System Description of Team IRLAB for SASV Challenge 2022

Jungwoo Heo[1], *Ju-ho Kim*[1], *Hyun-seo Shin*[1]

[1]School of Computer Science, University of Seoul, Republic of Korea

jungwoo4021@gmail.com, wngh1187@naver.com, gustjtls123@naver.com

Abstract

This report introduces the submission of the team IRLAB for the Spoofing-Aware Speaker Verification (SASV) challenge 2022. We explored two back-end models for the SASV task based on the score and embedding fusion methods. As a score fusion solution, we devised a score fusion model that performs speaker verification in spoofing scenarios, and finally derived a SASV score by aggregating speaker verification, anti-spoofing, and the score fusion model scores using a fully connected DNN. On the other hand, from the perspective of embedding fusion solution, we designed an integrated embedding projector that casts speaker and countermeasure embeddings to an SASV embedding, and calculated final score between the SASV embeddings based on the cosine similarity. Each of the proposed systems achieved equal error rates of 0.56% and 1.32% for the SASV evaluation protocol.

Index Terms: speaker verification, spoofing attacks, spoofing-aware speaker verification (SASV)

1. Introduction

This technical report describes the proposed back-end integrated models based on the score and embedding fusion approaches. Our score fusion solution uses the scores of the speaker verification (SV) and spoofing countermeasure (CM) subsystems directly as in baseline1 [1]. Furthermore, we introduced a score fusion model trained on the SV task in spoofing scenarios to improve the score of the SV subsystem. Then, the scores of the SV, CM, and the score fusion models are integrated using a multi-layer perceptron (MLP), and a final score for SASV is derived. Additionally, from the perspective of embedding fusion solution, we further proposed an integrated embedding projector that convert SV and CM embeddings into SASV embeddings. The embedding projector is trained using a metric learning loss, and the SASV score is directly calculated based on cosine similarity through comparison between the integrated embeddings.

2. SV and CM subsystems

To devise a back-end system for SASV task, we used the pre-trained ECAPA-TDNN [2] and AAIST [3] models provided by the challenge organizer as the subsystems of SV and anti-spoofing tasks [1]. Both subsystems used the pre-trained versions found in this link [1]. We modified the SV subsystems to extract a speaker embedding using the full utterance instead of 400 frames of the utterance.

3. Proposal1: MLP score fusion model

The proposed score fusion solution is based on the motivation from several observations of the experiments. Figure 1 shows the distribution of the ECAPA-TDNN's SV score for all development trials. The distribution of target and non-target are clearly separated, which means the SV subsystem reliably performs in bonafide scenarios. However, the distribution of spoofed utterances is widely dispersed, overlaying both the target and non-target distributions. This phenomenon is due to the degradation of the discrimination of the SV subsystem in the spoofing scenario.

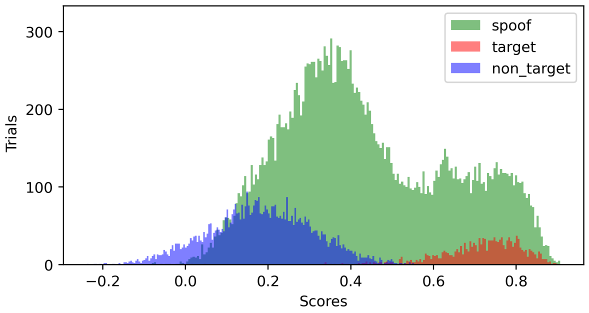

Figure 1: *The SV score distribution of the pre-trained ECAPA-TDNN subsystem for all development trials. The red, blue, and green colors denote the distributions of target, non-target, and spoofed.*

Therefore, to improve the SV score, we propose a speaker verification score fusion model (SVSF) that learns speaker verification in spoofing scenarios. The SVSF model is fed by the embeddings of the SV and CM subsystems and outputs a speaker verification score vector. The structure of SVSF is shown in figure 2 and each module is described in table 1. It consists of two embedding fusion blocks ($u1$, $u2$) and a score calculation block (pj). SV and CM embeddings extracted from test and enrollment utterances are concatenated and fed to $u1$. In the same way, the embeddings extracted from test utterances are fed to $u2$. $u1$ and $u2$ combine the speaker and spoofing information in the embeddings and output 160-dimensional feature vectors. Then the block pj converts the feature vector into a speaker verification score vector. The first node of the output score layer represents the non-target score, and the second node represents the target score. We used only the value of the second node as a SVSF score.

For the score fusion, we used the multi-layer perceptron score fusion (MLPSF) that is fed by SV, CM, and SVSF scores and output a SASV score vector. The structure of the MLPSF is described in the left column of Table 1. It consists of three fully-connected (FC) layers and two exponential linear unit (ELU) layers.

SVSF and MLPSF are trained simultaneously using categorical cross entropy (CCE) criterion. The loss is calculated as follows. Here, t, l are the ground truths of SV and SASV, and s, v denote score vectors output from SVSF and MLPSF, which

[1] https://github.com/sasv-challenge/SASVC2022_Baseline

has 2 nodes.

$$\mathcal{L}_{SVSF} = -\sum_{i}^{2} t_i \log(softmax(s)_i), \tag{1}$$

$$\mathcal{L}_{MLPSF} = -\sum_{i}^{2} l_i \log(softmax(v)_i), \tag{2}$$

$$\mathcal{L}_{TOTAL} = \mathcal{L}_{SVSF} + \mathcal{L}_{MLPSF}. \tag{3}$$

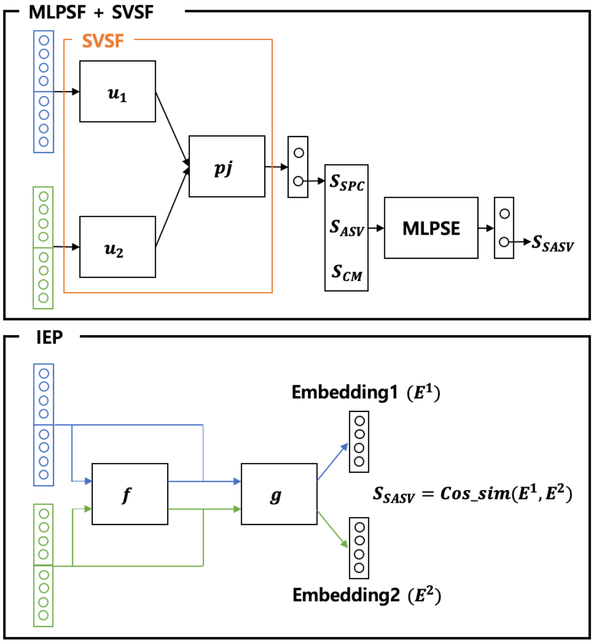

Figure 2: *Description of the structure of the proposed frameworks. Blue and green boxes are the feature vectors that concatenated SV and CM embeddings extracted from enrollment and test utterances. The* $\mathbf{S_{task}}$ *means the score of each* **task**.

Training the SASV system requires the train pairs which contain the enrollment utterances, test utterances, and SASV labels. However, ASV spoof 2019 LA train set provides only the utterances, speaker identities, and spoof keys. Therefore, it was necessary to construct training pairs. We designed a training set based on four scenarios: bonafide target speakers, bonafide non-target speakers, spoofed target speakers, and spoofed non-target speakers. The ratios for each scenario are 0.45, 0.25, 0.15, and 0.15.We use 2000 samples per epoch.

4. Proposal2: Integrated embedding projector

We designed the back-end model called integrated embedding projector (IEP) that transforms SV and CM embeddings into SASV embeddings using a metric learning. As shown in figure 2, IEP consists of two modules, f and g, and the structure of each module is described in table 1. The feedforward process of IEP is as follows:

$$z = g(f(x, y), x, y), \tag{4}$$

where z denotes the output of the IEP (SASV embedding) and x, y denote the SV, CM embeddings. We iteratively fed x and y embedding to the g model to reaggregate information of each task that can be distorted during the embedding fusion process. The IEP model is trained using cosine similarity-based triplet loss [5] to explore the embedding space for SASV. Considering the characteristics of the SASV task, we construct a triplet for the training as follows:

Table 1: *Description of the structure of the modules used in the proposed framework. The left column denotes the modules of SVSF and MLPSF, and the right column shows the modules of IEP. SVSF consists of the two embedding fusion blocks ($u1$, $u2$) and a score calculation block (pj). ELU is the exponential linear unit activation function proposed in [4].*

Layer	Structure	Layer	Structure

u_1	FC(352 ×128)
	ELU
	FC(128 ×128)
	ELU
	FC(128 ×64)
	ELU
	FC(64 ×160)
		f	FC(352 ×256)
			ELU
			FC(256 ×256)
			ELU
			FC(256 ×128)
			ELU

u_2	FC(352 ×128)
	ELU
	FC(128 ×128)
	ELU
	FC(128 ×64)
	ELU
	FC(64 ×160)
		g	FC(480 ×128)

pj	FC(320 ×128)
	ELU
	FC(128 ×64)
	ELU
	FC(64 ×2)
			-

MLPSF	FC(2 or 3 ×16)
	ELU
	FC(16 ×16)
	ELU
	FC(16 ×2)
			-

- Anchor (A_i): i-th speaker's bonafide embeddings.
- Positive pair (P_i): i-th speaker's bonafide embeddings which are extracted from different utterances other than the anchor.
- Negative pair (N): i-th speaker's any spoof embeddings or other speaker's any bonafide embeddings.

Therefore, our proposed IEP is trained to optimize the triplet loss as follows:

$$\mathcal{L}_{TRIPLET} = \frac{1}{c}\sum_{i=1}^{c} max(0, cos(A_i, N) - cos(A_i, P_i) + m), \tag{5}$$

where, c is the number of triplets per single mini-batch and m is the margin, set to 0.5.

5. Results

In Table 2, we compared our models with the challenge baseline systems. The proposed MLPSF outperformed the baseline1, which just sums the scores. The MLPSF achieved EER of 0.72% for the evaluation protocol, which shows that our proposed method is effective. Moreover, when integrated with the

Table 2: *Experimental results (EER, %) of speaker verification (SV), anti-spoofing (SPF), and spoofing-aware speaker verification (SASV) tasks for the SASV 2022 challenge development and evaluation protocols. MLPSF indicates the proposed score fusion method using MLP, and SVSF denotes the SV model considering the spoofed utterances. Also, IEP means integrated embedding projector.*

	SV EER		SPF EER		SASV EER	
	Dev	Eval	Dev	Eval	Dev	Eval
Baseline1 [1]	32.88	35.32	**0.06**	0.67	13.07	19.31
Baseline2 [1]	12.87	11.48	0.13	0.78	4.85	6.37
MLPSF	1.3	0.96	0.17	0.44	**0.60**	0.72
MLPSF + SVSF	**1.11**	**0.73**	0.13	**0.43**	0.67	**0.56**
IEP	2.51	1.58	2.7	1.12	1.55	1.32

SVSF score, performance improved by 30%, achieving the EER of 0.56%. On the other hand, the proposed IEP showed an EER of 1.32%, which indicated poor performance than the MLPSF with SVSF, but improved performance compared to baseline2.

6. References

[1] 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.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” in *Interspeech2020*. International Speech Communication Association (ISCA), 2020, pp. 3830–3834.

[3] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” *arXiv preprint arXiv:2110.01200*, 2021.

[4] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” *arXiv preprint arXiv:1511.07289*, 2015.

[5] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification.” *Journal of machine learning research*, vol. 10, no. 2, 2009.