Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection

Jiakang Li, Meng Sun and Xiongwei Zhang
Army Engineering University, Nanjing, China
E-mail: {jkangli@163.com, sunmengccjs@gmail.com, xwzhang9898@163.com}

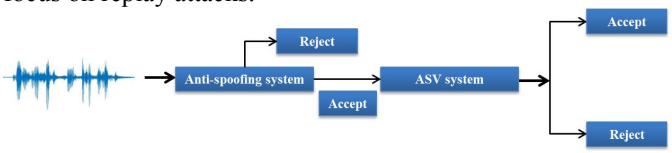
***Abstract*— With the development of spoofing technologies, *automatic speaker verification* (ASV) systems have encountered serious challenges on security. In order to address this problem, many anti-spoofing countermeasures have been explored. There are two intuitive recipes to protect an ASV system from spoofing. The first one is to use a cascaded structure where spoofing detection is performed firstly and ASV is subsequently conducted only on the attempts which have passed the spoofing detection. The other one is to perform spoofing detection and ASV jointly. The discriminate reliably of the joint system has been proven to be more advantageous than cascaded systems with traditional methods, not only in accuracy, but also in convenience and computational efficiency. In this paper, we proposed a multi-task learning approach based on deep neural network to make a joint system of ASV and anti-spoofing. The performance of different acoustic features and structures of deep neural networks has been investigated on the ASVspoof 2017 version 2.0 dataset. The experimental results showed that the joint *equal error rate* (EER) of our approach was reduced by 0.55% compared to a joint system with Gaussian back-end fusion baseline.**

Index Terms— anti-spoofing, speaker recognition, replay detection, multi-task learning, joint detection

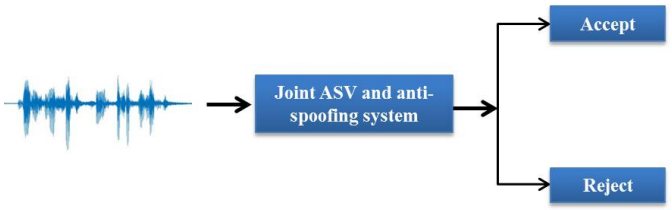
I. INTRODUCTION

The role of biometric authentication in data security is increasingly important these years. Although some commonly used biometric technologies, such as fingerprint, face recognition and voiceprint recognition, have been well applied to authentication scenarios, the security of these recognition systems is still an urgent problem in the case of various spoofing attacks and need to be addressed as soon as possible. In fact, any biometric authentication system has some particular weaknesses that are vulnerable to spoofing attacks [1], where the most accessible ones are sensor and transmission level attacks [2]. For example, using a photo of an authenticated user to attack a face recognition system, or using a playback recording to attack *automatic speaker verification* (ASV) systems. These are the common scenes that can happen to us. In this paper, we study anti-spoofing in text-independent speaker verification systems.

Generally, speech spoofing attacks can be categorized into four types: impersonation, synthesis, conversion, and replay [3]. In order to promote the research on anti-spoofing, the *automatic speaker verification spoofing and countermeasures challenge* (ASVspoof) was first launched in 2015, which focused on discriminating between synthesized or converted voices and those uttered by a human [4]. The ASVspoof 2017 challenge focused on the detection of replay spoofing to discriminate whether the given speech was the voice of an in-person human or the replay of a recorded speech [5]. Among the four kinds of spoofing attacks, replay attack refers to when an attacker uses a high-fidelity recording device to record the voice of a legitimate authentication system user and then uses the recorded playback through the device on the ASV system, thereby achieving an attack behavior [6]. Since replay attacks are easy to implement and highly similar to genuine speech, it is difficult to detect and bring serious threats to speaker verification systems [7]. In this paper, we focus on replay attacks.



(a) A cascaded system



(b) A joint system

Fig. 1: (a) The cascaded system of ASV and anti-spoofing. (b) The joint system of ASV and anti-spoofing.

With increasingly attention being paid to the security of the ASV system, a large number of anti-spoofing methods have been proposed and have achieved quite good results. Early studies [6, 8] proved that the replay detection method is effective for ASV systems. Recent studies [9-11] designed independent anti-spoofing systems by using deep learning methods and achieved promising results in ASVspoof 2017. Most of the current anti-spoofing systems are designed separated from the ASV system because it is relatively easy to conduct as a stand-alone classifier. The structure diagram of these two systems is shown in Fig. 1. When performing speaker verification, replay detection is performed at first, and the speaker verification is conducted subsequently, which is called a cascaded detection system. Another way to improve

the security of the ASV system is to design a joint model to prevent spoofing attacks while verifying the speakers, which is called a joint system. To the best of our knowledge, there is little research on joint systems than cascaded ones till now, where [2] proposed and studied the performance of a joint decision system in an i-vector framework, which demonstrated the advantage of joint systems.

In this paper, we studied the joint text-independent ASV and anti-spoofing system in a deep learning framework. It is believed that any speech can be represented by a feature vector containing speaker and spoofing information, no matter the voice comes from an in-person human or is replayed by a recording device. We may consider that a replay utterance as a distorted version of natural human speech by passing through some specific channels. Such spoofing properties will be reflected in the cepstral features and the features derived by deep neural networks from them, which will enable the detection.

The proposed joint system with deep learning has several advantages. Firstly, the speaker verification task and the anti-spoofing task share the same spectrum features. Secondly, joint model shares a number of common network layers and does not require the establishment of two completely independent anti-spoofing and ASV models. These two properties will reduce the unnecessary redundant computational complexity to a certain extent. Thirdly, the effectiveness of the joint system has been verified in traditional methods, e.g. in i-vector space. It is also expected to be workable in a deep learning fashion.

The remainder of the paper is organized as follows. Section □ describes the database and the evaluation metrics. The joint anti-spoofing and ASV system is presented in Section □. Detailed experimental setup and results are reported in Section □ and □. Section □ concludes the paper.

II. Dataset and Protocols

A. Dataset

In this paper, we used the ASVspoof 2017 version 2.0 corpus [12], which was released for the ASVspoof 2017 challenge and designed based on the RedDots corpus [13, 14] under various environments and unseen scenarios. The full database contains three subsets: Training set, Development set and Evaluation set. Each subset contains the voice of in-person human and the replay of a recorded speech. The sampling rate of the entire database is 16 kHz with sample precision of 16 bits. All three subsets are disjoint in terms of speakers and are also some differences in terms of data collection sites [5]. Detailed information of each subset is presented in Table 1.

B. Evaluation Metrics

Equal error rate (EER) is adopted to evaluate the performance of the joint system on speaker verification and spoofing detection respectively. As a classic evaluation metric, 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. EER was calculated using the MSR Identity Toolkit Version 1.0 [15].

Table 1: Profile of the ASVspoof 2017 version 2.0 corpus

Subset	# Speakers	# Utterances	
		Non-replay	Replay
Training	10	1508	1508
Development	8	760	950
Evaluation	24	1298	12008
Total	42	3566	14466

III. Joint Anti-spoofing and ASV System

Being different from the cascaded anti-spoofing and ASV system, the joint system needs to make speaker verification and anti-spoofing simultaneously. Under this condition, each utterance contains two parts of attributes: information about the speaker — X and information about the genuine/spoofing — Ψ. The hypothesis $H_{(\mathrm{X},\Psi)}$ means the utterance is a genuine utterance from the target speaker X, the complementary hypothesis $H_{\overline{(\mathrm{X},\Psi)}} = H_{(\overline{\mathrm{X}},\Psi)} \bigcup H_{(\mathrm{X},\overline{\Psi})}$ referring to the case that genuine utterance from any non-target speaker or any recordings from target speaker X.

Based on this assumption and inspired by the recent x-vector method [16] in speaker verification, we used deep neural network to extract an embedding that contains both speaker and spoofing information, and used back-end classifiers to make the speaker classification and spoofing detection jointly based on the extracted embeddings.

As shown in Fig. 2, the joint ASV and anti-spoofing system based on deep neural network can be divided into three parts: (a) Pre-training, (b) Re-training and (c) Enrollment. The detailed explanation is as follows:

A. Embedding Extraction

In order to train an embedding which contains both speaker and spoofing information, we proposed the embedding extraction method based on *convolutional neural network* (CNN) and *deep neural network* (DNN). The frame-level features with front-end processing are used as the inputs of pre-training networks. As for speaker verification, we used softmax loss to train the embedding. The softmax loss is,

$$L = -\sum_{j=1}^{T} y_j \log s_j \ , \tag{1}$$

where s_j is the j-th value of the output vector s, which indicates the probability that the current sample belongs to the j-th category; y is a T-dimensional vector which represent the speaker label.

As for anti-spoofing, we also used softmax loss to train the embedding. Spoofing detection is a binary classification problem: genuine or spoofing.

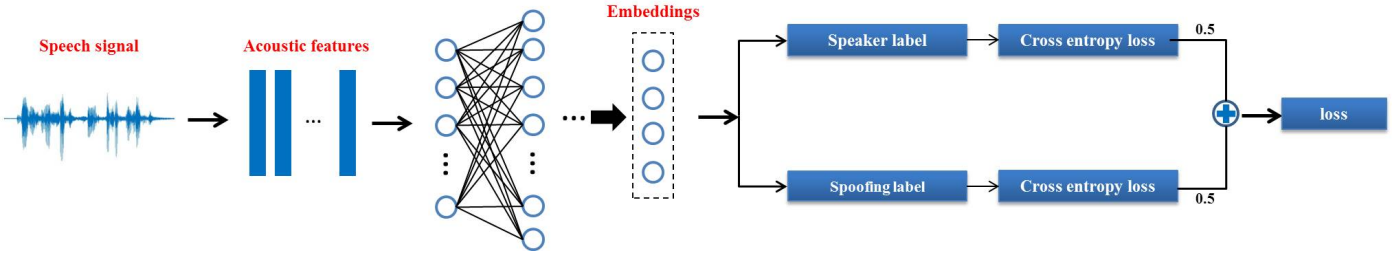
The two losses are then weighted equally to get an overall loss for the whole pre-training network. Through such a pre-training network, we can get embeddings that contain speaker and spoofing information, and could provide relatively representative attributes for subsequent classification and discrimination. Fig. 2(a) shows the detailed structure of this part.

B. *Re-training on Speaker Verification and Spoofing Detection*

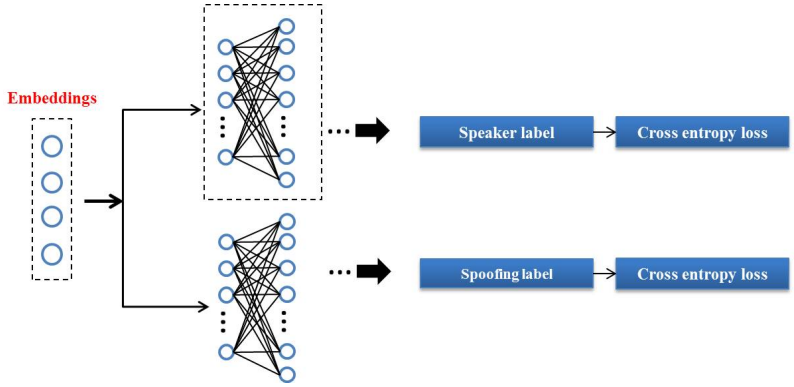
After extracting embeddings with a unified pre-training network, two different DNN discriminators are trained on its backend to identify the speaker and to detect spoofing attacks separately. By using pre-training to extract embeddings, and by utilizing two DNN classifiers on its backend, the entire joint ASV and anti-spoofing system has been established. Fig. 2(b) illustrates this part.

C. *Adaptive Classifier*

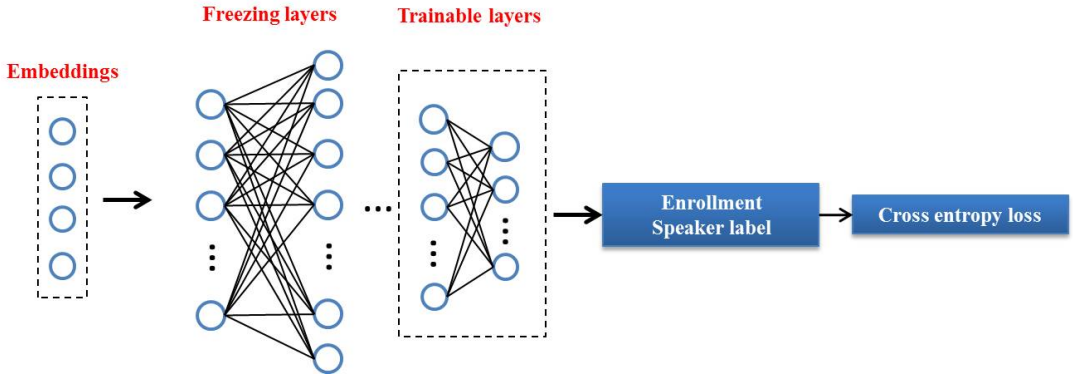
As shown in Fig. 2(c), in speaker enrollment stage, some layers of the trained DNN classifier are retrained to make the previously trained network suitable for the currently registered speakers.



(a) Pre-training stage. This is the structure of the network that shared by both ASV and anti-spoofing tasks.



(b) Re-training stage. The trained embeddings from stage (a) are treated as the inputs of the re-training stage.



(c) Enrollment stage. The trained embeddings from stage (b) are treated as inputs of the enrollment stage. The freezing and trainable parts are from the dotted part in stage (b). The trainable layers (dotted part) are the layers that need to be optimized at this stage.

Fig. 2: The framework of our proposed joint ASV and anti-spoofing system.

IV. EXPERIMENTAL SETUP

A. Data Preparation

The dataset we used was ASVspoof 2017 version 2.0 corpus. 10 speakers from Training and 8 speakers from Development subset (18 speakers, 4726 utterances in total) were utilized to train the joint model, and 17 speakers from Evaluation subset (12376 utterances) were used to verify the performance of the joint system. The EER of speaker verification and anti-spoofing were calculated separately.

B. Features

Two different features were extracted for comparative experiments, *log mel filter bank* (Fbank) and *mel-frequency cepstral coefficients* (MFCC). After removing the silent parts by 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 1st and 2nd order delta features (57-dimension in total). For the problem that each utterance had different numbers of frames, we used a 10 frames' length with 3 frames' sliding window on the frame-level features to divide each utterance into several fragments with the same size.

C. Pre-training

Two different networks were used and softmax layers were taken as the metric for pre-training. One of the pre-training networks was CNN followed by a two-layer fully-connected network for transforming features into one-dimensional vectors. The structure is called CNN-DNN in this paper, whose details are given in Table 2. Another pre-training network was a six-layer fully-connected DNN network. Table 3 illustrates the DNN architecture we used to extract embeddings.

Table 2: Architecture of CNN-DNN pre-training

Layer name	Structure	Stride	# Parameters
Conv1	2×2 ,64	1×1	0.32K
Conv2a	2×2 ,128	1×1	32.9K
Conv2b	2×2 ,64	1×1	32.8K
Conv3a	2×2 ,64	1×1	16.4K
Conv3b	2×2 ,32	1×1	8.2K
Conv4	2×2 ,64	1×1	8.3K
Dense1	1024	-	839K
Embedding	512	-	525K

Table 3: Architecture of DNN pre-training

Layer name	Structure	# Parameters
DNN1	2048	2.62M
DNN2	2048	4.19M
DNN3	1024	2.10M
DNN4	1024	1.05M
DNN5	512	0.52M
Embedding	512	0.26M

D. Re-training

Two 3-layer DNNs were utilized to train two sub-networks on speaker verification and spoofing detection separately. The structures of DNNs for the two parts were both $512\times256\times128$. The total numbers of parameters were 854K.

E. Adaptation

As for testing, 20 utterances per speaker were taken for enrollment (340 utterances). The final output layer softmax was adjusted and retrained in conjunction with the last layer of DNN, making the classifier to fit the number of enrollment speakers.

V. EXPERIMENTAL RESULTS

Table 4 presents the *accuracy values* (ACC) and EER results obtained from different features with CNN-DNN pre-training. Table 5 illustrates the performance with DNN pre-training.

For the results of speaker verification, the best performance came from the DNN based pre-trainining architecture with MFCC features, its EER was 9.76% relatively lower than that of the CNN-DNN architecture. While the best result of anti-spoofing came from the CNN-DNN architecture with MFCC features, 7.83% relatively lower than that from the Fbank with DNN in EER.

We compared our joint system's performance with the cascaded and joint method of ASV and anti-spoofing in [17]. As for speaker verification subtask, our joint system achieved the best EER of 5.27% in speaker verification on evaluation subset by using MFCC with DNN pre-training while [17] achieved the best EER of 4.92% with *infinite impulse response constant Q mel cepstral coefficients* (ICMCs) features by using cascaded combination. For anti-spoofing task, our system achieved an EER of 13.19% by using MFCC with CNN-DNN pre-training, while [17] achieved 21.28% with *linear frequency cepstral coefficients* (LFCCs) features. For the joint system in [17], the best EER of the two tasks are 2.90% and 17.98% separately. Detailed comparison results are shown in Table 6.

By comparing the average results in Table 6, we can see that the joint system we proposed achieved an EER of 6.24% in ASV task, a little worse than the EER of baseline in [17]. But in terms of anti-spoofing, our system achieved a better EER than [17].

Moreover, in our joint system, the MFCC features were better in terms of ACC and EER of speaker verification for both CNN-DNN and DNN. But for anti-spoofing, Fbank achieved better results in DNN but MFCC in CNN-DNN, so it is still not quite clear which feature is better and further research needs to be explored for this.

Finally, we combined the results of ASV and anti-spoofing, the false reject and false accept samples of two sub-tasks were discriminated as misjudgment samples. In other words, as long as one of the two indicators was wrong, the attempt was rejected by the system. Therefore, final results were able to be obtained intuitively.

As shown in Table 7, the results of our ASV and anti-spoofing were combined together to get the final discrimination (accept or reject). Compared with the results

of Gaussian back-end fusion baseline in [17], our joint system reduced by 0.55% in EER

Table 4: The performance of CNN-DNN pre-training architecture with adaptation DNN classifier on Fbank and MFCC features

Features	Speaker Verification		Anti-spoofing	
	EER[%]	ACC[%]	EER[%]	ACC[%]
Fbank	7.47	78.16	14.55	90.73
MFCC	**5.84**	82.32	**13.19**	91.25

Table 5: The performance of DNN pre-training architecture with adaptation DNN classifier on Fbank and MFCC features

Features	Speaker Verification		Anti-spoofing	
	EER[%]	ACC[%]	EER[%]	ACC[%]
Fbank	6.39	80.02	**14.31**	91.05
MFCC	**5.27**	84.23	15.76	88.43

Table 6: The comparison of cascaded/tandem combination and Gaussian back-end fusion method in [17] with our joint ASV and anti-spoofing system based on deep learning framework. All the results of each system are the average value.

System	Speaker Verification	Anti-spoofing
	EER[%]	EER[%]
Cascaded combination	5.89	24.75
Gaussian back-end fusion	**3.65**	20.74
Our joint system	6.24	**14.45**

Table 7: The final discriminate result of Gaussian back-end fusion method in [17] and our joint ASV and anti-spoofing system.

System	EER[%]
Gaussian back-end fusion	11.41
Our joint system	10.86

VI. CONCLUSIONS

In this paper, a multi-task learning approach based on deep neural network was proposed and experimented to make a joint system of ASV and anti-spoofing. Firstly, embeddings contained speaker and spoofing information were extracted by two kinds of pre-training networks. Then, two DNN sub-networks were trained based on the extracted embeddings. Finally, adaptation for the DNN sub-networks was conducted during the speaker enrollment stage. The performance of our joint system was evaluated on the ASVspoof 2017 v2.0 database. The results showed that the EER of our joint system was reduced by 0.55% compared to the Gaussian back-end fusion baseline.

It can be seen that DNN pre-training architecture was better than CNN-DNN in speaker verification, no matter what kind of acoustic feature was used. For anti-spoofing, the CNN-DNN was more prominent than DNN. As for features, MFCC was better than Fbank for speaker verification.

This work validated the feasibility of the deep learning method in the joint ASV and anti-spoofing system and provided a primary framework for solving the multi-task learning of the joint system. Further work will study how to improve the performance of ASV and anti-spoofing jointly with other deep learning methods by using different acoustic features.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Jiangsu Province with grant number BK20180080.

REFERENCES

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

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

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, et al., "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, 2015, 66, pp.130-153.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, et al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Top. Signal Process.,* 2017, 11, pp. 588-604.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, et al., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 20-24 August 2017.

[6] W. Shang and M. Stevenson, "A playback attack detector for speaker verification systems," *IEEE International Symposium on Communications*, 2008.

[7] F. Alegre, A. Janicki and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," *IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 10-12 September 2014.

[8] W. Shang and M. Stevenson, "Score normalization in playback attack detection," *IEEE International Conference on Acoustics, Speech & Signal Processing*, Dallas, TX, USA, 14-19 March 2010.

[9] L. Galina, N. Sergey, M. Egor, K. Alexander, K. Oleg, et al., "Audio replay attack detection with deep learning frameworks," *Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 20-24 August 2017.

[10] H. A. Patil, M. R. Kamble, T. B. Patel and M. Soni, "Novel variable length teager energy separation based instantaneous frequency features for replay detection," *Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 20-24 August 2017.

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

[12] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, et al., "ASVspoof 2017 Version 2.0: metadata analysis and baseline enhancements," *Odysssey, the Speaker & Language Recognition Workshop*, Les Sables D'Olonne, France, 2018.

[13] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, et al., "The reddots data collection for speaker recognition," *Annual Conference of the International Speech Communication Association*, Dresden, Germany, 6-10, September, 2015, pp. 2996-3000.

[14] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R.G. Hautamaki, et al., "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research,"

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5-9, March 2017.

[15] S. O. Sadjadi, M. Slaney and L. Heck, "MSR Identity Toolbox: A MATLAB Toolbox for speaker recognition research," *Speech Lang. Tech. Comm. Newsl.*, 2013, 1, pp. 1-32.

[16] S. David, G. R. Daniel, P. Daniel and K. Sanjeev, "Deep neural network embeddings for text-independent speaker verification," *Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 20-24 August 2017.

[17] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, et al., "Integrated presentation attack detection and automatic speaker verification: Common feaures and Gaussian back-end fusion," *Annual Conference of the International Speech Communication Association*, Hyderabad, Pakistan, 2-6 September 2018.