Spoofing-aware Attention Back-end with Multiple Enrollment and Novel Trials Sampling Strategy for SASVC 2022

Chang Zeng[1,2], *Lin Zhang*[1,2], *Meng Liu*[3], *Junichi Yamagishi*[1,2]

[1]National Institute of Informatics, Japan
[2]SOKENDAI, Japan [3]Tianjin University, China

{zengchang,zhanglin,jyamagis}@nii.ac.jp, liumeng2017@tju.edu.com

Abstract

The spoofing aware speaker verification challenge (SASVC) 2022 has been organized to explore the relation between automatic speaker verification (ASV) and spoof countermeasure (CM). In this paper, we will introduce our proposed spoofing-aware attention back-end developed for SASVC 2022. First, we design a novel sampling strategy for simulating real verification scenario. Then, in order to fully leverage information derived from multiple enrollments, a spoofing-aware attention back-end has been proposed. Finally, a joint decision strategy is aggregated to introduce mutual interaction between ASV module and CM module. Compared with the trial sampling method used in baseline systems, our proposed sampling method shows effective improvement without any attention modules. The experimental result shows our proposed spoofing-aware attention back-end improves the performance from 6.37% of best baseline system on evaluation dataset to 1.19% in term of SASV-EER (equal error rate) metric.

Index Terms: speaker verification, spoofing aware, attention, multiple enrollment

1. Introduction

Automatic speaker verification (ASV) aims to tell whether the test utterance comes from the claimed speaker who has registered. However, the ASV system is also vulnerable to several conditions: (1) from the environment site, the performance of ASV can be badly affected by background noise, transmission channel, data quality, etc. [1, 2]. (2) from the speech itself site, the test utterance can be spoofing impostor and generated by text-to-speech (TTS), voice conversion (VC), replayed, or DeepFakes [3, 4, 5, 6]. Although technologies has been largely improved for above two sites, most researchers only focus on building independent ASV or countermeasure (CM) without interaction.

Thus, the Spoofing Aware ASV Challenge 2022 (SASVC 2022) [7] is organized to promote the development of models that can ensemble ASV and CM and protect systems from both non-target and spoofing impostor. To achieve this, it is essential to build an ensemble model to integrate ASV and CM systems, or a single model that can detect non-target and spoofing impostor simultaneously.

Several existing studies has been discussed to develop a spoofing aware speaker verification (SASV) system [8, 9, 10, 11] from above two directions: (1) Ensembled model: [10] proposed a novel ensemble back-end model to integrate the decision of CM and decision of ASV systems; (2) Single model: [11] optimized a neural network model to minimize both ASV loss and CM loss simultaneously in multi-task learning style. [12] promotes ASV and CM through Reinforcement Learning (RL).

We can notice that the above mentioned single model is time consuming, and sometimes the models need to be trained from scratch. Thus, we focus on building an ensemble back-end model to integrate ASV and CM embeddings. This back-end model is a box-out method which can be flexibly utilized in any pre-trained ASV and CM models.

In our previous work [13], we have conducted an attention back-end model to make full use of multiple enrollment utterances. But it only focuses on ASV to distinguish non-target speaker. In this paper, we extend this attention back-end for the SASV scenario. In this new ensemble spoofing-aware back-end model, we introduce a new branch with trainable parameters for CM embedding, and propose a novel trial sampling method with considering on two different impostor cases. Concretely, the spoofing-aware attention back-end can individually generate two scores for ASV and CM sub-tasks simultaneously. For ASV score, it is generated in a same way in [13] except using the different trial sampling method, which will be described in Section 3 in detail. For CM score, it is obtained by a simple linear transformation followed by a sigmoid function on CM embeddings. These two scores are concatenated as a 2-dimensional vector and projected to a probability for the final decision, *where mutual interaction between ASV and CM is introduced in training stage of this model*. Finally, our proposed attention back-end model can achieve 1.19 % in term of SASV-EER metric for the evaluation set.

The rest of this paper is organized as below. In Section 2, two baseline systems provided by the organizers of this challenge and evaluation metrics are described in detail. The proposed spoofing-aware attention back-end with score-level fusion as well as the novel trial sampling strategy are illustrated in Section 3. Experimental result will be reported in Section 4. And this paper is concluded in Section 5.

2. Baseline systems and evaluation metrics

Two baseline systems are provided by SASVC 2022. Both of them operate on the embeddings from pretrained models including ECAPA-TDNN [14] model for ASV task and AASIST [15] model for CM task.

2.1. Score fusion baseline

Baseline1 simply sums scores generated by the separate pre-trained ECAPA-TDNN model and AASIST model. Thus, no data is used for this baseline as it does not involve any training nor fine-tuning.

2.2. Embedding fusion baseline

Baseline2 involves the fusion of three types of embeddings: one extracted from an ASV enrollment utterance using the ECAPA-

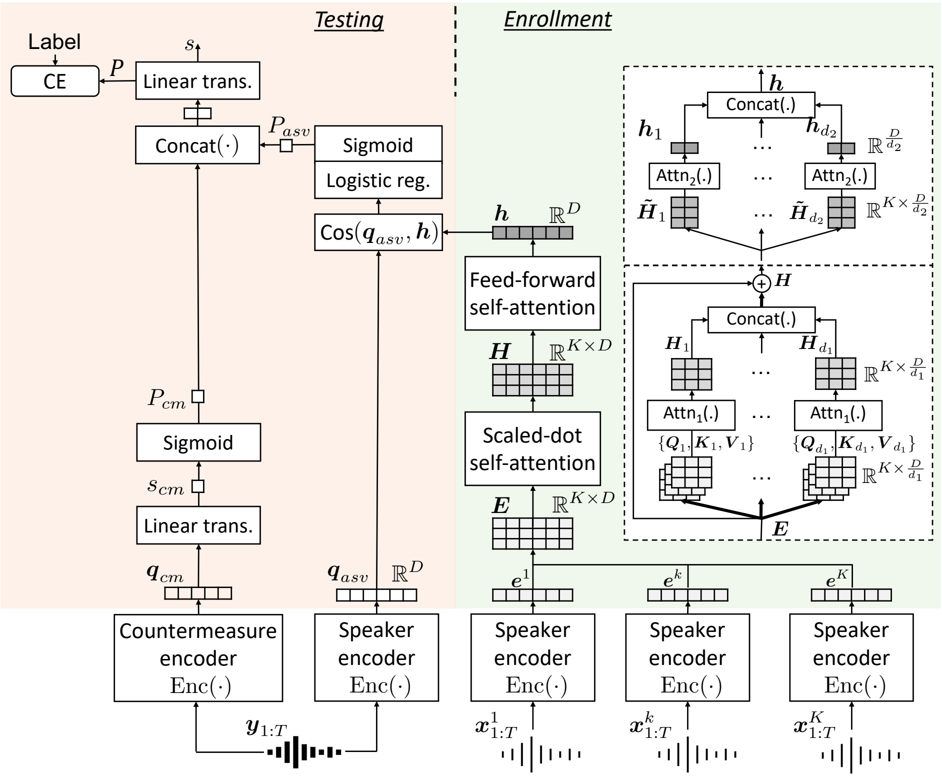

Figure 1: *Architecture of extended attention back-end with score-level fusion.*

TDNN system; the second extracted in an identical fashion from a test utterance; the third extracted from the same test utterance using the AASIST spoofing CM system. The model is a vanilla multi-layer perception with three hidden layers, trained using the ASVspoof 2019 LA train partition [7].

2.3. Evaluation metrics

As for evaluation metric, SASV performance is assessed using the SASV-EER as the primary metric without distinguish between different speaker and spoofed access attempts [7]. Besides, SV-EER and SPF-EER are also used to evaluate the performance of ASV and CM sub-tasks respectively.

3. Spoofing-aware attention back-end

Neural attention back-end [13] was proposed to handle the case of enrollment with multiple utterances reasonably, which is common in realistic scenario. But this model does not consider the condition where the testing utterance is spoofed by speech synthesis or voice conversion. So we extend it to spoofing-aware version for SASV scenario by introducing a CM branch as well as a novel sampling strategy. Note that the proposed spoofing-aware attention back-end is independent with frozen pre-trained models, and only the attention-based back-end are updated during training. Additionally, the sampling strategy can be flexibly applied for any other ensemble back-end model trained on trial pairs. Detail of the back-end module and sampling strategy are described in this section.

3.1. Model architecture

The detailed architecture of the proposed spoofing-aware attention back-end for SASVC is illustrated by Figure 1. Suppose a speaker has K enrollment utterances $\{\boldsymbol{x}^1_{1:T}, \cdots, \boldsymbol{x}^K_{1:T}\}$ and one test utterance $\boldsymbol{y}_{1:T}$. K enrollment ASV embeddings $\{\boldsymbol{e}^1_{asv}, \cdots, \boldsymbol{e}^K_{asv}\}$, one test ASV embedding $\boldsymbol{q}_{asv}$ and one CM embedding $\boldsymbol{q}_{cm}$ can be derived from the pre-trained ASV and CM models respectively. Those embeddings are treated as input for our proposed ensemble back-end. Within the back-end, K enrollment ASV embeddings $\{\boldsymbol{e}^1_{asv}, \cdots, \boldsymbol{e}^K_{asv}\}$ are firstly converted to a single speaker representative vector $\boldsymbol{h}$ through scaled-dot self-attention (SDSA) [16] and feed-forward self-attention (FFSA) [17, 18] modules step by step as the description in [13]. Then, this back-end module will produce probability scores based on the mentioned embeddings $\boldsymbol{q}_{cm}$, $\boldsymbol{q}_{asv}$ as well as $\boldsymbol{h}$ for two different hypothesises respectively.

One hypothesis is that whether the utterance $\boldsymbol{y}_{1:T}$ is bonafide or not, which can be formulated by the following equation:

$$P_{cm}(s_{cm}) = \frac{1}{1 + \exp^{-s_{cm}}}, \tag{1}$$

where P_{cm} is the CM probability how likely the input can be bonafide. CM score s_{cm} is transformed from the CM embedding $\boldsymbol{q}_{cm}$ by a linear transformation layer as shown in Figure 1.

The other hypothesis is whether $\boldsymbol{y}_{1:T}$ is uttered by the speaker who enrolled his or her identity with $\{\boldsymbol{x}^1_{1:T}, \cdots, \boldsymbol{x}^K_{1:T}\}$ utterances. As these multiple enrollment utterances has been transformed into a single speaker representative vector $\boldsymbol{h}$ by the pretrained ASV model and attention modules in our back-end model, the probability of this hypothesis can be calculated by the following formulas:

$$\begin{aligned} P_{asv}(\boldsymbol{q}_{asv}, \boldsymbol{h}) &= \frac{1}{1+\exp^{-s_{asv}}} \\ &= \frac{1}{1+\exp^{-a\,\mathrm{Cos}(\boldsymbol{q}_{asv},\boldsymbol{h})-b}}, \end{aligned} \tag{2}$$

$$\mathrm{Cos}(\boldsymbol{q}_{asv}, \boldsymbol{h}) = \frac{\boldsymbol{q}_{asv} \cdot \boldsymbol{h}}{||\boldsymbol{q}_{asv}|| \cdot ||\boldsymbol{h}||}, \tag{3}$$

where $P_{asv}(\boldsymbol{q}_{asv}, \boldsymbol{h})$ denotes the probability of $\boldsymbol{q}_{asv}$ and $\boldsymbol{h}$ belonging to the same speaker, and a and b are trainable parameters.

Unlike baseline1, that apply linear combination on the probability of ASV and CM modules only in inference stage, our proposed spoofing-aware back-end module can benefits from the interaction between ASV module and CM module during training stage through backward propagation. For example, the spoofing information contained in the CM embedding may have impact on the enrollment process of multiple utterances. Therefore, the CM probability P_{cm} and the ASV probability which is denoted as P_{asv} in Figure 1 are concatenated as a 2-dimensional vector and it is projected to the final probability by a simple linear transformation as well as a sigmoid function as shown below.

$$\begin{aligned} P(P_{cm}, P_{asv}) &= \frac{1}{1+\exp^{-s}} \\ &= \frac{1}{1+\exp^{-(w_1 * P_{cm} + w_2 * P_{asv} + v)}}, \end{aligned} \tag{4}$$

where $P(P_{cm}, P_{asv})$ denotes the probability of a joint decision between CM (eq. 1) and ASV (eq. 2). s is the final score used for the decision making. w_1, w_2 and v are trainable parameters in the linear transformation.

3.2. Trials sampling strategy

To train the proposed back-end module for SASVC, a novel trials sampling method is proposed and demonstrated in this section. For each mini batch, assuming it has M speakers, each speaker has K ASV embeddings and K CM embeddings, the size of one mini batch is $M \times K$ for both speaker embeddings and countermeasure embeddings. In our experiment, M and K are set as 16 and 10, respectively. Considering the number of bonafide audios is limited in the training dataset, we control the number of bonafide audios and spoof audios for each speaker in one mini-batch to be equal, which means one speaker in a mini-batch has $K/2$ bonafide audios and $K/2$ spoof audios. When one mini batch of data is fed into the back-end module, ASV embeddings will be rearranged to form speaker verification trials which have multiple enrollment ASV embeddings.

Table 1 illustrates one example to form the pairs of *(test-speaker-embedding, enrollment-data)* for ASV sub-task. In this example, one mini batch of data has $M = 3$ speakers A, B, and C, and each speaker has $K = 4$ embeddings consisting of 2 bonafide audios and 2 spoof audios whose index ranges from 1 to 4. Index with underline indicate from bonafide audios. There are numerous ways to compose the pairs, but we only consider the following cases. For positive pairs of ASV sub-task where test and enrollment utterances are from the same speaker, one test speaker embedding is selected from the speaker's data, and the rest are left for enrollment. For negative pairs of ASV sub-task where the speaker of the test utterance is different from that of enrollment data, we only consider pairs marked by (test-speaker-embedding=✓, enroll=(×, ×, ×)) of other speakers included in a mini batch. And the the trials sampling method will mask the spoof enroll out by setting these embeddings as zero vectors following other attention mechanisms [16, 17, 18]. As for the label of triple of (test-CM-embedding, test-speaker-embedding, enrollment-data), an AND operation denoted as $\bigotimes$ between CM label of the test utterance and ASV label of the pair (test-speaker-embedding, enrollment-data) is conducted to determine the final label for training.

3.3. Loss function

Binary cross-entropy (BCE) is used as loss function to train the proposed extended attention back-end model. According to the above mentioned trials sampling method, let us respectively define $\boldsymbol{q}_{cm}^{lm}$ as a countermeasure embedding and $\boldsymbol{q}_{asv}^{lm}$ as a speaker embedding vector extracted from the m-th test trial of speaker l, $\boldsymbol{h}^{nm}$ is based on an enrollment set m as $\boldsymbol{h}$ shown in Figure 1, which contains multiple audio files uttered by speaker n [13]. Superscript l and n indicate the testing and enrollment speaker respectively. The BCE loss can be computed as below,

$$\begin{aligned} \mathcal{L}_{\text{bce}} = -\sum_{\forall l,m,n} & [\mathcal{I}\left(l=n\right) \log P(\boldsymbol{q}_{asv}^{lm}, \boldsymbol{h}^{nm}, \boldsymbol{q}_{cm}^{lm}) \\ & + \mathcal{I}\left(l \neq n\right) \log \left(1 - P\left(\boldsymbol{q}_{asv}^{lm}, \boldsymbol{h}^{nm}, \boldsymbol{q}_{cm}^{lm}\right)\right)], \end{aligned} \tag{5}$$

where $\mathcal{I}(\cdot)$ is an indicator function that returns one when its argument is true and zero otherwise.

Due to the proposed trials sampling method, the extremely unbalanced ratio of positive samples and negative samples in a mini-batch is equal or even less than $1 : (M - 1)$. To alleviate the impact of unbalanced mini-batch and emphasize the contribution of hard training samples [19], we select all positive samples and top H negative samples with large values to implement backward propagation.

4. Experiments

4.1. Datasets

In this challenge, the organizers provide VoxCeleb2 [20], ASVspoof 2019 LA [21] train partition and development partition as training data. Since our proposed model focuses on integrating ASV and CM information from a back-end view, we leverage speaker and CM embeddings extracted from the provided pretrained ECAPA-TDNN [14] and AASIST [15] models as the training data. The training data we used contains bonafide and spoofed utterances, with a total amount of more than 25,000 utterances from 20 speakers. The development set contains around 25,000 speaker and CM embeddings from the ASVspoof 2019 LA development partition. The ASVspoof 2019 LA evaluation partition is used for evaluated, with more than 70,000 utterances from 48 different speakers.

4.2. Training strategy

Before feeding data to the proposed model, training trials are sampled from training embeddings by using the above mentioned method. In our experimental setting, one mini-batch contains 16 speakers and each speaker has 5 bonafide embeddings and 5 spoofing embeddings. In order to optimize the model, SGD optimizer with 0.0001 learning rate, 0.9 momentum and 0.00001 weight decay is utilized to train the model for 40 epochs. The learning rate is decayed by 0.95 at the end of

Table 1: *Composition of triples of (test-CM-embedding, test-speaker-embedding, enrollment-data) for training back-end model and ground-truth labels from mini-batch. A, B, and C are speaker IDs, and 1, 2, 3 and 4 are his or her audio IDs (IDs with underline indicate the bona fide case). ✓ and × denote test and enrollment audio files, respectively.*

	A				*B*				*C*									
	$\underline{1}$	2	$\underline{3}$	4	1	$\underline{2}$	3	$\underline{4}$	1	2	$\underline{3}$	$\underline{4}$	Test	Enroll	ASV Label		CM Label	Label
Test to be used for training	✓	×	×	×									q_{A1}	h_{A1}	P	$\otimes$	P	P
	✓					×	×	×					q_{A1}	h_{B1}	N	$\otimes$	P	N
	✓									×	×	×	q_{A1}	h_{C1}	N	$\otimes$	P	N
	×	✓	×	×									q_{A2}	h_{A2}	P	$\otimes$	N	N
		✓			×		×	×					q_{A2}	h_{B2}	N	$\otimes$	N	N
		✓							×		×	×	q_{A2}	h_{C2}	N	$\otimes$	N	N
							⋮									⋮		
	×	×	×									✓	q_{C4}	h_{A4}	N	$\otimes$	P	N
					×	×	×					✓	q_{C4}	h_{B4}	N	$\otimes$	P	N
									×	×	×	✓	q_{C4}	h_{C4}	P	$\otimes$	P	P

Table 2: *The three different EERs (%) for the SASV 2022 development and evaluation partitions. SASV-EER for all systems are calculated using the entire protocol that includes trials used to measure the SV-EER (target vs. non-target) and those used to measure the SPF-EER (target vs. spoof). Results shown for a conventional ASV system (ECAPA-TDNN), the two baseline solutions as well as our proposed extended attention back-end with score-level fusion.*

	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
ECAPA-TDNN	1.88	1.63	20.30	30.75	17.38	23.83
Baseline1	32.88	35.32	0.06	0.67	13.07	19.31
Baseline2	12.87	11.48	0.13	0.78	4.85	6.37
Proposed	1.41	1.32	0.61	1.14	0.81	1.19

each epoch. In addition, due to the heavily unbalanced positive and negative trials within one mini-batch described in Section 3.2 and Section 3.3, only the top 100 largest loss values of negative trials and all positive trials are selected for backward propagation.

4.3. Results and analysis

Table 2 demonstrates our best SV-EER, SPF-EER and SASV-EER performance of each system. The first line of this table shows how vulnerable the SOTA ECAPA-TDNN ASV model is when attacked by spoofing data. The EER of ECAPA-TDNN on eval dataset has been heavily degraded from 1.63% to 23.83%. Compared the result of baseline1 with that of ECAPA-TDNN, just simply introducing CM information in inference stage can alleviate the vulnerability of ASV model. However, due to no learnable parameter in baseline1 model, the result is still unacceptable compared with the case of no spoofing attacks. Different with baseline1, baseline2 trained a vanilla MLP model to solve this binary classification problem and the result is largely improved compared with baseline1. As for the result of score-level fusion attention back-end on the fourth line, due to the interaction between ASV and CM modules in training stage, both metrics of SASV-EER and SV-EER far surpass baseline2 system more than 80% relatively. Another SPF-EER metric is slightly worse than baseline2.

4.4. Ablation analysis

In order to verify the effectiveness of our proposed trial sampling method, a simple ablation study was conducted and its result is shown in Table 3. Compared with the spoofing-aware attention back-end, all attention modules are substituted by a simply average operation to aggregate multiple enrollment speaker embeddings into a single speaker representative vector which is equivalent to $\boldsymbol{h}$ in Figure 1. The result in the second line of Table 3 shows even if the model only has a few learnable parameters without using any attention mechanism, it can realize a surprising result as long as feeding the training trials sampled by the proposed sampling method to the model. On the contrary, baseline2 adopted a random sampling method [7] which reults in much worse result even if it used a vanilla MLP model with 3 layers, whose number of parameters is far beyond the average model. Additionally, when comparing the second line and the third line, it proves that these attention modules in our proposed model is necessary to improve the performance further.

Table 3: *The three different EERs (%) for ablation study of the proposed trial sampling method. Average means simply average enrollment speaker embeddings to a single speaker representative vector. Attention denotes the system uses attention mechanisms to aggregate multiple enrollment embeddings.*

	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
Baseline2	12.87	11.48	0.13	0.78	4.85	6.37
Average	2.09	1.97	0.07	0.76	1.15	1.53
Attention	1.41	1.32	0.61	1.14	0.81	1.19

5. Conclusions

In this paper we proposed a spoofing-aware attention back-end model for the SASV task. In addition, we also designed a novel and general sampling strategy for the model which is trained based upon trial-pairs. On the ASVspoof 19 evaluation partition data, our model outperformed the best baseline system provided by the organizers of SASVC more than 80% relatively. And the ablation analysis also shows the effectiveness of our proposed method.

Acknowledgements This study is partially supported by JST CREST Grants (JPMJCR18A6 and JPMJCR20D3), MEXT KAKENHI Grants (21K17775, 21H04906, 21K11951, 18H04112), and Google AI for Japan program.

6. References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.

[4] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[5] 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.

[6] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[7] 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.

[8] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[9] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2020.

[10] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*. ISCA, 2018.

[11] J. Li, M. Sun, X. Zhang, and Y. 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.

[12] A. Kanervisto, V. Hautamäki, T. Kinnunen, and J. Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, p. 477–488, jan 2022. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3138681

[13] C. Zeng, X. Wang, E. Cooper, X. Miao, and J. Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," *arXiv preprint arXiv:2104.01541*, 2021.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[15] 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. ICASSP*, 2022.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[17] C. Raffel and D. P. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv preprint arXiv:1512.08756*, 2015.

[18] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[19] R. Li, N. Li, D. Tuo, M. Yu, D. Su, and D. Yu, "Boundary discriminative large margin cosine loss for text-independent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6321–6325.

[20] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[21] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech and Language*, vol. 64, p. 101114, 2020.