Circuits, Systems, and Signal Processing (2022) 41:4068–4089
https://doi.org/10.1007/s00034-022-01974-z

Multi-task Learning-Based Spoofing-Robust Automatic Speaker Verification System

Yuanjun Zhao[1] · **Roberto Togneri**[1] · **Victor Sreeram**[1]

Received: 14 December 2020 / Revised: 22 January 2022 / Accepted: 22 January 2022 /
Published online: 18 February 2022

Abstract
Spoofing attacks posed by generating artificial speech can severely degrade the performance of a speaker verification system. Recently, many anti-spoofing countermeasures have been proposed for detecting varying types of attacks from synthetic speech to replay presentations. While there are numerous effective defenses reported on standalone anti-spoofing solutions, the integration for speaker verification and spoofing detection systems has obvious benefits. In this paper, we propose a spoofing-robust automatic speaker verification system for diverse attacks based on a multi-task learning architecture. This deep learning-based model is jointly trained with time-frequency representations from utterances to provide recognition decisions for both tasks simultaneously. Compared with other state-of-the-art systems on the ASVspoof 2017 and 2019 corpora, a substantial improvement of the combined system under different spoofing conditions can be obtained.

Keywords Automatic speaker verification · Spoofing-robust · Multi-task learning · Anti-spoofing countermeasures

1 Introduction

Prior to the consideration of spoofing, speaker/voice recognition systems have been designed and widely used for commercial and forensic applications by identifying

✉ Yuanjun Zhao
zhaoyuanjun1122@hotmail.com

Roberto Togneri
roberto.togneri@uwa.edu.au

Victor Sreeram
victor.sreeram@uwa.edu.au

[1] Department of Electrical, Electronic and Computer Engineering, The University of Western Australia, Perth, WA 6009, Australia

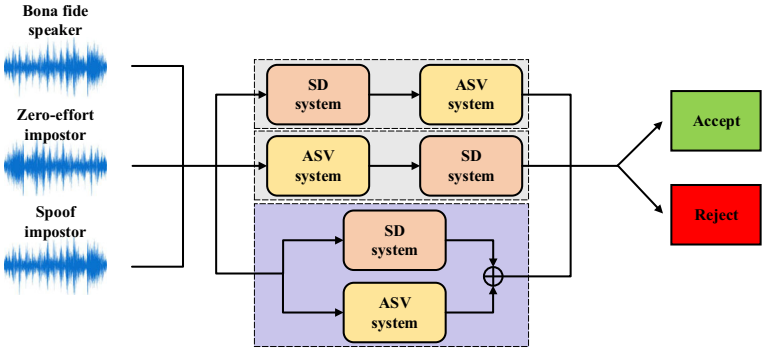


Fig. 1 Integrated systems of different combination methods (top two are cascaded systems, bottom is a parallel system)

and verifying the claimed identity of a speaker [39]. However, for instant and convenient authentications, issues of malicious interference and manipulations on speaker recognition systems are coexisting [46]. The potential for speaker recognition systems to be spoofed is now well-recognized [7, 8, 42]. Urgent needs are suggested to address spoofing in numerous vulnerability studies. Generally, approaches involved for anti-spoofing concentrate on proposing specific or generalized standalone spoofing detectors. From economical and practical perspectives, designing spoofing-robust systems which integrate the functions of spoofing detection and speaker recognition make sense. While both options are vital in the community, integrated systems can assist to streamline recognition processes, reduce costs and ensure efficiency.

Until now, most relevant studies only focus on the framework and evaluation of standalone countermeasures. However, there are a number of reasons why integration of the spoofing detection and speaker recognition system is important [30]. First, since the recognition system and its corresponding spoofing detector are trained to solve two different tasks, a standard linear fusion on the score level is not appropriate. Second, the performance of a spoofing detection system is critical for the final output decision. An imperfect spoofing detector can increase the false alarm rate by rejecting genuine speakers [29]. Thirdly, in a framework which contains separated spoofing detection and speaker recognition modules, it has not been confirmed whether improvements in standalone countermeasures should improve the overall system as a whole. A perfect anti-spoofing system will fail to protect a recognition system which is not properly calibrated [24].

The impact of standalone spoofing detection systems is difficult to be gauged unless they are evaluated when integrated with an ASV system [9]. For example, as binary classification problems, ASV and spoofing detection systems have a common goal of distinguishing unauthorized access attempts. For ASV systems, zero-effort impostors are rejected. While for spoofing detectors, forged trials from the spoof impostors will be detected. As shown in Fig. 1, there are two conventional and simple solutions to jointly combine an ASV and a spoofing detector. The first is to cascade these two modules in alternative orders to protect ASV from spoofing attacks [1, 5]. Obviously, two recognition thresholds are required in the cascaded approach to be applied to each

module. The final decision is obtained by comparing the produced scores with two thresholds. Only the trials with scores that are not less than both the thresholds can be accepted. In addition to the cascaded method, another solution is the parallel approach which is also shown in Fig. 1. A single threshold is used in the parallel approach to compare with fused scores for the final decision. For integrated systems, only Bona fide speech samples can be accepted for correct authentication while samples from both the zero-effort and spoof impostors will be rejected as illegal access.

Research in the area of spoofing-robust integrated recognition system is still in its relative infancy and greater attention is needed in the future. In [29], several state-of-the-art spoofing countermeasures were integrated with ASV systems. Selected countermeasures were combined in a cascaded or parallel framework and evaluated with the ASVspoof 2015 corpus. Experimental results indicated that when ASV systems were integrated with a diverse set of countermeasures, the performance can remain robust in the presence of varying attack approaches. A subsequent exploration on the ASVspoof 2017 V2.0 corpus was given in Todisco et al. [37]. A Gaussian back-end fusion approach was presented to combine the spoofing detector and ASV system. A variety of different features were used to assess the performance of the integrated system. The proposed combination approach was shown to generalize particularly well across independent development and evaluation subsets. In [6], a joint modeling approach was introduced to detect spoofing attacks while also performing the speaker verification task. Factor analysis methods were adopted such that the spoof variability subspace and the speaker variability subspace are jointly trained. Experiments were performed using the speaker and spoofing (SAS) database [43]. Compared to a baseline system integrated in the conventional method, the proposed approach provided substantial improvements for spoof detection as well as speaker verification.

Unlike conventional fusion methods, the problem of ASV and spoofing detection integration is that these two systems are designed with different objectives. It has been demonstrated that the performance of a spoofing detector naturally impacts the performance of the ASV system; either the false alarm rate or the false reject rate will be influenced [16]. With progress in standalone anti-spoofing research continuing, we should also concentrate on integrated spoofing-robust ASV systems which are optimized jointly. Another problem relates to the degraded performance of ASV systems which are expected to be attacked by varying types of spoofed speech. Even if a speaker verification system is integrated with spoofing detection, it is still troublesome to overcome the performance loss caused by diverse attacks. Although integrated spoofing detection and ASV systems have been proposed no system has been designed to handle diverse attacks, that is both logical access (using machine generated spoofed speech) and physical access (using replayed spoofed speech) attacks.

Different to the previous works, in this paper, we pursue solutions for integrated spoofing-robust ASV (SR-ASV) systems which are aware of the logical and physical access attacks simultaneously. To extend the generalization ability of the model used, sequential residual convolutional blocks with Max-Feature-Map activations (MFM) [41] are applied. So far, however, there has been limited discussion about this type of versatile anti-spoofing countermeasure which can handle both condition attacks. The work presented here provides the first investigation on how to jointly optimize both the spoofing detection and ASV tasks for diverse attacks. The proposed integrated system

is evaluated on the newly released ASVspoof 2017 Version 2.0 and 2019 corpora and compared with other state-of-the-art systems. Results demonstrate that the proposed SR-ASV system can overwhelm the other state-of-the-art integrated systems for both spoofing conditions. This also indicates the model used is efficient for both speech processing tasks. Detailed discussion and analysis of the experimental results are given in Sect. 5. More details of the proposed SR-ASV system can be found in Sect. 3.

The contributions of this paper are as follows:

- In this paper, it is the first time that an integrated spoofing-robust ASV system is proposed with a generalization for both logical and physical condition attacks. By adopting the multi-task learning, the system introduced is optimized jointly to obtain an effective representation based on the combined information from anti-spoofing and speaker verification tasks. The auxiliary relations between these two tasks are utilized in the training process.
- The discriminative information of speakers and artifacts caused by spoofing attacks in acoustic features are crucial for building an effective verification system. To obtain the abilities of reducing spectral variations and modeling spectral correlations in acoustic features, we adopt sequential residual convolutional blocks with MFM activations. These network units are used for the first time in the training for integrated spoofing-robust ASV systems based on the multi-task learning.

The rest of the paper is organized as follows. In sect. 2, several related works are introduced. The proposed SR-ASV system is introduced in Sect. 3 and the experiment settings are provided in Sect. 4. The experimental results and relevant analysis are given in Sect. 5 followed by our conclusions in Sect. 6.

2 Related Works

In this paper, we introduce a novel integrated solution for a spoofing-robust ASV system based on deep learning techniques. Generally, deep neural networks (DNNs) are used for extracting discriminative embeddings for each speaker [32, 33] and as part of an end-to-end system for speaker verification [12]. Applying a deep learning-based architecture to speaker verification is a relatively new endeavor [25]. The proposed system is based on the multi-task learning (MTL) approach, which has been used successfully across all applications of machine learning, from natural language processing [35] and speech recognition [4] to computer vision [15] and acoustic event detection [44]. Recently, multi-task learning-based architectures have been adopted in biometrics recognition and anti-spoofing (especially for presentation attack detection (PAD) which is also known as replay spoofing). In [3], a multi-task PAD approach was proposed to simultaneously perform iris detection and iris presentation attack detection. A convolution neural network (CNN) used for general object detection was leveraged to build a multi-task learning framework. With this approach, a bounding box defining the spatial location of the iris can be predicted and a presentation attack score denoting the probability can be generated. The MTL framework was also used for joint face recognition and PAD [14]. Convolutional layers were applied for fea-

ture extraction and two parallel output networks were used for face recognition and classification of PAD.

Due to the effective representation learned by MTL models, this architecture has also been adopted in speaker anti-spoofing. In [22], an MTL network was used for improving anti-spoofing performance with a proposed helpful butterfly unit (BU). The authors achieved the evaluation EER of 2.39% from the best single system on ASVspoof 2019 PA. In [40], a siamese neural network (SNN) was used to build an MTL network that can yield improvement with additional reconstruction loss. In addition, multi-task outputs can also be applied to predict spoofing labels and replay configuration labels as in Yang et al. [45]. The sum of the multi-task outputs in both Bona fide nodes was regarded as the detection score. A similar work was Shim et al. [31], in which an MTL network was used to classify the noise of playback devices, recording environments and recording devices as well as the spoofing detection.

In [37] it was the first work to fuse anti-spoofing and speaker verification on the score level by using a cascade manner, while with the MTL network the integration is made at the model level in this work. Applying MTL models for integrating anti-spoofing and speaker verification is still in the early stages. In [21], contrastive loss was used in an MTL framework in order to improve the cascaded decision approach. A modified triplet loss was constructed for extracting embeddings containing information of both speaker identity and spoofing. However, this work only focused on the physical condition attacks and the logical condition attacks were not considered, which can also pose a serious threat to ASV systems. In addition, the deep learning architectures adopted in Li et al. [21] were conventional networks in speaker recognition, such as sequential fully connected layers, convolutional layers and time-delay DNNs. Although these networks have been proved as efficient solutions for the ASV task, their effectiveness in spoofing detection is still unknown.

In this paper, we provide a more effective MTL framework by employing the Max-Feature-Map activation (MFM) [41] which has been used in high-performing LCNN-based spoofing detection systems [17, 18]. We also design the entire MTL network according to the ResBlock used for ASV in Li et al. [20]. With these architectures, the proposed network can be trained to extract speech embeddings of high discrimination and precise representation. The type of embeddings extracted include both the information used for spoofing detection and ASV tasks and will not cause biased evaluation results. It is the first time that residual convolutional blocks and MFM activation layers are used in a multi-task learning framework for jointly training a spoofing-robust speaker verification system. More introductions of the network used in this paper are given in Sect. 3.

The multi-task learning approach is performed with hard parameter sharing of hidden layers in this network. The tasks of spoofing detection and speaker verification are optimized jointly with an auxiliary effect. Hidden layers are shared between these two tasks during the training while several task-specific output layers are cascaded afterwards. This is the most commonly used approach in multi-task learning to greatly reduce the risk of over-fitting. Since the tasks of spoofing detection and speaker verification are learned simultaneously, the model is forced to find a representation that captures both tasks and generalizes better than either single task. After training, the model will be used for extracting embeddings applied with a probabilistic linear

discriminant analysis (PLDA) back-end for speaker verification. The decisions for spoofing detection can then be directly achieved by adding a softmax output layer to the network or adopting a GMM back-end classifier.

3 Proposed SR-ASV System

3.1 Framework of SR-ASV System

In this section, the detailed framework of the proposed SR-ASV system is introduced, which is depicted in Fig. 2. In the preprocessing phase, the training samples are converted to time-frequency representation-based features. Extracted time-frequency representation (TFR) features are used to train the deep learning-based embedding extractor with corresponding labels which contain the speaker IDs and detection keys. The MTL network used in this work is designed according to the high-performing LCNN architecture [17, 18] and the ResBlock employed in the Deep Speaker [20].

For the task of ASV, a softmax layer is applied in the training process and the number of nodes equals the number of training speakers. Following several fully connected layers, this output layer can provide discriminative embeddings for each speaker. These representations are centered and length-normalized before modeled by a PLDA back-end. The scores are normalized using the adaptive s-norm [34]. To

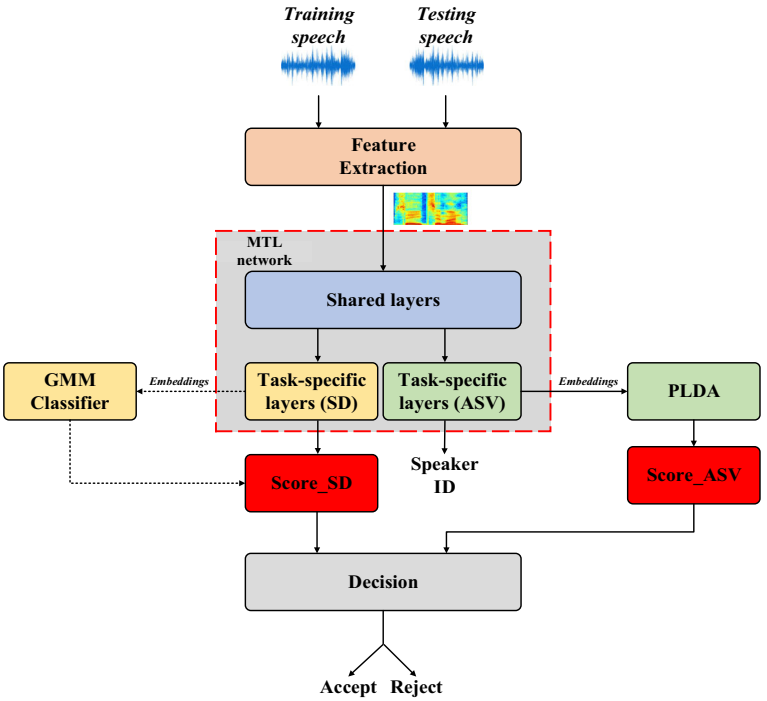


Fig. 2 The scheme of the proposed SR-ASV system

achieve the final decision, we use the t-DCF [16] as the performance metric which will be introduced later.

Compared to shallow networks, deep networks can learn complex representations from acoustic features and even original time-domain waveforms. However, it is commonly known that deep networks tend to be more difficult to train due to problems like vanishing gradient. To overcome this, stacked residual blocks are used to force the gradient information to be passed to deeper layers in the network. Each ResBlock contains direct links between the shallower layer outputs and the deeper layer inputs. Networks built with this architecture have a stronger ability to provide an efficient high-level feature representation. This is important for establishing an integrated speaker verification system which can be aware to potential spoofing attacks. As introduced in Sect. 2, there are several shared ResBlocks for extracting discriminative representations, which are then used for both tasks. If we ignore the information and the content carried by a voice, a spoofed speech can be thought of as a Bona fide speech from a 'fake' speaker. With this hypothesis, a spoofing detection task is similar to a two-speaker classification and a speaker verification task is a multi-speaker classification by its definition. Shared ResBlocks in the proposed multi-task learning network are used for capturing auxiliary relationships from both tasks. The experimental results in Sect. 5 demonstrate the effectiveness and the benefit of this architecture.

Another effective method employed to enhance the generalization of the proposed SR-ASV system is the MFM activation. For a multi-task learning-based system, if the artifacts in spoofed speech are not properly handled, a biased result will be yielded. The commonly-used ReLU activation leads to a zero output value if a node is not active in the network. This might cause a loss of some information especially for the first several convolutional layers [41]. By using the MFM activation, optimal feature at each location of different kernels are selected. A model with MFM can obtain a compact and robust representation while the gradients of MFM layers are sparse. Competitive nodes learning more generalized information can be output from the MFM layer. This property of MFM helps to force the network to extract effective embeddings for both tasks.

By rebuilding and modifying the essential network layers like the MFM and residual blocks, the proposed network can be trained to extract speech embeddings of high discrimination and precise representation.

3.2 Multi-task Learning

The multi-task learning network is used in this paper to train the embedding extractor jointly. The tasks of spoofing detection and ASV are integrated at the model-level. The training space for the joint model can be expressed as:

$$H = \left\{ \boldsymbol{X}_i, \boldsymbol{y}_i^{\mathrm{SD}}, \boldsymbol{y}_i^{\mathrm{ASV}} \right\} \tag{1}$$

where $\boldsymbol{X}_i$ is the input of ith speaker. The $\boldsymbol{y}_i^{\mathrm{SD}}$ and $\boldsymbol{y}_i^{\mathrm{ASV}}$ are the corresponding one-hot encoded labels indicating the spoofing detection key and speaker ID, respectively.

To optimize the joint model, we adopt the angular margin-based softmax loss (A-softmax) which has been used for face recognition [23] and speaker embeddings extraction [26]. Recently, the A-softmax is also applied in spoofing detection to train deep learning-based architectures [18]. A-softmax is a well-regularized loss function by forcing learned features to be discriminative on a hypersphere manifold. This loss function can be described as:

$$L_A(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \frac{1}{N}\sum_i -\log\left(\frac{e^{\|x_i\|\cos(m\theta_{i,y_i})}}{e^{\|x_i\|\cos(m\theta_{i,y_i})} + \sum_{i\neq y_i} e^{\|x_i\|\cos(m\theta_{i,y_i})}}\right) \tag{2}$$

where N is the number of training samples $\{x_i\}_{i=1}^N$ and labels $\{y_i\}_{i=1}^N$. The angle between a sample x_i and the corresponding column y_i of the fully connected classification layer weights $\boldsymbol{W}$ is denoted as θ_{i,y_i}. In addition, m is an integer that controls the size of an angular margin between classes.

If we use $L_{\text{SD}}\left(\boldsymbol{X}, \boldsymbol{y}^{\text{SD}}, \boldsymbol{\theta}^{\text{SD}}\right)$ and $L_{ASV}\left(\boldsymbol{X}, \boldsymbol{y}^{\text{ASV}}, \boldsymbol{\theta}^{\text{ASV}}\right)$ to denote the loss functions for spoofing detection and ASV, we can obtain the total cost function as below for simplicity:

$$J\left(\boldsymbol{X}, \boldsymbol{y}^{\text{SD}}, \boldsymbol{y}^{\text{ASV}}, \boldsymbol{\theta}^{\text{SD}}, \boldsymbol{\theta}^{\text{ASV}}\right) = L_{\text{SD}}\left(\boldsymbol{X}, \boldsymbol{y}^{\text{SD}}, \boldsymbol{\theta}^{\text{SD}}\right) + L_{\text{ASV}}\left(\boldsymbol{X}, \boldsymbol{y}^{\text{ASV}}, \boldsymbol{\theta}^{\text{ASV}}\right) + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 \tag{3}$$

where the λ is the regularization parameter which is optimized on the development subset.

4 Experimental Settings

4.1 Database

To compare the proposed SR-ASV system with the state-of-the-art, it is evaluated on the ASVspoof 2017 and 2019 databases which are released by the challenge organizers.[1] Detailed statistical summaries of these two corpora can be found as in Tables 1 and 2.

4.1.1 ASVspoof 2017 Corpus

The ASVspoof 2017 corpus has been used in the second challenge and was collected for countermeasures to replay spoofing attacks. Bona fide utterances are a sub-set of the RedDots corpus [19] and spoofed utterances are the result of replaying and recording bona fide utterances using a variety of heterogeneous devices and acoustic environments. There are two versions for this database. Version 1.0 was used as the official corpus in the 2017 challenge while Version 2.0 was released afterwards. In

[1] https://www.asvspoof.org/database.

Table 1 Statistics of the ASVspoof 2017 Version 2.0 database

Subset	#Speaker	#Replay sessions	#Replay config.	#Utterances	
				Bona fide	Replay
Training	10	6	3	1507	1507
Dev.	8	10	10	760	950
Eval.	24	161	57	1298	12,008
Total	42	177	61	3565	14,465

Table 2 Statistics of the ASVspoof 2019 database

Subset	#Speaker		#Utterances			
	M	F	LA		PA	
			Bona fide	Spoof	Bona fide	Spoof
Training	8	12	2580	22,800	5400	48,600
Dev.	8	12	2548	22,296	5400	24,300
Eval.	21	27	7355	63,882	18,090	116,640
Total	37	51	12,483	108,978	28,890	189,540

this updated version, a number of data anomalies in the Version 1.0 of the corpus were removed to avoid potential influence on the detection results. In this paper we will utilize the Version 2.0 corpus to evaluate the proposed SR-ASV system in an impartial and objective manner. Furthermore, the protocol[2] used for the enrollment with regard to the ASVspoof 2017 corpus is the same as in [37].

4.1.2 ASVspoof 2019 Corpus

The other database is the ASVspoof 2019 corpus and it is designed for the third challenge. This database encompasses two partitions: logical access (LA) and physical access (PA) scenarios, which are all derived from the VCTK base corpus.[3] All the spoofed speech are generated from the Bona fide data using diverse spoofing algorithms. More details of the spoofing approaches have been released in the challenge evaluation plan.[4]

There is another dataset included with the ASVspoof 2019 corpus and used for the enrollment of the baseline ASV system in the ASVspoof 2019 challenge. We also adopt this dataset in this paper to enroll the SR-ASV system. The details of the enrollment partition subset are given in Table 3. All these speakers in the table are presented in the corresponding subsets in the ASVspoof 2019 corpus given in Table 2. Detailed numbers of speech samples from different genders are also provided for both partitions.

[2] https://www.asvspoof.org/index2017.html.

[3] http://dx.doi.org/10.7488/ds/1994.

[4] https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf.

Table 3 Statistics of the enrollment partition of the ASVspoof 2019 database

Subset	#Speaker		#Utterances			
	M	F	LA		PA	
			M	F	M	F
Enrollment_Dev.	4	6	76	66	2052	1782
Enrollment_Eval.	21	27	399	297	10,773	8019
Total	25	33	475	363	12,825	9801

Similar to the baseline ASV system [38] used in the challenge, the VoxCeleb corpus[5] is used to pre-train the shared hidden layers of the MTL architecture used in this work. Then the entire network is trained with the training subsets of the challenge corpora. PLDA adaptation is performed with the enrollment subsets, which include disjoint, Bona fide, and in-domain speech samples.

4.2 Front-End Processing

Two types of time-frequency representation-based speech features including the constant Q transformation (CQT) and log linear-filterbank (LLFB) are adopted in this work. To extract the CQT feature for each utterance, we apply the open-source Matlab toolkit.[6] The maximum and the minimum frequency in the constant Q transform are set as $F_{\max} = F_{\text{sample}}/2$ and $F_{\min} = F_{\max}/2^9$ respectively. The Nyquist frequency of the database is $F_{\text{sample}} = 16$ kHz. The number of octaves is 9 and the number of bins per octave B is set to 96, which results in a time shift of 8 ms. The parameter γ is set to $\gamma = \Gamma = 228.7 * \left(2^{(1/B)} - 2^{(-1/B)}\right)$. The re-sampling period is $d = 16$. After the CQT is applied on an utterance, we take the same truncating process as in [17, 18] to get spectral features with a size of 864×400. The LLFB features are extracted by a classical pipeline for filterbank-based features. Specifically, a signal goes through a pre-emphasis filter; is segmented into (overlapping) frames and a Hamming window function is applied to each frame (the frame length is 25 ms and the frame step size is 10 ms); afterwards, we use a 256 points short-time Fourier Transform (STFT) on each frame and calculate the power spectrum; and subsequently compute and apply a linear-scale filter banks with 80 triangular overlapping windows where center frequencies of the windows are equally spaced along a Hz scale [13]. By taking the logarithm of the power spectrogram and truncating these utterance-level features with the same processing used for the CQT feature, LLFB features with unified shapes of 80×400 (where 400 is the number of time frames) can be obtained.

[5] An audio-visual dataset contains over 100K utterances for 1251 celebrities, which are extracted from videos uploaded to YouTube. This dataset is available at http://www.robots.ox.ac.uk/~vgg/data/voxceleb/.

[6] http://audio.eurecom.fr/content/software.

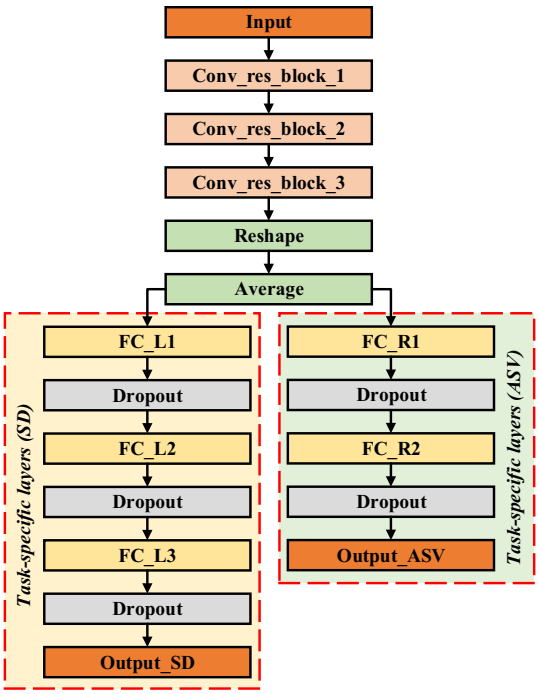


Fig. 3 The multi-task learning architecture used in the SR-ASV system

4.3 Network Architectures and Configurations

The shared hidden layers in the MTL network are built by a series of residual convolutional blocks. Separate task-specific fully connected layers are used after the shared hidden layers. The architecture of the entire network[7] is demonstrated in Fig. 3. Due to the relatively large dimension of speech features and the limited computing power, in each block we apply the Max-Feature-Map (MFM) activations to simplify the deep learning architecture. The MFM activation adopts a competitive relationship to obtain a compact representation and performs feature filter selection. More details can be found in [41].

As shown in Fig. 3, there are three residual convolutional blocks used in the MTL network, of which the outputs are reshaped and averaged before the task-specific layers. The detailed components of the residual convolutional block are shown in Fig. 4. For the whole network, we use 6 convolutional layers, 6 Network in Network (NIN) layers, 12 MFM layers and 5 fully connected layers. We modify the connections in the residual blocks to perform the full pre-activation for the optimal gradient flow [11]. For the shared residual convolutional blocks, we insert the Leaky ReLU (LReLU) activation function and batch normalization to stabilize the model training. The number of filters used in the three residual convolutional blocks are 32, 64 and 128, respectively.

Detailed configurations and statistics of the network parameters are listed in Table 4 and Table 5. The He normal initializer [10] and the L2 regularizer are used with a

[7] The corresponding codes used for reimplementing the experiments will be released later.

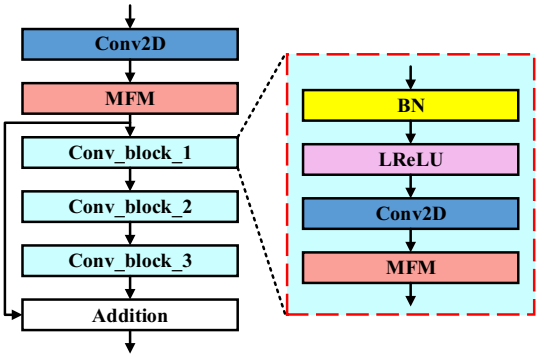


Fig. 4 The residual convolutional block used in the MTL network

Table 4 An example of the network configurations of the MTL network

Type	Output	#Params
Input	864 × 400 × 1	–
Conv_res_block_1	432 × 200 × 16	6K
Conv_res_block_2	216 × 100 × 32	32K
Conv_res_block_3	108 × 50 × 64	128K
Reshape	108 × 3200	–
Average	3200	–
FC	512 (512)	1.6M (1.6M)
Dropout	512 (512)	–
FC	128 (128)	65K (65K)
Dropout	128 (128)	–
FC	64 (–)	8K (–)
Dropout	64 (–)	–
Output	2 (78)	130 (10K)
Total	–	3.6M

Statistics of the task-specific layers for ASV are given in brackets

regularization parameter $\lambda = 0.001$. To avoid the over-fitting issue, dropout 0.7 is used in the network. The Adam optimizer is adopted. Class weights are needed to address the imbalanced training data.

4.4 Performance Metrics

4.4.1 Tandem Decision Cost Function (t-DCF)

In [16], the tandem decision cost function (t-DCF) was proposed for the assessment of combined spoofing countermeasures and ASV. The t-DCF has been adopted as the official primary performance metric.[8] The basic form of t-DCF is expressed as below:

$$\text{t-DCF}(s) = C_1 P_{\text{miss}}^{\text{SD}}(s) + C_2 P_{\text{fa}}^{\text{SD}}(s) + C_0 \tag{4}$$

[8] www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf.

Table 5 Configurations of residual convolutional blocks used in the MTL network

Type	Filter/stride	Output	#Params
Conv2D	3×3 / 2×2	$432 \times 200 \times 32$	320
MFM	–	$432 \times 200 \times 16$	–
BatchNormalization	–	$432 \times 200 \times 16$	64
LReLU	–	$432 \times 200 \times 16$	–
Conv2D	1×1 / 1×1	$432 \times 200 \times 32$	544
MFM	–	$432 \times 200 \times 16$	–
BatchNormalization	–	$432 \times 200 \times 16$	64
LReLU	–	$432 \times 200 \times 16$	–
Conv2D	3×3 / 1×1	$432 \times 200 \times 32$	4640
MFM	–	$432 \times 200 \times 16$	–
BatchNormalization	–	$432 \times 200 \times 16$	64
LReLU	–	$432 \times 200 \times 16$	–
Conv2D	1×1 / 1×1	$432 \times 200 \times 32$	544
MFM	–	$432 \times 200 \times 16$	–
Addition	–	$432 \times 200 \times 16$	–
Total	–	–	6K

Table 6 t-DCF cost function parameters assumed in ASVspoof 2019

Priors			SD costs		ASV costs	
π_{tar}	π_{non}	π_{spoof}	$C_{\text{miss}}^{\text{SD}}$	$C_{\text{fa}}^{\text{SD}}$	$C_{\text{miss}}^{\text{ASV}}$	$C_{\text{fa}}^{\text{ASV}}$
0.9405	0.0095	0.05	1	10	1	10

where $P_{\text{miss}}^{\text{SD}}(s)$ and $P_{\text{fa}}^{\text{SD}}(s)$ are the false rejection rate (FRR) and the false alarm rate (FAR) of the spoofing detection system at threshold s, respectively. The constants C_0, C_1 and C_2 can be calculated with the t-DCF costs, priors and the ASV system detection errors:

$$\begin{cases} C_0 = \pi_{\text{tar}} C_{\text{miss}}^{\text{ASV}} P_{\text{miss}}^{\text{ASV}} + \pi_{\text{non}} C_{\text{fa}}^{\text{ASV}} P_{\text{fa}}^{\text{ASV}} \\ C_1 = \pi_{\text{tar}} C_{\text{miss}}^{\text{SD}} - C_0 \\ C_2 = C_{\text{fa}}^{\text{SD}} \pi_{\text{spoof}} \left(1 - P_{\text{miss,spoof}}^{\text{ASV}}\right) \end{cases} \tag{5}$$

In (5), $C_{\text{miss}}^{\text{SD}}$, $C_{\text{fa}}^{\text{SD}}$, $C_{\text{miss}}^{\text{ASV}}$ and $C_{\text{fa}}^{\text{ASV}}$ are the costs of the spoofing detection and ASV systems respectively for rejection (miss) of a positive (Bona fide or target) trial and false acceptance (fa) of a negative (spoof or nontarget) trial. Furthermore, we assert a priori probabilities of target (π_{tar}), nontarget (π_{non}) and spoof (π_{spoof}) classes. Note that $\pi_{\text{tar}} + \pi_{\text{non}} + \pi_{\text{spoof}} = 1$. In this work, we adopt the same costs and prior probabilities used in the ASVspoof 2019 challenge, which are shown in Table 6.

In this work, the normalized t-DCF is adopted as the primary performance metric and given as below:

$$\text{t-DCF}_{\text{norm}}(s) = \frac{C_1}{C_2} P_{\text{miss}}^{\text{SD}}(s) + P_{\text{fa}}^{\text{SD}}(s) \tag{6}$$

Note that in the ASVspoof 2019 challenge, the decision scores for the baseline ASV system were not made available during the evaluation. The final t-DCFs of every spoofing detection system were calculated by the organizers with the spoofing detection scores submitted from all teams and the ASV scores pre-calculated. In this work, the t-DCFs of the proposed SR-ASV system are computed by the resultant ASV and spoofing detection scores and used for assessing the performance of spoofing detection.

4.4.2 Equal Error Rate (EER)

The EER is used to predetermine the threshold values for the FAR and the FRR. When these two rates are equal, the corresponding value is referred to as the EER. The lower the EER value, the higher the accuracy of a biometric system. In this work, EER is used to measure the system performance for both the spoofing detection and the ASV tasks.

4.5 Baseline and Benchmark Systems

The BS1 system [37] is used as the baseline integrated system in this work. In addition, the BM1 system [21], which is a well-performing joint decision-based integration system, is employed as a benchmark system to compare with the proposed approach.

For spoofing detection, we adopt several cepstral coefficients-based systems, including the Constant Q Cepstral Coefficient (CQCC), the Linear Frequency Cepstral Coefficient (LFCC), the Short-time Fourier Transform Cepstral Coefficient (SFTCC) and the Inverted Mel-Frequency Cepstral Coefficient (IMFCC). We also included benchmark systems using time-frequency-based features such as those based on the CQT and STFT. The BS2 [36] and BS3 [2] are the two baseline systems and BM2-BM5 [17, 18, 28] are the four top performing benchmark systems submitted to the ASVspoof 2019 challenge.

The baseline ASV system used is from the ASVspoof 2019 challenge [38]. It uses x-vector speaker embeddings [34] together with a PLDA back-end [27]. More detailed configurations for the baseline ASV system are given in [18, 38].

5 Experimental Results and Analysis

5.1 Results on the ASVspoof 2017 Corpus

The comparison between different integrated systems on the evaluation subset of the ASVspoof 2017 corpus is given in Table 7. The baseline system BS1 adopts the

Table 7 EERs(%) and t-DCF comparison between different integrated systems on the evaluation subset of the ASVspoof 2017 Version 2.0 corpus

Systems	ASV	SD		Average
		EER	t-DCF	
BS1	4.71	18.11	–	11.41
BM1	–	–	–	8.97
SR-ASV (CQT)	3.16	8.22	0.2022	5.69
SR-ASV (LLFB)	3.02	8.76	0.2173	5.89
Fusion	**2.87**	**8.05**	**0.1974**	**5.46**

ASV denotes the speaker verification and SD denotes the spoofing detection task. Results not provided by the authors are denoted by '–'

Table 8 The t-DCF and EER results for baseline systems and the proposed SR-ASV system on the LA partition for the spoofing detection task

Systems	Front-ends	Back-ends	Dev.		Eval.	
			t-DCF	EER	t-DCF	EER
AS1	CQT	CNN	0.0003	0.03	0.1127	5.04
AS2	LLFB	CNN	0.0057	0.19	0.1863	6.92
BS2	CQCC	GMM	0.0123	0.43	0.2366	9.57
BS3	LFCC	GMM	0.0663	2.71	0.2116	8.09
BM2	STFCC	GMM	0.0000	0.00	0.1400	5.97
BM3	IMFCC	GMM	0.0002	0.03	0.2198	9.49
BM4	FFT	LCNN	0.0009	0.04	0.1028	**4.53**
BM5	CQT	LCNN	0.0000	0.00	0.1014	4.58
SR-ASV	CQT	CNN	0.0000	0.00	**0.1009**	4.55
SR-ASV	LLFB	CNN	0.0005	0.02	0.1034	4.60
Fusion	–	–	0.0000	0.00	**0.0517**	**1.73**

Information not applicable is denoted by '–'

conventional method with the Gaussian back-end fusion [37], while the BM1 system applies the multi-task learning approach based on contrastive loss [21]. For the ASV task, our proposed SR-ASV system with the LLFB feature achieves the best EER of 3.02%. Furthermore, for the spoofing detection system, the proposed system with the CQT feature can provide a low EER of 8.22% and a t-DCF of 0.2022 on the evaluation subset. The results can be further improved by fusing the proposed systems with the Bosaris toolkit.[9] From the fusion of the SR-ASV systems using CQT and LLFB features, an averaged EER of 5.46% can be obtained.

5.2 Results on the ASVspoof 2019 Corpus

Since the ASVspoof 2019 corpus is newly released and contains both the logical and physical access attacks, we investigate more on this general database. The performance of our proposed SR-ASV and other state-of-the-art systems for the spoofing detection

[9] https://sites.google.com/site/bosaristoolkit/home.

Table 9 The t-DCF and EER results for baseline systems and the proposed SR-ASV system on the PA partition for the spoofing detection task

Systems	Front-ends	Back-ends	Dev.		Eval.	
			t-DCF	EER	t-DCF	EER
AS1	CQT	CNN	0.0522	1.50	0.7521	3.29
AS2	LLFB	CNN	0.0661	3.31	0.1380	5.35
BS2	CQCC	GMM	0.1953	9.87	0.2454	11.04
BS3	LFCC	GMM	0.2554	11.96	0.3017	13.54
BM2	STFCC	GMM	0.1462	7.09	0.2129	9.07
BM3	IMFCC	GMM	0.1464	7.04	0.2128	9.04
BM1	MFCC/CQCC	TDNN	–	–	–	8.55
BM4	FFT	LCNN	0.0759	3.92	0.6713	2.75
BM5	CQT	LCNN	0.0197	0.80	0.0295	1.23
SR-ASV	CQT	CNN	0.0186	0.73	**0.0287**	**1.15**
SR-ASV	LLFB	CNN	0.0346	2.02	0.3597	1.62
Fusion	–	–	0.0005	0.03	**0.0108**	**0.55**

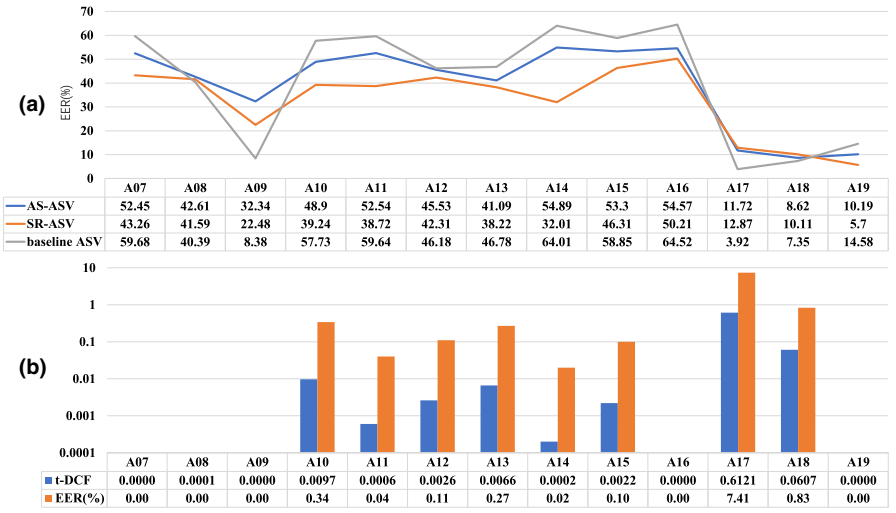
Information not applicable or not provided is denoted by '–'

task are presented in Tables 8 and 9 for the LA and PA partitions, respectively. The BS2 and BS3 are the two baseline systems used in the 2019 challenge. In addition to that, the BM2 to BM5 are the four top performing benchmark systems submitted to the challenge. Detailed introductions of these systems have been given in Sect. 4.5.

Compared to the traditional cepstral coefficients, experiment results for deep learning-based systems display a better detection ability and verification accuracy. This proves that a more comprehensive and discriminative representation can be learned by deep learning-based architectures with fine-tuned parameters. The best t-DCF results of single system on both partitions are obtained by the proposed SR-ASV system using CQT features. Moreover, the performance can be further improved by fusing the SR-ASV (CQT) and SR-ASV (LLFB) scores. The corresponding fusion results are listed in the last rows in Tables 8 and 9. Specifically, for the LA partition, the t-DCFs of baseline systems (BS2 and BS3) used in the challenge are reduced almost 80%. Furthermore, a significant improvement is also observed for the PA partition. The lowest t-DCF and EER on the evaluation subset of the PA partition are 0.0108 and 0.55, respectively.

Detailed performance comparison of the best single SR-ASV system and the challenge baseline ASV system across all spoofing attacks in the evaluation subsets are illustrated in Figs. 5 and 6 for the LA and PA partitions, respectively. Note that the results shown are all from single ASV systems. All the results of the challenge baseline ASV system are released in [38].

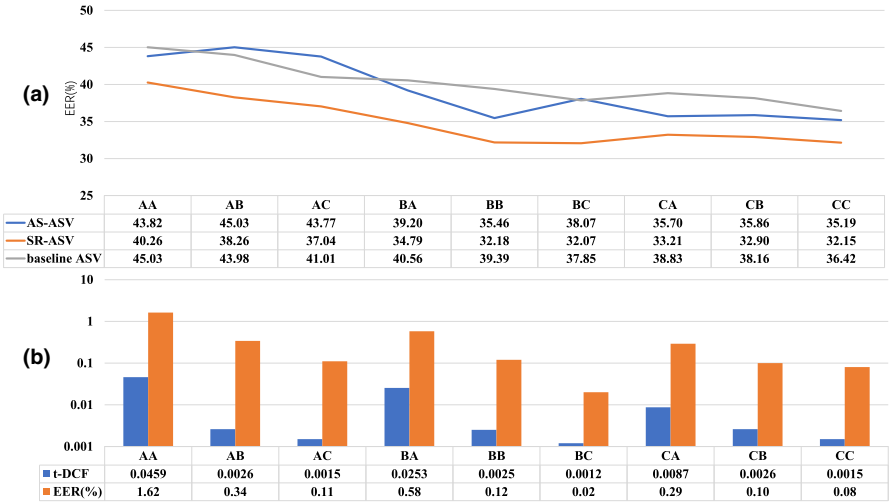
To investigate if the proposed system efficiently alleviates the negative influence from spoofing attacks to ASV, we present the EER results when the SR-ASV and the challenge baseline ASV system are under attack while no external anti-spoofing systems are deployed. The curves of the verification performance are given in (a) of



	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
AS-ASV	52.45	42.61	32.34	48.9	52.54	45.53	41.09	54.89	53.3	54.57	11.72	8.62	10.19
SR-ASV	43.26	41.59	22.48	39.24	38.72	42.31	38.22	32.01	46.31	50.21	12.87	10.11	5.7
baseline ASV	59.68	40.39	8.38	57.73	59.64	46.18	46.78	64.01	58.85	64.52	3.92	7.35	14.58

	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
t-DCF	0.0000	0.0001	0.0000	0.0097	0.0006	0.0026	0.0066	0.0002	0.0022	0.0000	0.6121	0.0607	0.0000
EER(%)	0.00	0.00	0.00	0.34	0.04	0.11	0.27	0.02	0.10	0.00	7.41	0.83	0.00

Fig. 5 Performance of ASV systems across varying LA spoofing attack types in the evaluation subset. In **a** the curves indicate the resultant EERs when the SR-ASV and the challenge baseline ASV system are under attack and no anti-spoofing detectors are used. In **b** the t-DCFs and EERs of the SR-ASV system on different LA attacks are provided



	AA	AB	AC	BA	BB	BC	CA	CB	CC
AS-ASV	43.82	45.03	43.77	39.20	35.46	38.07	35.70	35.86	35.19
SR-ASV	40.26	38.26	37.04	34.79	32.18	32.07	33.21	32.90	32.15
baseline ASV	45.03	43.98	41.01	40.56	39.39	37.85	38.83	38.16	36.42

	AA	AB	AC	BA	BB	BC	CA	CB	CC
t-DCF	0.0459	0.0026	0.0015	0.0253	0.0025	0.0012	0.0087	0.0026	0.0015
EER(%)	1.62	0.34	0.11	0.58	0.12	0.02	0.29	0.10	0.08

Fig. 6 Performance of ASV systems across varying PA spoofing attack types in the evaluation subset. In **a** the curves indicate the resultant EERs when the SR-ASV and the challenge baseline ASV system are under attack and no anti-spoofing detectors are used. In **b** the t-DCFs and EERs of the SR-ASV system on different PA attacks are provided

Figs. 5 and 6. As shown in (a) of Fig. 5, for the LA scenario there are several spoofing attacks that degrade ASV systems heavily and are difficult to detect, such as attacks A10, A13 and A15. They are all based on neural waveform or waveform concatenation skills. The proposed SR-ASV system performs worse than the baseline ASV system

on the A09, A17 and A18 attacks, which are mounted with vocoder and waveform filtering. This is probably caused by the different embedding extraction approaches used in the networks. In the baseline ASV system, the x-vector-based embedding is extracted from a statistical pooling layer with frame-level representations, while in our proposed system the embeddings are generated from utterance-level input features. The different perceptibilities of these embeddings lead to different sensibilities depending on the type of LA spoofing techniques. Specifically, artifacts caused by varying types of spoofing techniques can be hidden in the hierarchical structures of a speech signal. This raises an expectation of embeddings being robust to diverse spoofing attacks, and being sufficiently generalized for both the speaker verification and the spoofing detection tasks. In this work, embeddings are extracted from a multi-task learning system and these embeddings contain a more robust task-specific representation than those extracted from single-task frame-level x-vector systems. Furthermore, the utterance-level information included in the proposed embeddings help to capture the discontinuous artifacts hidden in waveform concatenated-based spoof attacks. In general it is the temporal dynamics in speech signals which improve the performance of x-vector-based speaker recognition systems. While in spoofing attacks, especially for the LA scenario, it is common for independent and irregular artifacts to be produced which are not present in Bona fide speech. This has always been used as an important clue to detect spoofing attacks and is the reason for extracting utterance-level embeddings in this work.

Interestingly, A17, a VAE-based voice conversion with waveform filtering, poses little threat to ASV systems while it is the most difficult to detect (as seen in (b) of Fig. 5). This indicates that the spoofed speech generated from A17 is more of a threat as it tends to be verified as genuine target samples by the ASV systems under attack.

Compared to the LA partition, the EER curves of these two ASV systems in (a) of Fig. 6 on the PA scenario are more consistent and stable. The performance gap between tasks of spoofing detection and speaker verification is fixed across all the nine replay configurations, i.e. replay attacks. When high-quality replay attacks are mounted, the EERs of these two ASV systems increase expectedly. This confirms that the quality of replay attack is the principal factor to be considered in anti-spoofing countermeasures. The overall performance of the SR-ASV system is better than the baseline ASV system. This indicates that the proposed integrated framework can improve the resistance to spoofing attacks for ASV.

To assess the spoofing detection performance of the proposed integrated system on different attacks, the resultant t-DCFs and EERs are given in (b) of Figs. 5 and 6. For cases of some unknown spoofing attacks in the LA partition, the performance is degraded at varying levels as shown in (b) of Fig. 5. The proposed system performs poorly especially for the A17 and A18 spoofing attacks, which apply voice conversion with waveform filtering and vocoder. In contrast, for spoofed samples generated by speech synthesis techniques (A07–A12), relatively stable results are achieved.

In (b) of Fig. 6, a clear trend on the t-DCF and EER can be seen with different replay attacks. The system performance is mainly affected by the replay configurations which include replay device quality, distances to the original speakers and to ASV system and reverberation characteristics. Poor system performance is observed on high-quality

replay attacks. This type of attack is mounted by using high-quality loudspeakers and recorders at a small distance to talkers in a quiet room.

5.3 Ablative Study for the Proposed System

To explore the cause of the performance gains achieved and for a more comprehensive justification of the proposed SR-ASV system, we present the results for single spoofing detection systems in Tables 8 and 9. The systems in the first two rows with name of AS1 (CQT) and AS2 (LLFB) are two spoofing detection systems for ablative study. The backbone network used in either system is similar to the proposed SR-ASV system. However, the task-specific layers for the ASV task are pruned to remove the benefits of MTL on the speaker verification. By comparing the AS1 and AS2 with SR-ASV systems, it is clearly shown that the embeddings extracted when the task-specific layers for ASV are used can help to further improve the performance on anti-spoofing. There is a 10–30% improvement on both the EER and t-DCF scores.

We also give an ablative study on the ASV task. By using the same pruning manner on the backbone MTL, we cut the task-specific layers for the spoofing detection task to remove any influence on the ASV. The results of this single ablative study ASV system (AS-ASV) are shown in Figs. 5a and 6a. Without the auxiliary information provided from the spoofing detection task, there is a clear performance gap between the AS-ASV and the SR-ASV systems. The AS-ASV system performs consistently worse compared to the proposed SR-ASV system on different types of spoofing attacks. Despite the steady performance on most other attacks, the increased EERs on the A17 and A18 still indicate the limitations of the proposed model for spoofed speech generated by novel voice conversion techniques.

6 Conclusion

In this paper, we investigated on integrated spoofing-robust speaker verification systems which can effectively resist varying attacks. We proposed a multi-task learning-based framework to jointly optimize the model used for extracting discriminative embeddings. This SR-ASV system was evaluated on the newly released ASVspoof 2017 Version 2.0 and 2019 corpora with experimental protocols of logical and physical access scenarios. By comparing with the other state-of-the-art systems on varying types of attacks, it was proved that the SR-ASV system can offer impressive performance for the spoofing detection and speaker verification tasks. However, for some powerful spoofing techniques derived from generative models and neural networks, the performance still needs to be improved. We also make the observation that, the different detection performance across the different types of attacks motivates the demand for more research on generalized detection systems at the forefront of cutting-edge spoofing technology. For example, exploring deep learning-based systems with both a stronger modeling ability which incorporates improved feature engineering is one of the future works. In addition, deploying anti-spoofing systems, and especially integrated systems, on practical hardware for real-time applications, is still in its infancy.

It is important to deploy and evaluate the proposed integration system to measure if there is a performance gap between the laboratory simulations and practical implementations. Therefore, another future work is exploring the possibility of implementing integrated systems in practical applications. To make it possible, general methods such as reducing the number of model parameters and model quantization should be explored. .

Acknowledgements We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used for this research.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability statement The data that support the findings of this study are openly available at https://github.com/zhaoyj1122/SRASV.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. F. Alegre, A. Amehraye, N. Evans, Spoofing countermeasures to protect automatic speaker verification from voice conversion, in *2013 IEEE International Conference on Acoustics (Speech and Signal Processing (ICASSP)* (IEEE, 2013), pp. 3068–3072
2. F. Alegre, R. Vipperla, A. Amehraye, N. Evans, A new speaker verification spoofing countermeasure based on local binary patterns, in *Interspeech* (2013), pp. 940–944
3. C. Chen, A. Ross, A multi-task convolutional neural network for joint iris detection and presentation attack detection, in *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (IEEE, 2018), pp. 44–51
4. D. Chen, B.K.W. Mak, Multitask learning of deep neural networks for low-resource speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. TASLP **23**(7), 1172–1183 (2015)
5. P.L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, I. Saratxaga, Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE/ACM Trans. Audio Speech Lang. Process. **20**(8), 2280–2290 (2012)
6. B. Dhanush, S. Suparna, R. Aarthy, C. Likhita, D. Shashank, H. Harish, S. Ganapathy, Factor analysis methods for joint speaker verification and spoof detection, in *2017 IEEE International Conference on Acoustics (Speech and Signal Processing (ICASSP)* (IEEE, 2017), pp. 5385–5389
7. N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, P. De Leon, Speaker recognition anti-spoofing, in *Handbook of Biometric Anti-spoofing* (Springer, 2014), pp. 125–146
8. N.W. Evans, T. Kinnunen, J . Yamagishi, spoofing and countermeasures for automatic speaker verification, in *Interspeech* (2013), pp. 925–929
9. A. Hadid, N. Evans, S. Marcel, J. Fierrez, Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. IEEE Signal Process. Mag. **32**(5), 20–30 (2015)

10. K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1026–1034
11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778
12. G. Heigold, I. Moreno, S. Bengio, N. Shazeer, End-to-end text-dependent speaker verification, in *2016 IEEE International Conference on Acoustics (Speech and Signal Processing (ICASSP)* (IEEE, 2016), pp. 5115–5119
13. Y. Jung, Y. Kim, H. Lim, H. Kim, Linear-scale filterbank for deep neural network-based voice activity detection, in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)* (IEEE, 2017), pp. 1–5
14. T. Kambič, *Multi-task Learning for Joint Face Recognition and Presentation Attack Detection, Tech. rep.* (University of Ljubljana, 2019)
15. A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 7482–7491
16. T. Kinnunen, K.A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, D.A. Reynolds, t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification, in *Proceedings of the Odyssey 2018 the Speaker and Language Recognition Workshop* (Springer, Berlin, 2018), pp. 312–319. https://doi.org/10.21437/Odyssey.2018-44
17. G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, V. Shchemelinin, Audio replay attack detection with deep learning frameworks, in *Interspeech* (2017), pp. 82–86
18. G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, A. Kozlov, STC Antispoofing Systems for the ASVspoof2019 Challenge, in *Interspeech* (2019), pp. 1033–1037
19. K.A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D.v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, et al., The reddots data collection for speaker recognition, in *Interspeech* (2015), pp. 2996–3000
20. C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, Z. Zhu, Deep speaker: an end-to-end neural speaker embedding system (2017). arXiv preprint arXiv:1705.02304
21. J. Li, M. Sun, X. Zhang, Y. Wang, Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss. IEEE Access **8**, 7907–7915 (2020)
22. R. Li, M. Zhao, Z. Li, L. Li, Q. Hong, Anti-spoofing speaker verification system with multi-feature integration and multi-task learning, in *Interspeech* (2019), pp. 1048–1052
23. W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 212–220
24. H. Muckenhirn, P. Korshunov, M. Magimai-Doss, S. Marcel, H. Muckenhirn, P. Korshunov, M. Magimai-Doss, S. Marcel, Long-term spectral statistics for voice presentation attack detection. IEEE/ACM Trans. Audio Speech Lang. Process. TASLP **25**(11), 2098–2111 (2017)
25. A.B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, Speech recognition using deep neural networks: a systematic review. IEEE Access **7**, 19143–19165 (2019)
26. S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, V. Shchemelinin, On deep speaker embeddings for text-independent speaker recognition, in *Proceedings of the Odyssey 2018 the Speaker and Language Recognition Workshop* (2018), pp. 378–385
27. S.J. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2007), pp. 1–8
28. M. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection, in *Interspeech* (2015), pp. 2087–2091
29. M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, Z.H. Tan, Integrated spoofing countermeasures and automatic speaker verification: an evaluation on asvspoof 2015, in *Interspeech* (2016), pp. 1700–1704
30. M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, K.A. Lee, Introduction to voice presentation attack detection and recent advances, in *Handbook of Biometric Anti-Spoofing* (Springer, 2019), pp. 321–361

31. H.J. Shim, J.W. Jung, H.S. Heo, S.H. Yoon, H.J. Yu, Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes, in *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (IEEE, 2018), pp. 172–176
32. D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, S. Khudanpur, Deep neural network-based speaker embeddings for end-to-end speaker verification, in *2016 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2016), pp. 165–170
33. D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification, in *Interspeech* (2017), pp. 999–1003
34. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in *2018 IEEE International Conference on Acoustics (Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 5329–5333
35. A. Søgaard, Y. Goldberg, Deep multi-task learning with low level tasks supervised at lower layers, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2 (Short Papers) (2016), pp. 231–235
36. M. Todisco, H. Delgado, N. Evans, Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. Comput. Speech Lang. **45**, 516–535 (2017)
37. M. Todisco, H. Delgado, K.A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, J. Yamagishi, Integrated presentation attack detection and automatic speaker verification: common features and Gaussian back-end fusion, in *Interspeech* (2018), pp. 77–81
38. M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T.H. Kinnunen, K.A. Lee, ASVspoof 2019: future horizons in spoofed and fake audio detection, in *Interspeech* (2019), pp. 1008–1012
39. R. Togneri, D. Pullella, An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag. **11**(2), 23–61 (2011)
40. P. von Platen, F. Tao, G. Tur, Multi-task siamese neural network for improving replay attack detection (2020). arXiv preprint arXiv:2002.07629
41. X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. **13**(11), 2884–2896 (2018)
42. Z. Wu, A. Larcher, K.A. Lee, E. Chng, T. Kinnunen, H. Li, Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints, in *Interspeech* (2013), pp. 950–954
43. Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, S. King, SAS: A speaker verification spoofing database containing diverse attacks, in *2015 IEEE International Conference on Acoustics (Speech and Signal Processing (ICASSP)* (IEEE, 2015), pp. 4440–4444
44. X. Xia, R. Togneri, F. Sohel, Y. Zhao, D. Huang, Multi-task learning for acoustic event detection using event and frame position information. IEEE Trans. Multimedia **22**(3), 569–578 (2020)
45. Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, K. Yu, The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge, in *Interspeech* (2019), pp. 1038–1042
46. Y. Zhao, X. Xia, R. Togneri, Applications of deep learning to audio generation. IEEE Circuits Syst. Mag. **19**(4), 19–38 (2019)