IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 17, 2019, accepted January 2, 2020, date of publication January 6, 2020, date of current version January 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964048

Joint Decision of Anti-Spoofing and Automatic Speaker Verification by Multi-Task Learning With Contrastive Loss

JIAKANG LI[1], MENG SUN[1], XIONGWEI ZHANG[1], AND YIMIN WANG[2]

[1]Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing 210007, China
[2]Communications Engineering College, Army Engineering University, Nanjing 210007, China

Corresponding authors: Meng Sun (sunmengccjs@gmail.com) and Xiongwei Zhang (xwzhang9898@163.com)

This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180080, and in part by the National Natural Science Foundation of China under Grant 61471394.

ABSTRACT Automatic speaker verification (ASV) is an emerging biometric verification technique with more and more applications. However, both verification accuracy and anti-spoofing should be considered carefully before putting ASV into practice, where anti-spoofing is also called replay detection in which voice is recorded, stored and replayed to deceive ASV systems. Cascaded decision of anti-spoofing and ASV is a straightforward solution to tackle the two issues. In this paper, joint decision of anti-spoofing and ASV was investigated in a multi-task learning framework with contrastive loss in order to improve the cascaded decision approach. A modified triplet loss was firstly constructed to supervise deep neural networks to extract embedding vectors containing information of both speaker identity and spoofing. The embedding vectors were subsequently taken as input features by back-end classifiers towards speaker and spoofing classification. The experimental results on both ASVspoof 2017 and ASVspoof 2019 showed that the proposed joint decision approach with triplet loss outperformed the corresponding baselines, a recent work on joint decision with Gaussian back-end fusion and our previous joint decision approach with cross-entropy loss.

INDEX TERMS Anti-spoofing, speaker verification, replay detection, multi-task learning, triplet loss.

I. INTRODUCTION

With the development of engineering applications of artificial intelligence, biometric authentication is becoming popular in scenario of protecting the security of computers, smart devices, and networks, such as fingerprint and face recognition. Voiceprint is an emerging biometric with potential advantages given its hands-free, liveliness and dynamic nature. Automatic speaker verification (ASV) is a conventional way to put voiceprint into practical usage, where it verifies the claimed identity of a speaker by recording voices, extracting voiceprints and computing similarities.

However, spoofing is a great threaten to the safety of biometric authentication, which is attributed to the biometrics can be copied [1]. Among spoofing attacks, the most accessible ones are attacks at the sensor and transmission levels [2], which are called physical access (PA), e.g. attacking a face recognition system by showing a photo of an authenticated user to the camera, or attacking an ASV system by playing back a recording of a verified user [3], [4]. The replayed voices are not only from recordings but also generated by state-of-the-art professional tools of text-to-speech synthesis (TTS) and voice conversion (VC). Since replay attacks are easy to implement and highly similar to bona fide speech, it is difficult to detect and bring serious threats to ASV systems [5]. Therefore, anti-spoofing should be considered carefully before putting ASV into practical usage.

In recent years, many works have been done to study anti-spoofing problems, among which the automatic speaker verification spoofing and countermeasures challenge (ASVspoof) is the most comprehensive one. ASVspoof 2015 focused on the discrimination between bona fide speech and voices generated by TTS or VC [6]. ASVspoof 2017 focused on the detection of PA attacks to discriminate whether the given speech was the voice of an in-person human or the replay of a recorded

The associate editor coordinating the review of this manuscript and approving it for publication was Huawei Chen.

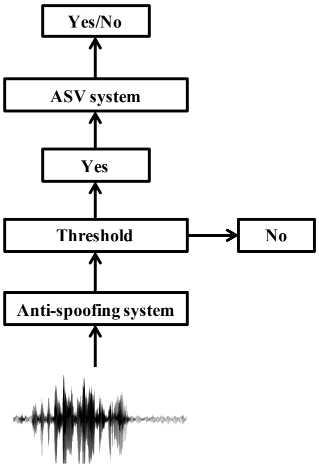


FIGURE 1. A cascaded system of anti-spoofing and ASV.

speech [7]. ASVspoof 2019 considered tasks from both ASVspoof 2015 and ASVspoof 2017 [8]. The challenges provided us extensive data to make thorough comparison and evaluation.

A large number of anti-spoofing methods have been proposed and have achieved quite good results in ASVspoof challenge these years. By introducing deep learning into anti-spoofing, deep neural networks (DNN) have achieved promising results in anti-spoofing of ASVspoof 2017 [9]–[11] and ASVspoof 2019 [12]. ASV as a standalone task has also gained great improvement from deep learning [13], [14]. Given those achievements and in order to make ASV and anti-spoofing a step forward to practical usage, some early studies have proposed that a separately designed anti-spoofing system is implement before ASV, only the utterances which have passed spoofing detection are verified again by a ASV system [4], [14], which is in fact a *cascaded* structure as is illustrated in FIGURE 1.

Though it is straightforward to cascade two classifiers to accomplish the tasks of ASV and anti-spoofing, redundant computation is actually introduced when analyzing the input voices twice, i.e. one for anti-spoofing and the other for ASV. Furthermore, possible mutual enhancement of the two tasks cannot be explored when treating them separately, which seems not elegant in the point of machine learning. Therefore, a joint decision system, which is able to conduct ASV and anti-spoofing simultaneously, is elegant in theory and concise in practice, as is depicted in FIGURE 2. Recently, a joint ASV and anti-spoofing system was studied in an i-vector framework by A. Sizov, et al. [2]. Promising results were obtained on the dataset of NIST 2006, which demonstrated the feasibility and advantage of the joint decision of ASV and anti-spoofing, which is one of the motivation of this paper.

In light of the success of deep learning on both ASV [15] and anti-spoofing [2], a joint ASV and anti-spoofing system with deep learning has been proposed and studied in our previous work [16]. The joint system has shown its effectiveness by obtaining better results than both the cascaded system with Gaussian back-end fusion in [17] and the joint

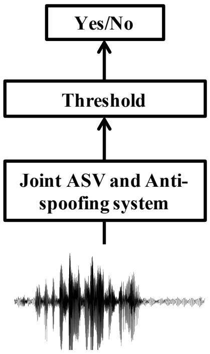


FIGURE 2. A joint system of anti-spoofing and ASV.

system with i-vector features [2]. In this paper, inspired by the great improvement of contrastive loss on speaker verification accuracy in [18], triplet loss [19] is introduced to replace the conventional cross-entropy loss on classification tasks. As far as we know, this is the first time that contrastive loss is investigated on anti-spoofing tasks as well as on the joint ASV and anti-spoofing tasks.

By contrasting FIGURE 1 and 2, it is straightforward to see there are at least two obvious advantages of the proposed joint decision approach of ASV and anti-spoofing by using deep learning with triplet loss. Firstly, only one decision procedure is required which reduces the number of thresholds to make subjective decisions on yes or no. Secondly, speakers' voices are analyzed only once to extract time-frequency features as inputs of classifiers, which reduce the computational complexity. Moreover, the two tasks may learn from each other to perform mutual enhancement as observed in many other multi-task learning works [20]. By introducing triplet loss to maximize the margin between bona fide target speaker and the spoofing or non-target ones, discriminative and representative features on both speaker identification and spoofing classification would be extracted.

The remaining part of the paper is organized as follows. The proposed joint anti-spoofing and ASV system based on triplet loss are presented in Section II, where the necessary parts of the deep learning baseline with cross-entropy loss in our previous work is also introduced. In Section III, the implementation of the proposed joint decision approach is described. The experimental results and their discussion are given in Section IV. Section V is the conclusion of the paper.

II. JOINT DECISION ON ANTI-SPOOFING AND ASV

A. HYPOTHESES

As a biometric, every spoken utterance has two attributes, one of which is the speaker identity, and the other of which is whether the audio is bona fide or spoofing. In order to define the decision problem formally, let observation O denotes the feature sequence of an utterance from speaker s, χ the target speaker and η the bona fide speech. The decision on ASV and anti-spoofing is thus equivalent to a hypothesis testing problem where the null hypothesis $H_{(\chi,\eta)}$ represents that O is bona fide speech from the target speaker s and the alternative

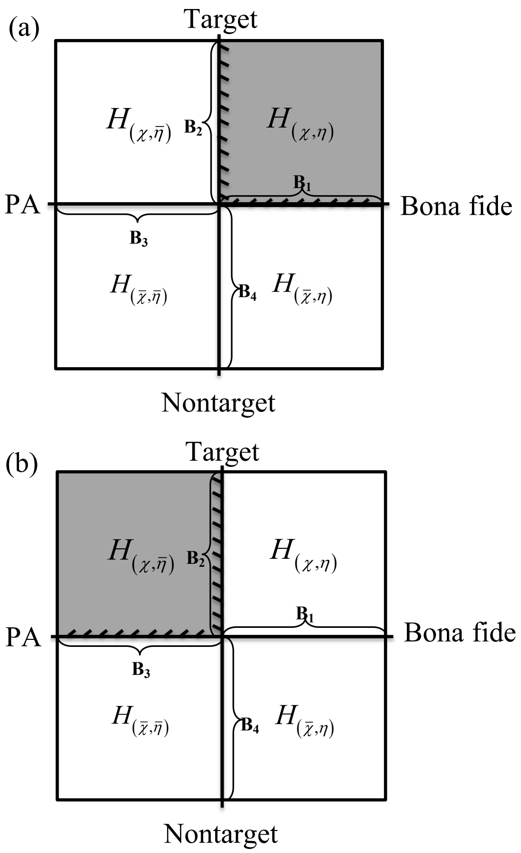


FIGURE 3. Illustration of the four hypotheses and two configurations of joint decision problems.

hypothesis $H_{\overline{(\chi,\eta)}}$ is the union of three choices,

$$H_{\overline{(\chi,\eta)}} = H_{(\bar\chi,\eta)} \cup H_{(\chi,\bar\eta)} \cup H_{(\bar\chi,\bar\eta)}, \tag{1}$$

where $(\bar\chi, \eta)$ represents bona fide speech from a non-target speaker, $(\chi, \bar\eta)$ spoofed speech from the target speaker, and $(\bar\chi, \bar\eta)$ spoofed speech from a non-target speaker.

ASV aims at optimizing the boundary $B_1 + B_3$ while anti-spoofing tries to find the boundary $B_2 + B_4$ in FIGURE 3. However, the joint decision on ASV and anti-spoofing boils down to find the classification boundary $B_1 + B_2$ as shown in (a) of FIGURE 3. Towards this goal, mapping function from the observed feature sequence O to $H_{(\chi,\eta)}$ and $H_{\overline{(\chi,\eta)}}$ should be constructed directly. Considering the imbalance of the number of samples between $H_{(\chi,\eta)}$ and $H_{\overline{(\chi,\eta)}}$, a mirror joint decision problem to distinguish $(\chi, \bar\eta)$ and the union of (χ, η),$(\bar\chi, \eta)$, and $(\bar\chi, \bar\eta)$ is also configured for training purpose, which would strengthen the learning performance of B_2.

By taking deep convolutional or neural networks as mapping functions, the extracted vectors at their final layers should contain information of both speaker identity and spoofing, multi-task learning is thus a natural choice to fulfill this requirement. In subsection B, multi-task learning with cross-entropy loss will be introduced firstly together with its joint decision procedures. Triplet loss which is able to maximize the margin between $H_{(\chi,\eta)}$ and $H_{\overline{(\chi,\eta)}}$ in (a) of FIGURE 3 (or $H_{(\chi,\bar\eta)}$ and $H_{(\chi,\eta)} \cup H_{(\bar\chi,\eta)} \cup H_{(\bar\chi,\bar\eta)}$ in (b) of FIGURE 3) will be utilized to replace the cross-entropy loss in subsection C.

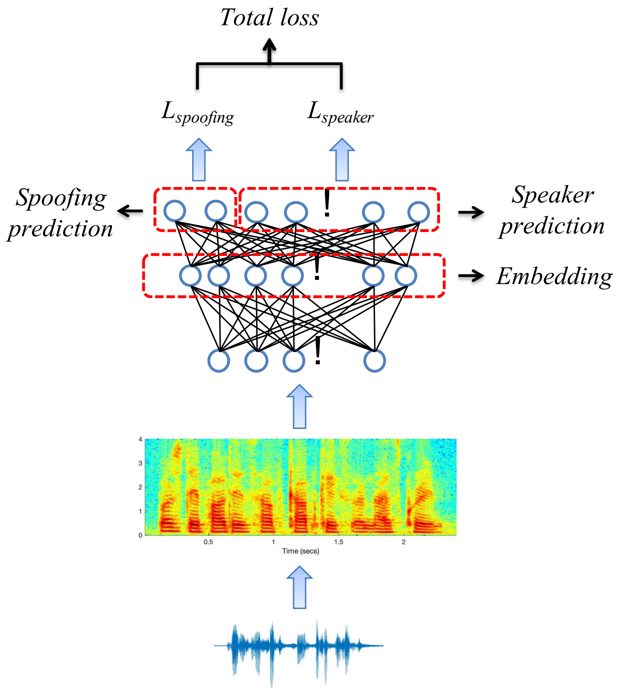


FIGURE 4. A multi-task learning network to extract embedding vectors from speech that contain information of both speaker identity and spoofing.

B. JOINT DECISION BY MULTI-TASK LEARNING WITH CROSS-ENTROPY LOSS

Towards the goal of joint decision, a unified network is designed to extract embedding vectors that contain information of both speaker identity and spoofing, which is the network performs two tasks in a single learning framework. Cross-entropy is a commonly utilized loss for classification, where softmax is usually taken as the activation of the final layer to yield exclusive activation probabilities. Therefore, cross-entropy with softmax is chosen as loss functions on both tasks. The cross-entropy loss for ASV task is,

$$L_s = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{w_{y_i}^T X_i}}{\sum_{j=1}^{C} e^{w_j^T X_i}}, \tag{2}$$

where N is the number of training utterances, C is the number of speakers, w_j is the weight vector corresponding to the j-th category, X_i is the embedding vector of the i-th utterance, y_i is the true label index of the i-th utterance. In anti-spoofing task, the cross-entropy loss is basically the same as (2) by just replacing the speaker labels by spoofing labels and $C = 2$.

Let $L_{speaker}$ represents the cross-entropy loss for ASV, and $L_{spoofing}$ represents the cross-entropy loss for anti-spoofing.

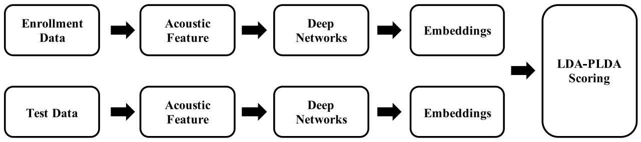


FIGURE 5. The enrollment and testing stages of ASV and anti-spoofing.

The total loss for the multi-task learning is hereby,

$$L_{total} = \alpha_1 L_{speaker} + \alpha_2 L_{spoofing}, \tag{3}$$

where α_1 and α_2 represent the weight of $L_{speaker}$ and $L_{spoofing}$ in the total loss, and satisfy $\alpha_1 + \alpha_2 = 1$.The two cross-entropy losses on ASV and spoofing detection are fused to obtain an unified loss, which works together at the end of the last layer of deep convolutional or neural networks, and then the loss is propagated back to the entire network. In our experiments, we set $\alpha_1 = \alpha_2 = 0.5$, because the two subtasks have the same status in the entire task. Through training, embedding vectors would be obtained which preserve the information of both speaker identity and spoofing. In the network, voice signals are analyzed only once and only one embedding vector is extracted to represent both kinds of information. Therefore, the complexity is reduced in the multi-task learning scheme w.r.t. that of the learning of two separate tasks.

In the enrollment and testing stages, by putting the embedding vectors into linear discriminative analysis (LDA), feature dimension is further reduced. A probabilistic LDA (PLDA) is subsequently learned as the backend classifier to facilitate the speaker verification and anti-spoofing, as shown in FIGURE 5. For the detailed usage of embedding vectors and the adaptation in the enrollment of new speakers, please refer to [16] for more details.

C. JOINT DECISION BY MULTI-TASK LEARNING WITH TRIPLET LOSS

Contrastive loss has been studied extensively for speaker verification. With the learning or training with contrastive loss, embedding vectors of the utterances from the same speaker would be pushed together and embedding vectors of the utterances from different speakers would be allocated apart. Triplet loss is one of the conventional realizations of contrastive loss, which is defined over triplets of embedding vectors: an anchor sample (i. e. an utterance in this paper) u^a, a positive sample u^p sharing the same class label of the anchor, and a negative sample u^n holding a different class label from the anchor. The loss is thus derived by following the idea that u^n should be further away from the anchor u^a than u^p by some margin, that is to minimize,

$$\begin{aligned} &L(u^a, u^p, u^n) \\ &\quad = \max(\left\|f(u^a) - f(u^p)\right\|_2^2 - \left\|f(u^a) - f(u^n)\right\|_2^2 + margin, 0) \end{aligned} \tag{4}$$

where $f(u)$ denotes the embedding vectors of utterance u, $\|f(a) - f(b)\|_2^2$ the Euclidean distance between the two embedding vectors $f(a)$ and $f(b)$, and *margin* a constant.

However, for the multi-task learning of ASV and anti-spoofing, each utterance contains two kinds of labels, the speaker identity and spoofing or not. Therefore, the original triplet loss in (4) cannot be used for discrimination purpose directly. A modification on (4) should firstly be made to fit to the multi-task problem. Based on the hypotheses introduced in Section II A, two choices of creating triplets are given as follows.

1) Two bona fide utterances from the same speaker are randomly selected as anchor and positive samples, respectively. A spoofing utterance from the same speaker, or an utterance from a different speaker no matter which is bona fide or spoofing, is randomly selected as negative sample, as is shown in (a) of FIGURE 3.

2) Two spoofing utterances from the same speaker are randomly selected as anchor and positive samples, respectively. A bona fide utterance from the same speaker, or an utterance from a different speaker no matter which is bona fide or spoofing, is randomly selected as negative sample, as is shown in (b) of FIGURE 3.

By summarizing 1) and 2), the negative sample is always chosen from three sets. Different from the task in speaker verification that the triplet loss in speaker verification only needs to select an anchor, a positive type sample and a negative type sample each time, a triplet in multi-task learning of anti-spoofing and ASV has three types of negative samples. When selecting negative samples, it is necessary to select three types of negative samples at the same time, and calculate the distance between the three types of negative sample and one positive sample respectively to obtain the loss. Therefore, the triplet loss in (4) was updated to,

$$\begin{aligned} &L_m(u^a, u^p, u^n) \\ &\quad = L(u^a, u^p, u^{n_1}) + L(u^a, u^p, u^{n_2}) + L(u^a, u^p, u^{n_3}), \end{aligned} \tag{5}$$

where n_1, n_2 and n_3 denote the three kinds of negative samples. The derivatives in back propagation are subsequently given (6), as shown at the bottom of this page.

$$\begin{cases} \frac{\partial L_m}{\partial f(u^a)} = \frac{\partial L(u^a,u^p,u^{n_1})}{\partial f(u^a)} + \frac{\partial L(u^a,u^p,u^{n_2})}{\partial f(u^a)} + \frac{\partial L(u^a,u^p,u^{n_3})}{\partial f(u^a)} = 2f(u^{n_1}) + 2f(u^{n_2}) + 2f(u^{n_3}) - 6f(u^p) \\ \frac{\partial L_m}{\partial f(u^p)} = \frac{\partial L(u^a,u^p,u^{n_1})}{\partial f(u^p)} + \frac{\partial L(u^a,u^p,u^{n_2})}{\partial f(u^p)} + \frac{\partial L(u^a,u^p,u^{n_3})}{\partial f(u^p)} = 6(f(u^p) - f(u^a)) \\ \frac{\partial L_m}{\partial f(u^{n_1})} = \frac{\partial L(u^a,u^p,u^{n_1})}{\partial f(u^{n_1})} + \frac{\partial L(u^a,u^p,u^{n_2})}{\partial f(u^{n_1})} + \frac{\partial L(u^a,u^p,u^{n_3})}{\partial f(u^{n_1})} = 2(f(u^a) - f(u^{n_1})) \\ \frac{\partial L_m}{\partial f(u^{n_2})} = \frac{\partial L(u^a,u^p,u^{n_1})}{\partial f(u^{n_2})} + \frac{\partial L(u^a,u^p,u^{n_2})}{\partial f(u^{n_2})} + \frac{\partial L(u^a,u^p,u^{n_3})}{\partial f(u^{n_2})} = 2(f(u^a) - f(u^{n_2})) \\ \frac{\partial L_m}{\partial f(u^{n_3})} = \frac{\partial L(u^a,u^p,u^{n_1})}{\partial f(u^{n_3})} + \frac{\partial L(u^a,u^p,u^{n_2})}{\partial f(u^{n_3})} + \frac{\partial L(u^a,u^p,u^{n_3})}{\partial f(u^{n_3})} = 2(f(u^a) - f(u^{n_3})), \end{cases} \tag{6}$$

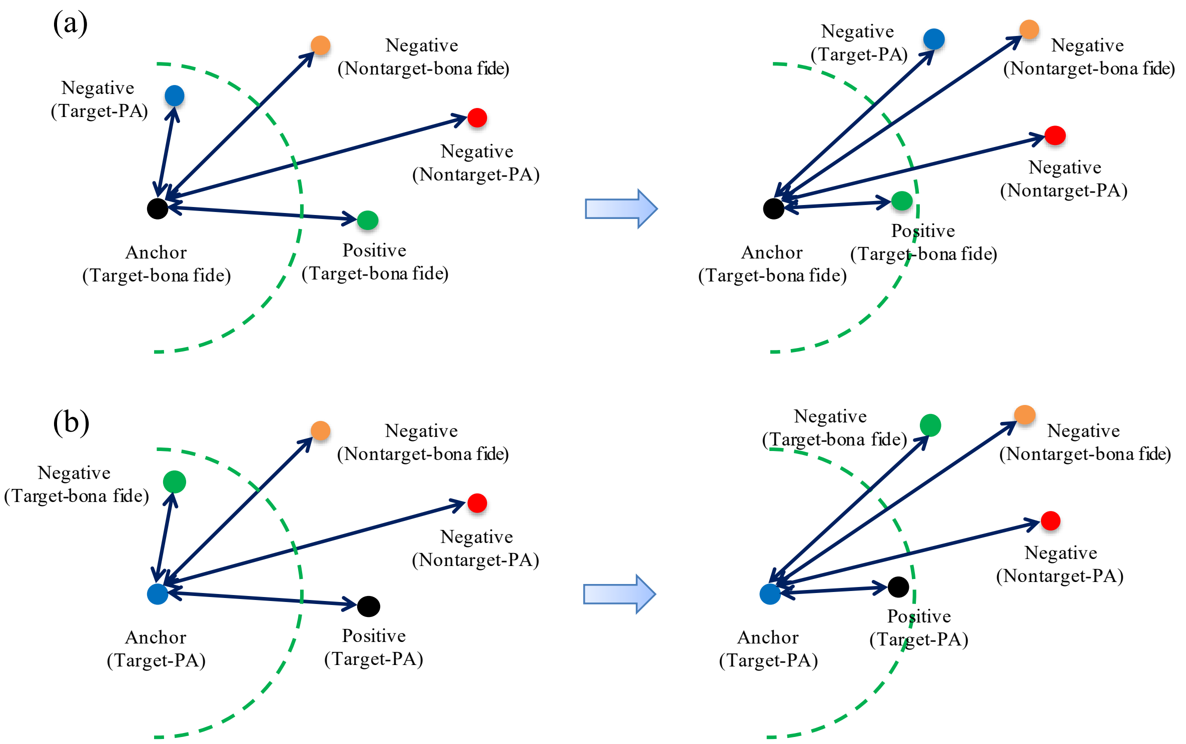


FIGURE 6. The role of triplet loss in training: pulling positive samples closer while pushing negative ones further. (a) and (b) correspond to subfigure (a) and (b) in FIGURE 3, respectively.

TABLE 1. Profile of ASVspoof 2017 Version 2.0.

Subset	# Speakers	# Utterances		
		Bona fide	Spoofed	Total
Training	10	1508	1508	3016
Development	8	760	950	1710
Evaluation	24	1298	12008	13306
Total	42	3566	14466	18032

TABLE 2. Profile of ASVspoof 2019 PA.

Subset	# Speakers	# Utterances		
		Bona fide	Spoofed	Total
Training	20	5400	48600	54000
Development	20	5400	24300	29700
Evaluation	48	18090	116640	134730
Total	88	28890	189540	218430

FIGURE 6 visually illustrated a set of triplet selections and the expected results through optimization. As for the training stage of joint ASV and anti-spoofing with triplet loss, since the triplet loss integrates the speaker information and the spoofing information into the same target, the embedding vectors obtained by deep neural/convolutional networks contain the information of both speaker identity and spoofing.

III. EXPERIMENTAL SETUP

A. DATASETS

Two datasets, ASVspoof 2017 version 2.0 from [21] and ASVspoof 2019 from [8] were taken for the evaluation of the ASV and anti-spoofing tasks. ASVspoof 2017 was designed based on RedDots [22], [23] under various environments. The detailed information is listed in TABLE I. ASVspoof 2019 was designed to evaluate logical access (LA) and physical access (PA), both of which were derived from VCTK. LA contains bona fide speech of real-world recordings and spoofed speech data generated by using 17 different TTS and VC systems. Speech data of PA is assumed to be captured by a microphone in a physical, reverberant space. Replay spoofing attacks are recordings of bona fide voices which are assumed to be captured, possibly surreptitiously, and then re-presented to the microphone of an ASV system using a replay device. In contrast to ASVspoof 2017, PA of ASVspoof 2019 was constructed from a far more controlled simulation of replay spoofing attacks, e.g. smart home devices, whose detailed information is presented in TABLE 2.

B. EVALUATION METRICS

Equal error rate (EER) is a commonly used criterion to evaluate the performance of the joint system on speaker verification and spoofing detection. As for speaker verification task, EER is the error rate for a specific value of a threshold where the false rejection rate (FRR) is equal to the false acceptance rate (FAR). False rejection is a target speaker that erroneously classified as an impostor. False acceptance is the opposite case when an imposter is misclassified as a target. In anti-spoofing task, false rejection is a bona fide utterance that is

TABLE 3. CNN Architecture of the unified network.

Network	Layer	Structure	Stride
CNN	Conv1	2×2 , 64	1×1
	Conv2	2×2 , 128	1×1
	Conv3	2×2 , 64	1×1
	Conv4	2×2 , 64	1×1
	Conv5	2×2 , 32	1×1
	Conv6	2×2 , 64	1×1
	Dense	1024	-
	Embedding	512	-

classified as spoofing, while false acceptance is a spoofing utterance that is discriminated as bona fide.

For the joint ASV and anti-spoofing system, it can only be accepted when an utterance is classified as target speaker and bona fide. The false rejection is thus the case that a bona fide utterance from the target speaker is discriminated as the hypothesis $H_{\overline{(\chi,\eta)}}$ in Section II. A. The false acceptance is an utterance from the hypothesis $H_{\overline{(\chi,\eta)}}$ is discriminated as a bona fide utterance from the target speaker.

C. FEATURES

Three different features were extracted for ASV and anti-spoofing, log mel filter bank (Fbank), mel-frequency cepstral coefficients (MFCC) and constant Q cepstral coefficients (CQCC). After removing the silent parts by voice activity detection (VAD), a frame-length of 25ms and 15ms sliding window was applied to extract acoustic features. The Fbank feature was 128-dimensional and MFCC was 19-dimensional with 1$^{\text{st}}$ and 2$^{\text{nd}}$ order delta features (57-dimension in total). For the problem that each utterance had different numbers of frames, a 10 frames' length with 3 frames' sliding window was applied on the frame-level features to divide each utterance into several fragments with the same size. Given the good performance of CQCC on anti-spoofing, 30-dimensional CQCC with its 1st and 2nd order delta features (90-dimension in total) were also extracted as input features.

D. NETWORK ARCHITECTURES

We used three different network architectures with cross-entropy/triplet loss to extract embedding vectors. One of the network was convolutional neural network (CNN) followed by a two-layer fully-connected for classification purpose. The detailed architecture of CNN is given in TABLE 3.

Another network was a six-layer fully-connected DNN network. The output of the embedding layer is also called *d*-vector [24]. The architecture of DNN is shown in TABLE 4.

The third architecture was time-delay deep neural network (TDNN) [25]. The embedding vector extracted from TDNN was also called *x*-vector [26]. The TDNN structure we used was the same as that in [26] and was shown in TABLE 5. The embedding vectors from TDNN we used were from layer *Segment6*.

TABLE 4. DNN Architecture of the unified network.

Network	Layer	Structure
DNN	DNN1	2048
	DNN2	2048
	DNN3	1024
	DNN4	1024
	DNN5	512
	Embedding	512

TABLE 5. TDNN Architecture of the unified network.

Network	Layer	Layer context	Structure
TDNN	Frame1	$[t-2, t+2]$	120×512
	Frame2	$\{t-2, t, t+2\}$	1536×512
	Frame3	$\{t-3, t, t+3\}$	1536×512
	Frame4	$\{t\}$	512×512
	Frame5	$\{t\}$	512×1500
	Stats pooling	$[0, T)$	$1500T \times 3000$
	Segment6	$\{0\}$	3000×512
	Segment7	$\{0\}$	512×512

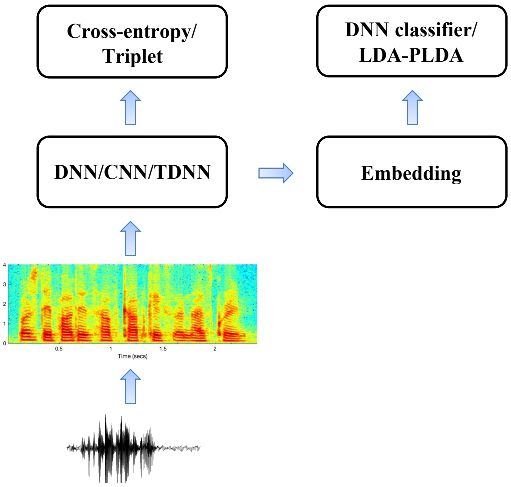


FIGURE 7. The overall framework and its options of joint system.

The overall framework of our joint system is shown in FIGURE 7. Boxes with slashes show the options of network architectures, objective losses or back-end classifiers.

After extracting the embedding vectors, two methods of classification were applied on the back-end of the extracted vectors (see the top-right part of FIGURE 7).

One was two 3-layer adaptive DNN classifiers which were utilized to train two sub-networks on speaker verification and spoofing detection tasks separately. In enrollment and testing stage, 20 utterances per speaker from the evaluation set were taken for speaker enrollment [16].

The other was LDA-PLDA. The LDA was used to reduce the dimension of extracted *x*-vectors to 200. For PLDA, there are two choices when performing ASV and anti-spoofing.

1) Train two PLDAs on the *x*-vectors, one for speaker verification and the other for anti-spoofing. The final results come from the fused discrimination of the two PLDAs;

TABLE 6. Results on Asvspoof 2017 (EER [%]) where The Baselines of Cascaded Combination and Gaussian back-end Fusion are from [17].

Loss	Model	Features	Speaker Verification	Anti-spoofing	Joint
-	Cascaded combination	MFCC	5.36	24.65	-
-	Gaussian back-end fusion	MFCC	**3.26**	24.35	13.81
Cross-entropy	CNN	Fbank	7.47	14.55	11.52
		MFCC	5.84	13.19	10.61
		MFCC+CQCC	5.21	12.23	10.13
	DNN	Fbank	6.39	14.31	11.24
		MFCC	5.37	15.76	11.37
		MFCC+CQCC	5.16	13.98	10.49
	TDNN	Fbank	5.62	14.83	11.67
		MFCC	5.27	14.01	11.15
		MFCC+CQCC	5.11	11.86	10.03
Triplet	TDNN+2PLDA	Fbank	5.09	12.24	10.51
		MFCC	4.62	11.89	10.16
		MFCC+CQCC	4.43	**11.16**	9.75
	TDNN+1PLDA	Fbank	-	-	10.23
		MFCC	-	-	9.84
		MFCC+CQCC	-	-	**8.97**

TABLE 7. Results on Asvspoof 2019 (EER [%]).

Loss	Model	Features	Speaker Verification	Anti-spoofing	Joint
Cross-entropy	CNN	Fbank	7.12	14.93	12.64
		MFCC	5.96	13.72	11.41
		MFCC+CQCC	5.79	12.68	11.24
	DNN	Fbank	7.31	14.75	12.47
		MFCC	6.30	15.61	11.92
		MFCC+CQCC	5.35	12.47	10.85
	TDNN	Fbank	5.96	13.85	12.02
		MFCC	4.85	12.72	10.64
		MFCC+CQCC	4.53	11.22	9.98
Triplet	TDNN+2PLDA	Fbank	4.34	9.81	8.63
		MFCC	3.52	9.73	6.64
		MFCC+CQCC	**3.17**	**8.55**	5.86
	TDNN+1PLDA	Fbank	-	-	8.32
		MFCC	-	-	6.43
		MFCC+CQCC	-	-	**5.65**

2) Train one PLDA on the x-vectors to discriminate whether the tested speech was the target speaker's bona fide utterance or not. It is clear to see that this choice is just what we want to make a joint final decision.

IV. RESULTS AND DISCUSSION

TABLE 6 presents the EER results obtained from different features with CNN, DNN or TDNN networks based on cross-entropy or triplet loss on ASVspoof 2017. A cascaded combination and joint decision of Gaussian back-end fusion were taken as baselines as reported in [17]. For ASV, our joint system achieved the best EER of 4.43% on speaker verification by using MFCC+CQCC and TDNN with triplet loss, while [17] achieved the EER of 5.36% by using MFCC with cascaded decision approach and 3.26% with a Gaussian back-end fusion. For anti-spoofing, our system achieved an EER of 11.89% by using MFCC and TDNN with triplet loss compared to 24.35% in [17]. Detailed comparison is shown in TABLE 6 and the results of the same group of experiments on ASVspoof 2019 are given in TABLE 7. By reading TABLE 6 and TABLE 7, several conclusions could be drawn as listed below.

1) For the results of speaker verification and anti-spoofing with cross-entropy and triplet loss, the best performance always came from MFCC+CQCC, which demonstrated the advantage of using multiple features w.r.t. using one single feature. ASV could benefit from CQCC while anti-spoofing could also benefit from MFCC.

2) The performance of TDNN was better than those of CNN and DNN in both joint or the two separate tasks

(i.e. ASV and anti-spoofing), thanks to its strong ability on modeling time series data. Therefore, only TDNN was experimented with triplet loss.

3) By replacing cross-entropy with triplet loss, the performance of TDNN was further improved, which validated the idea of the paper: multi-task learning with triplet loss improves joint decision.

4) Advantages were also overserved from training a single PLDA for joint decision over training two separate PLDAs for each task and fusing them. One possible reason could be speaker identity and spoofing information were already merged together in the extracted *x*-vectors from TDNN, which could confuse PLDA when the PLDA was only designed for each of the two tasks.

5) The consistent performance of the proposed approach on both ASVspoof 2017 and ASVspoof 2019 justified the robustness of the algorithm across data distributions.

V. CONCLUSION

In this paper, a multi-task learning approach based on contrastive loss was proposed and experimented to make a joint decision of ASV and anti-spoofing. Firstly, embedding vectors containing speaker identity and spoofing information were extracted by deep networks with triplet loss. LDA-PLDA was trained subsequently from the extracted embedding vectors to make the final discrimination. The performance of the proposed approach was evaluated on the ASVspoof 2017 v2.0 and ASVspoof 2019 PA datasets. The experimental results showed that the joint decision approach outperformed some recently proposed baselines.

The proposed approach validated the feasibility of the contrastive loss with deep learning on the multi-task learning of ASV and anti-spoofing. In a general view, the work provided a primary idea for modeling multi-task learning, joint decision and contrastive loss. The further work will study other ways of choosing triplets or other margin functions to further improve the performance on joint decision.

REFERENCES

[1] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Syst. J.*, vol. 40, no. 3, pp. 614–634, 2001.

[2] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and anti-spoofing in the *i*-vector space," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 821–832, 2015.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[4] W. Shang and M. Stevenson, "A playback attack detector for speaker verification systems," in *Proc. 3rd Int. Symp. Commun., Control Signal Process.*, Mar. 2008, pp. 1144–1149.

[5] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. IEEE Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2014, pp. 1–6.

[6] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017.

[7] 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," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 20–24.

[8] 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," Apr. 2019, *arXiv: 1904.05441*. [Online]. Available: https://arxiv.org/abs/1904.05441

[9] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 20–24.

[10] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Stockholm, Sweden, Aug. 2017, pp. 20–24.

[11] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," Jun. 2017, *arXiv: 1706.02101*. [Online]. Available: https://arxiv.org/abs/1706.02101

[12] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU replay detection system for the ASVspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," Jul. 2019, *arXiv:1907.02663*. [Online]. Available: https://arxiv.org/abs/1907.02663

[13] L. Xu, R. K. Das, E. Yilmaz, J. Yang, and H. Li, "Generative x-vectors for text-independent speaker verification," Sep. 2018, *arXiv:1809.06798*. [Online]. Available: https://arxiv.org/abs/1809.06798

[14] L. You, W. Guo, L. Dai, and J. Du, "Deep neural network embeddings with gating mechanisms for text-independent speaker verification," Mar. 2019, *arXiv: 1903.12092v2*. [Online]. Available: https://arxiv.org/abs/1903.12092v2

[15] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text–independent speaker verification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Stockholm, Sweden, Aug. 2017, pp. 20–24.

[16] J. Li, M. Sun, and X. Zhang, "Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection," in *Proc. APSIPA Annu. Summit Conf.*, 2019.

[17] 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 *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Hyderabad, Pakistan, Aug. 2018, pp. 2–6.

[18] C. Zhang and K. Koishida, "End-to-end text–independent speaker verification with triplet loss on short utterances," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Stockholm, Sweden, Aug. 2017, pp. 20–24.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 7–12.

[20] J. Cao, Y. Li, and Z. Zhang, "Partially shared multi-task convolutional neural network with local constraint for face attribute learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 18–22.

[21] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, Les Sables-d'Olonne, France, Jun. 2018.

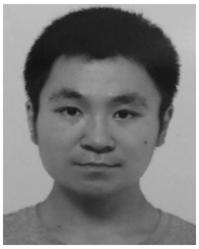
[22] K. A. Lee, A. Larcher, G. Wang, P. Kenny, and N. Brümmer, "The reddots data collection for speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 2996–3000.

[23] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamaki, and K. A. Lee, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5–9.

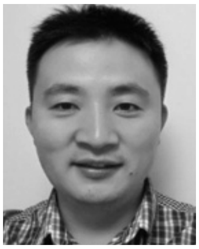
[24] E. Variani, X. Lei, E. Mcdermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 4–9.

[25] V. Peddinti, D. Povey, and D. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 6–10.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 15–20



JIAKANG LI received the M.S. degree from the Department of Science, Army Engineering University, Nanjing, China, in 2017, where he is currently pursuing the Ph.D. degree with the Laboratory of Intelligent Information Processing. His research interests include speech signal processing, anti-spoofing, speaker recognition, and machine learning.



MENG SUN received the Ph.D. degree from the Department of Electrical Engineering, Katholieke University Leuven. He is currently an Associate Professor with Army Engineering University, Nanjing, China. His research interests include speech processing, unsupervised/semi-supervised machine learning, and sequential pattern recognition.

XIONGWEI ZHANG received the Ph.D. degree in signal and information processing from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1992. He is currently a Professor with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, China. His research interests include speech signal processing, machine learning, and pattern recognition.

YIMIN WANG is currently a Researcher with Army Engineering University, Nanjing, China. Her research field includes quantum artificial intelligence.