SASV Challenge 2022 System Description

Jin Woo Lee, Eungbeom Kim, Junghyun Koo, and Kyogu Lee

Team MARG

{jinwlee, eb.kim, dg22302, kglee}@snu.ac.kr

Abstract

Text-to-speech (TTS) and voice conversion (VC) studies are constantly improving to the extent where they can produce synthetic speech almost indistinguishable from bona fide human speech in terms of naturalness and similarity. Although automatic speaker verification (ASV) systems have also shown remarkable progress, the need for enhancing vigilance against synthetic speech attacks is unavoidable. In this work, we propose a simple yet effective spoofing aware speaker verification (SASV) methodology called representation selective self-distillation (RSSD), which selectively performs speaker verification based on self-distillation using spoof countermeasures and speaker embeddings. Evaluation performed with the SASV Challenge 2022 database shows 1.15% of SASV equal error rate (SASV-EER), with 1.41% of speaker verification EER (SV-EER) and 0.76% of spoof EER (SPF-EER). Our experimental results indicate that RSSD takes advantage of the state-of-the-art speaker verification and spoofing detection models along with time and memory efficiency as well.

Index Terms: speaker verification, speech anti-spoofing, spoofing aware speaker verification

1. Methods

The main objective of SASV is to ensure that the speaker verification system has reliable robustness against spoofing attacks. A solution to carrying out SASV is to ensemble pre-trained ASV and CM systems, where we can at least expect their original performance on each separate task. Considering synthetic speech as negative samples of speech verification, the SASV system can first filter out fake voices as non-targets through spoofing detection and then operate ASV system only for utterances judged to be bona fide. From this intuition, we propose RSSD, an effective method to perform SASV inspired by representation distillation [1] and self-distillation [2]. In order to construct the representation space which preserve speaker information and spoofing information simultaneously, we leveraged state-of-the-art speaker verification model ECAPA-TDNN[1] [3] and spoofing countermeasure model AASIST[2] [4].

1.1. Representation selective self-distillation

Our method consists of four modules: a pre-trained speaker verification (SV) network E, a pre-trained spoof countermeasure (CM) network C, a feature transformation layer f, and a gating operator g. In order to allow our system to selectively transform the representation through self distillation, we design a system that measures the similarity between two samples as follows.

$$S(E(x_e), g(f(C(x_t)), E(x_t))) \tag{1}$$

Subscripts $\cdot_e$ and $\cdot_t$ represent enrollment and test samples, respectively. We use cosine similarity $S(\cdot, \cdot)$ as a measure between two embeddings: one extracted from the bona fide enrollment speech using E, the other transformed by our network. Our network f transform the CM embedding of test speech x_t extracted by CM network C, to modulate the speaker embedding of x_t. Then, gate operator g outputs self-distilled representation using the two embeddings $f(C(x_t))$ and $E(x_t)$. Finally, we measure spoof-aware speaker similarity as equation (1).

1.2. Objective function

Our main objective function L_{total} consists of two loss terms

$$L_{\text{total}} = L_{\text{r-distill}} + L_{\text{spoof}}, \tag{2}$$

where $L_{\text{r-distill}}$ denotes representation self-distillation loss, and L_{spoof} denotes spoofing countermeasure loss. The representation self-distillation term optimizes the network to preserve the representation of bona fide speakers by pre-trained speaker verification model. On the other hand, the spoofing countermeasure loss guides our network to transform the representation of spoofed speakers away from the representation of the bona fide speakers. In other words, our network is trained to be fully aware of countermeasure representation, and to adaptively transform the speaker representation based to it.

1.2.1. Representation self-distillation loss

Different from traditional distillation that distill the information through probabilistic outputs from teacher networks [5], RSSD distills the information through representations from pre-trained speaker verification network in the case of bona fide inputs. The representation self-distillation loss $L_{\text{r-distill}}$ of bona-fide example $x^{(b)}$ is defined as

$$L_{\text{r-distill}} = -S(\mathbf{e}_t^{(b)}, g(f(\mathbf{c}_t^{(b)}), \mathbf{e}_t^{(b)})), \tag{3}$$

where $\mathbf{e}_t^{(b)} = E(x_t^{(b)})$ denotes representation from the pre-trained speaker verification network, and $\mathbf{c}_t^{(b)} = C(x_t^{(b)})$ denotes representation from the pre-trained countermeasure network. The feature transformation layer $f(\cdot)$ is optimized to preserve the bona fide speaker representation from the pre-trained network.

The transformed countermeasure representation $f(C(\cdot))$ determines whether to preserve speaker representation $E(\cdot)$ using the gating operator g or not. That is, g decides to keep input speaker representation itself up when countermeasure representation contains bona fide feature.

1.2.2. Spoofing countermeasure loss

The spoofing loss L_{spoof} of spoof example $x^{(s)}$ with enrollment example $x^{(e)}$ is defined as

$$L_{\text{spoof}} = S(\mathbf{e}_e^{(b)}, g(f(\mathbf{c}_t^{(s)}), \mathbf{e}_t^{(s)})) \tag{4}$$

[1] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[2] https://github.com/clovaai/aasist

Table 1: *Comparison of various EERs for ASVspoof 2019 LA database.* `dev` *and* `eval` *denotes the development and evaluation sets of the database, respectively.*

System	SV-EER		SPF-EER		SASV-EER	
	`dev`	`eval`	`dev`	`eval`	`dev`	`eval`
Baseline 1	32.88	35.32	0.06	0.67	13.07	19.31
Baseline 2	12.87	11.48	0.13	0.78	4.85	6.37
Ours	**2.01**	**1.41**	0.17	0.76	**1.16**	**1.15**

In contrast to bona fide examples, the desired representation space should not distill the information of given spoof speaker representations so that transformed representation $g(E(x_s), f(C(x_s)))$ drifts apart from original enrollment speaker representation $E(x_e)$. The feature transformation layer $f(\cdot)$ changes spoof countermeasure representation $C(x_s)$ to displace spoof speaker representation far from enrollment speaker representation.

2. Experiments

2.1. Implementation details

To train RSSD, we used ASVspoof 2019 LA database [6], which is a standard database designed under consideration of logical attack scenario in ASVspoof 2019 challenge. It consists of bona fide speeches collected from VCTK corpus [7], and spoofed speeches generated using variety of TTS and VC methods. The training and development sets of ASVspoof 2019 LA database were constructed using 6 different algorithms, while the evaluation set consist of non-overlaping 13 different methods [6]. The training, development, and evaluation datasets consists of 20, 20, and 67 number of speakers, respectively. For more specific details on the experimental dataset, we refer to the SASV challenge protocols [8].

For the feature transformation layer f, we adopted 2 layer fully connected network with leakly ReLU activation function. The batch size is 32 and we used Adam optimizer [9] with learning rate 0.0001. We select the best model on development set for 20 epochs. We adopt two methods presented in the SASV challenge [8] as our baselines.

2.2. Evaluation metric

EER is a widely used measure to evaluate the performance of binary classification networks. It is used to show how well an ASV system can distinguish between target and non-target signals, but it is also used to show how well a CM system can distinguish between bona fide and spoofed signals. To evaluate the spoof-aware speaker verification performance of our model, we used a measure called SASV-EER which is a combination of the above two EERs: we show how well our system can distinguish "bonafide *and* target" signals from spoofed or (zero-effort) non-target ones. In addition to SASV-EER, which is the main evaluation measure in our study, we also analyzed the EER of the ASV systems (denoted by SV-EER) and the EER of the CM systems (denoted by SPF-EER).

2.3. Results

Empirical results of Table 1 showed the proposed method outperforms the SASV challenge baselines, resulting in 1.15% of SASV-EER. Our experimental results showed that RSSD effectively leveraged the state-of-the-art ASV and CM systems, along with efficiency as well. Figure 1 shows comparison of SPF-EERs for various spoofing attack types within ASVspoof 2019 LA evaluation set. Figure 2 shows score distribution of the proposed method for the development and evaluation sets.

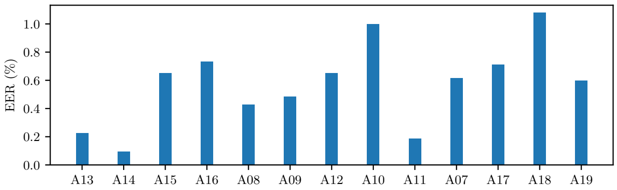

Figure 1: *Comparison of SPF-EERs for different attack types of ASVspoof 2019 LA database test set.*

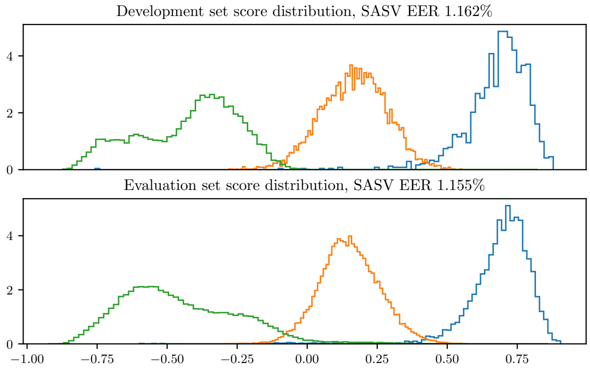

Figure 2: *Comparison of score distribution and corresponding SASV EERs for ASVspoof 2019 LA database. Blue, orange, and green lines represent score distributions of bona fide, zero-effort and spoofed samples, respectively.*

3. References

[1] Y. Tian, D. Krishnan, and P. Isola, “Contrastive representation distillation,” *arXiv preprint arXiv:1910.10699*, 2019.

[2] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be your own teacher: Improve the performance of convolutional neural networks via self distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3713–3722.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.

[4] 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.

[5] G. Hinton, O. Vinyals, J. Dean *et al.*, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, “Asvspoof 2019: Future horizons in spoofed and fake audio detection,” *arXiv preprint arXiv:1904.05441*, 2019.

[7] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” 2019.

[8] 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.

[9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.