabilistic fusion framework on top of pre-trained and fixed speaker embeddings and CM embeddings. We also describe our model implementation for joint optimizing the system under the probabilistic framework. The best single model in our experiments achieves a SASV-EER of 1.01% and 1.34% on the official development and evaluation trials, respectively.

Index Terms: spoofing aware speaker verification, spoofing countermeasure, probabilistic framework

1. Introduction

Automatic speaker verification (ASV) systems are vulnerable to spoofing attacks, where synthesized or replayed speech is presented to deceive the system on the speaker identity [1]. Spoofing countermeasure (CM) systems aim to detect whether the speech is bona fide, i.e. natural speech from humans. Recent progress has been made on a standalone CM [2, 3, 4, 5] and the best performing CM system can achieve an equal error rate (EER) of less than 1%. However, there CM is considered as a separate task and the improvement of CM may not benefit ASV since the two systems are not jointly optimized.

The SASV challenge [6] aims to build the gap between ASV and CM and encourage the joint optimization of the systems. The dataset is built on ASVspoof 2019 LA, where spoofing attacks are presented. The SASV task is a binary classification problem that aims to discern whether the test utterance is bona fide speech of the target speaker. Same as ASV, some bona fide utterances are provided to register the speaker in the enrollment stage. Positive labels represent the *target* class, i.e. the test utterance belongs to the target speaker and is bona fide speech. For negative labels, there are two cases: the *non-target* class, the test utterance is bona fide speech from a speaker other than the target speaker; the *spoof* class, where the test utterance is spoofing attacks. The challenge uses SASV-EER as the primary evaluation metric.

In this work, we employ our probabilistic fusion framework previously proposed in [7] to map the scores from ASV and CM subsystems to probabilities and calculate the posterior probability as the final score for SASV. We also report score average fusion results for different systems. Our best performing system achieves 1.34% SASV-EER on the evaluation trials, surpassing the baselines by a large margin.

2. Probabilistic Fusion Framework

In this section, we elaborate on our probabilistic fusion framework proposed in [7]. Suppose we have speaker embeddings

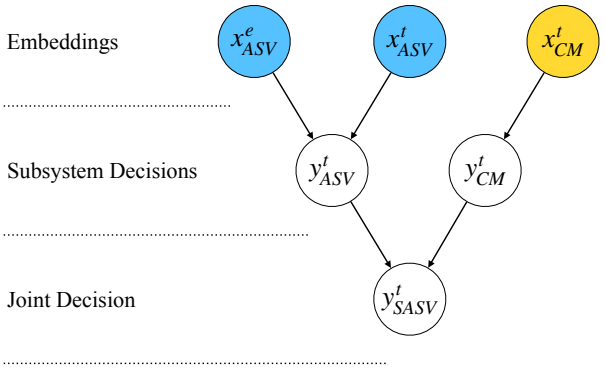

Figure 1: *The belief network of our proposed probabilistic fusion framework for SASV. The embeddings are computed from pre-trained and fixed ASV and CM embedding networks. "e" and "t" denote for enrollment and test, respectively.*

x^e_{ASV} and x^t_{ASV} for the enrollment utterance and test utterance respectively, computed by the ASV subsystem; and we also have access to the CM embedding x^t_{CM} of the test utterance, computed by the CM subsystem. We aim to calculate the posterior probability of the test utterance belongs to the *target* class, as $P(y^t_{SASV} = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM})$. Here $y^t_{SASV} \in \{0, 1\}$ as SASV is a binary classification problem. $y^t_{SASV} = 1$ and $y^t_{SASV} = 0$ indicate the ground-truth positive and negative labels as introduced in Section 1.

As our SASV system is a fusion of ASV and CM subsystems, we denote the underlying ground-truth labels from the ASV and CM aspects as y^t_{ASV} and $y^t_{CM} \in \{0, 1\}$. The positive label $y^t_{ASV} = 1$ indicates that the test utterance belongs to the target speaker, whereas the positive label $y^t_{CM} = 1$ means the test utterance is bona fide speech.

Figure 1 shows the belief network for our proposed probabilistic fusion framework. The joint decision of SASV is made by fusing the decisions from the ASV and CM subsystems. By definition, $y^t_{SASV} = 1$, if and only if $y^t_{ASV} = 1$ and $y^t_{CM} = 1$. Thus, we derive the posterior probability as follows:

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

The second equality is based on the chain rule and it treats y^t_{ASV} as a condition. 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. It could be counter-intuitive that the prediction of the CM subsystem depends on that of the ASV subsystem. A better way to interpret is by assuming conditional independence be-

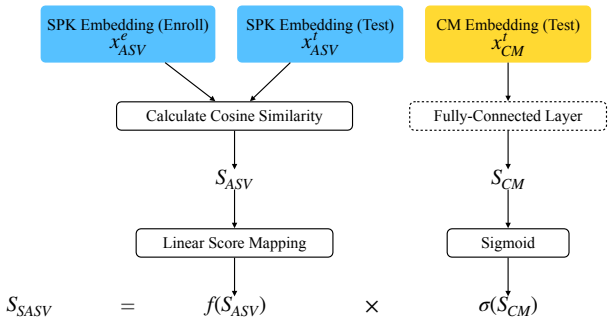

Figure 2: *Model structure for implementation of our proposed framework. Colored boxes represent the embeddings and the bordered boxes denote the operations. The dashed border denotes that the FC layer is trainable. Borderless symbols represent the outputs or intermediate outputs.*

tween y^t_{ASV} and y^t_{CM} and then slack it. Further details can be referred to in [7].

3. Model Implementation

We implement the SASV system based on our probabilistic fusion framework. We employ the posterior probability as the final decision score for the SASV system.

$$\mathcal{S}_{SASV} = P(y^t_{SASV} = 1 | x^e_{ASV}, x^t_{ASV}, x^t_{CM}). \quad (2)$$

3.1. Model Structure

The structure of our model is shown in Figure 2. The ASV subsystem computes the cosine similarity between the speaker embeddings x^e_{ASV} and x^t_{ASV}, as the ASV score $\mathcal{S}_{ASV} \in [-1, 1]$. We then perform a linear function $f(s) = (s + 1)/2$ to monotonically map the score to fit the probability range of $[0, 1]$. As such, we define

$$P(y^t_{ASV} = 1 | x^e_{ASV}, x^t_{ASV}) = f(\mathcal{S}_{ASV}). \quad (3)$$

For the CM subsystem, we use a fully-connected (FC) layer to map the CM embedding to a score and then apply the Sigmoid function to ensure the output is in the range of $[0, 1]$. We assign

$$P(y^t_{CM} = 1 | y^t_{ASV}, x^t_{CM}) = \sigma(\mathcal{S}_{CM}). \quad (4)$$

Note that we ignore the FC layer followed by the CM embedding in the original pre-trained CM subsystem, which is different from the Baseline1 method in [6]. Our FC layer can be interpreted as a re-initialization of that FC layer. The decision of y^t_{CM} is dependent on y^t_{ASV} since we train the FC layer with the information from the ASV branch. The training details is described in the next subsection.

Combining Eq. (1)-(4), the final decision score for our SASV system is denoted as:

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

3.2. Training details

Both the ASV and CM 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. We train the system with a prior-weighted binary cross-entropy loss for the joint decision score $\mathcal{S}_{SASV}$ and we set the prior probability for a target trial as 0.1. Therefore, the system is jointly optimized to fit the SASV ground-truth binary labels. The dependency of y^t_{CM} on y^t_{ASV} is thus realized by training the FC layer conditioned on the ASV output score.

Regarding the training data, we randomly select pairs of utterances from the training set of the ASVspoof 2019 logical access dataset and assign SASV labels by definition.

Table 1: *Four kinds of EERs for evaluation (Adapted from [6]). "+" 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	+		-
CM-EER	+	+	-

4. Experimental Results

In our experiments, we train our systems using Adam optimizer with an initial learning rate of 0.00003. The batch size is set to 1024. We train the model for 500 epochs and select the best epoch according to the SASV-EER on the development trials. The model in the best epoch is used for final evaluation. For evaluation on the official development and evaluation trials, our system's output decision scores are calculated with Eq. (5) and we then calculate the EERs in Table 1 for two sets of trials, respectively. Note that we also observe the CM-EER as it is the standard evaluation metric for CM systems, but excluded in the SASV challenge.

4.1. Performance of single systems with our method

We train and evaluate eight of our SASV systems by varying the random seed. The random seed controls (1) the random selection of pairs from the training data (2) the random initialization of the FC layer. As shown in Table 2, the system performance varies according to random seed. This observation is similar to anti-spoofing as observed in a comparative study [8]. The best performing system achieves a SASV-EER of 1.01% and 1.34% on the development and evaluation trials, respectively.

Comparing different random seeds, in general, a better SASV-EER is originated from a better SV-EER, showing that it is important to discriminate the speaker identity. However, on the CM aspect, a better SPF-EER or CM-EER may not result in a better SASV-EER. The CM-EER of the best-performing system is extremely high compared to the other systems. This shows that optimizing the CM subsystem and achieving a lower CM-EER might not benefit the SASV system. This observation might be of interest to the CM community.

4.2. Fusion of single systems

We also experiment with score average fusion for different single systems to try to improve the performance. The fusion results is shown in Table 3. The fusion of single systems did not generate better performance than single systems.

Table 2: *Comparison among different single systems of our proposed methods with varying the random seed. In each cell, we report the EER on the development trials (top) and that on the evaluation trials (bottom). The results of eight training-evaluation rounds are sorted by the SASV-EER on the development trials from high (I) to low (VIII).*

Seed	SASV-EER	SV-EER	SPF-EER	CM-EER
I	1.12	2.02	0.07	0.69
	1.56	1.97	0.82	2.00
II	1.11	2.02	0.07	0.65
	1.56	1.94	0.82	2.12
III	1.11	2.02	0.07	0.69
	1.53	1.94	0.81	2.21
IV	1.11	2.02	0.07	0.62
	1.49	1.88	0.80	2.27
V	1.08	1.95	0.07	0.66
	1.56	1.94	0.87	2.45
VI	1.08	2.02	0.07	0.62
	1.51	1.92	0.82	2.14
VII	1.08	2.02	0.09	0.76
	1.48	1.92	0.80	2.63
VIII	**1.01**	1.75	0.20	11.84
	1.34	1.70	1.08	18.16

Table 3: *Results for score average fusion of single systems in Table 2. In each cell, we report the EER on the development trials (top) and that on the evaluation trials (bottom).*

Fusion	SASV-EER	SV-EER	SPF-EER	CM-EER
I+II+III	1.11	2.02	0.07	0.69
	1.56	1.94	0.81	2.10
V+VI+VIII	1.08	2.02	0.07	0.68
	1.51	1.94	0.82	2.36
All	1.14	2.02	0.07	0.70
	1.49	1.94	0.82	2.40

4.3. Our system submission

We submit the single system with seed VIII in Table 2 since it achieves the lowest SASV-EER with 1.01% on the development trials. It achieves 1.34% SASV-EER on the evaluation trials.

We acknowledge that the random seed for our submitted single system could be a lucky number since it achieved the unusually best performance, compared to other random seeds. However, the other single systems in Table 2 all outperform the baselines in [6] by a large margin. We demonstrated the comparison of methods and performed an ablation study in [7], verifying the effectiveness of our proposed probabilistic fusion framework.

5. Conclusions

In this paper, we introduced UR-AIR system submission for the SASV challenge 2022. We adopted our previously proposed probabilistic fusion framework and trained a fully-connected layer to fuse the pre-trained and fixed ASV and CM embeddings. Our training method directly optimizes the joint SASV confidence score. Our submitted single system achieved 1.34% SASV-EER on the official evaluation trials.

6. Acknowledgments

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 would like to thank the synergistic activities provided by the NRT program on AR/VR funded by NSF grant DGE-1922591.

7. References

[1] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000788

[2] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

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

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

[5] 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," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[6] 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.

[7] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," *arXiv preprint arXiv:2202.05253*, 2022.

[8] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Interspeech 2021*, 2021, pp. 4259–4263.