Circuits, Systems, and Signal Processing
https://doi.org/10.1007/s00034-023-02314-5

Efficient Attention Branch Network with Combined Loss Function for Automatic Speaker Verification Spoof Detection

Amir Mohammad Rostami[1] · Mohammad Mehdi Homayounpour[1] · Ahmad Nickabadi[1]

Received: 2 July 2022 / Revised: 2 February 2023 / Accepted: 2 February 2023

Abstract

Many endeavors have sought to develop countermeasure techniques as enhancements on Automatic Speaker Verification (ASV) systems, in order to make them more robust against spoof attacks. As evidenced by the latest ASVspoof 2019 countermeasure challenge, models currently deployed for the task of ASV are, at their best, devoid of suitable degrees of generalization to unseen attacks. A joint improvement of components of ASV spoof detection systems including the classifier, feature extraction phase, and model loss function may lead to a better detection of attacks by these systems. Accordingly, the present study proposes the Efficient Attention Branch Network (EABN) architecture with a combined loss function to address the model generalization to unseen attacks. The EABN is based on attention and perception branches. The attention branch provides an attention mask that improves the classification performance and at the same time is interpretable from a human point of view. The perception branch, is used for our main purpose which is spoof detection. The new EfficientNet-A0 architecture was optimized and employed for the perception branch, with nearly ten times fewer parameters and approximately seven times fewer floating-point operations than the SE-Res2Net50 as the best existing network. The proposed method on ASVspoof 2019 dataset achieved EER = 0.86% and t-DCF = 0.0239 in the Physical Access (PA) scenario using the logPowSpec as the input feature extraction method. Furthermore, using the LFCC feature, and the SE-Res2Net50 for the perception branch, the proposed model achieved EER = 1.89% and t-DCF = 0.507 in the

✉ Mohammad Mehdi Homayounpour
homayoun@aut.ac.ir

Amir Mohammad Rostami
a.m.rostami@aut.ac.ir

Ahmad Nickabadi
nickabadi@aut.ac.ir

[1] Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran

Published online: 23 February 2023

Logical Access (LA) scenario, which to the best of our knowledge, is the best single system ASV spoofing countermeasure method.

Keywords Automatic speaker verification · Spoof detection · ASVspoof · Efficient attention branch network · Combined loss function · EfficientNet-A0

1 Introduction

Remote authentication has excited great interests in various academic circles and otherwise, given the increasing reliance on online applications as well as the onset of certain conditions such as the COVID-19 pandemic. Such circumstances call for an easy-to-use, accurate, and efficient authentication system. Along this thread, Automatic Speaker Verification (ASV) system and other biometric systems such as face recognition, electronic signatures, iris-based, and hybrid methods have been proposed as a means to satisfy user needs [11, 24]. Nevertheless, virtually all of these systems are vulnerable to spoof attacks (i.e., spoofable). A system based on face recognition for example, may be spoofed by simply displaying a person's image (photograph) to the system [10]. Likewise, a fingerprint system can be spoofed by copying a fingerprint. In particular, ASV systems are also vulnerable in the face of four types of attacks, including recording and replaying the voice of the authorized person (replay attack), text-to-speech systems that are trained with the voice of the targeted person, voice conversion systems, and speaker imitation [7]. The threats facing ASV systems in terms of spoof attacks are potentially high and may amass to serious implications [26]. In consequence, since 2015, ASVspoof challenges were held for research communities worldwide to try and enhance ASV systems so as to make them robust against spoofing attacks.

A total of four ASVspoof challenges have thus far been held, with the first instance in 2015, covering only speech synthesis and voice conversion (also called logical access scenario) attacks [32]. A variety of methods and systems were proposed and implemented by ASV organizers to produce spoof samples, exciting the interest of many researchers intrigued by both the challenge and the dataset provided therein. The second ASVspoof challenge held in 2017, focused more on the replay attack (also called physical access scenario) [6, 14]. In order to be able to test the performance of countermeasure systems in real conditions, the organizers produced the dataset in different environmental conditions and using different devices. Further comprehensive conditions were investigated in 2019 to account for all three attacks considered in previous challenges [30], ushering in the development of an extensive dataset using state-of-the-art voice conversion and speech synthesis systems. Spoofing samples in this challenge were more realistic and challenging in view of the improvements made to the spoof systems in previous years. For replay attacks, in particular, samples were produced with greater degrees of control, and a tandem detection cost function (t-DCF) metric was used as the primary metric to assess the efficiency of integrating countermeasures with ASV systems. The 2021 edition was considerably more complex than its predecessors, more challenging data that move ASVspoof nearer to more practical application scenarios and also add new deepfake task. Participants must only use 2019 edition train and development dataset to develop their models [34].

Inquiries made into the 2019 ASVspoof dataset results are suggestive of two primary drawbacks of the proposed methods. The first points to a lack of generalization and high error rate against unseen attacks, which is clearly observed given the difference between errors obtained for the training, development, and evaluation sets. In addressing this lack of generalization, numerous studies have tried to improve generalization by means of fusing several models (ensemble models) [15]. Such fusion and ensemble models and methods that use deep neural networks have led to considerable increases in model parameters as well as the necessary floating-point operations (FLOPS). Under such circumstances, it would be infeasible to use the proposed models in specific applications. This provides the required grounds for the integration of simple yet efficient countermeasure techniques with ASV systems to make them more robust. Moreover, the proposed body of research fails to provide a detailed understanding for how models detect spoof attacks or handle generalization issue. This ambiguity can be interpreted in terms of the incapacity of humans or rather human-oriented decision-making to differentiate between the spoofed and the bonafide samples detected by the final system. A detailed examination of this issue can provide further insight into the development of better systems.

The primary purpose of this work is to provide a model for detecting spoofing attacks on ASV systems. An interpretable attention mask in a new modular architecture is used for this purpose via the introduction of perception and attention branches in the model. Furthermore, for the first time in this domain, the EfficientNet-A0 [25] architecture was employed to achieve a system with low number of parameters and FLOPS. The proposed architecture along with the newly combined loss function and masks that provide a more human-oriented perspective, was used to obtain comparable and, in some cases, top-performing results in these spoofing attacks.

The following section provides a brief review of relevant studies conducted in recent years. The proposed countermeasures and the loss function are introduced in Sect. 3. Section 4 calls attention to the general configuration used for experiments and sect. 5 gives the analysis results along with a summarization of the work. This study is finally concluded in Sect. 6.

2 Related Work

This section reviews some of the research carried out on spoof attack detection, taking a look on the best-performing methods, as per results obtained on the ASVspoof 2019 dataset. Similar models and tasks were also investigated inclusive of new architecture and the application of attention mechanisms and attitude in the loss function.

Models proposed to assess the ASVspoof 2019 dataset can be categorized into two main classes: methods based on extraction and engineering of features and methods based on classifier architecture. Methods of the first category incorporate features such as Mel-filter Frequency Cepstral coefficients (MFCC), Inverted Mel-Filter Frequency Cepstral coefficients (IMFFC), Constant Q Cepstral Coefficients (CQCC), Group Delay (GD) gram, Instantaneous Amplitude (IA), Instantaneous Frequency

(IF), X-vectors, and features from deep learning models [1, 3, 13, 19, 23, 27, 29, 33]. Some methods also use raw signals to extract features using methods such as SincNet [35] or Variational Auto Encoder (VAE) [5]. The second category deals with a variety of classifiers such as Neural network-based methods including VGG [35], Squeeze-Excitation (SE), Residual network, Siamese networks [17], and recurrent networks [12], as well as other traditional GMM-based methods. Certain methods have also used end-to-end structures for this purpose.

Z. Wu et al. [31] propose a novel feature genuinization based light convolution neural network (LCNN) system for detection of synthetic speech attacks. They transform a genuine feature distribution more close to that of the genuine speech. They fed transformed features to proposed LCNN system for detecting synthetic speech attacks. In another work, X. Cheng et al. [4] proposed the replay detection system based on a novel CQT-based modify delay group (MGD) feature to utilize the phase of CQT. An 18-layer ResNeWt model is used to detect the replay attacks. Their models were evaluated on ASVspoof 2019 physical access challenge dataset and show a significant improvement on the ability to detect the distortion introduced by the playback device and the ability to detect the reverberation introduced by far-field recording, compared with CQCC-GMM baseline system.

Cheng-I Lai et al. proposed a deep model to obtain discriminative features in both time and frequency domains [16]. The proposed design includes a filter-based attention mechanism used to improve or ignore commonly extracted features implemented in the ResNet architecture to classify attended input maps. The reserved classifier used in their study (Residual Network) consists of a convolution layer equipped with dilated mechanism instead of a fully connected layer, which runs as an attentive filtering network; i.e., masks input features. The obtained results were suggestive of the relatively high performance of the model given the use of an attention mechanism to produce attention masks as well as an appropriate classifier.

X. Li et al. attempted to use the Res2Net architecture, which has achieved significant results in various computer vision tasks [20]. They proposed a new Res2Net architecture by revisioning ResNet blocks to allow for multi-scale features. In a Res2Net architecture, input feature maps of a block are divided into several groups of channels with a similar residual structure to the original ResNet. Using channels, feature map sizes can be different, increasing the covered area, and thereby yielding features with different scales. This modification improves system performance and the model's generalization against unseen attacks. In addition, using this architecture could reduce the size or number of model parameters relative to the original ResNet structure while improving model performance. The obtained results show that the Res2Net50 model outperforms the ResNet34 and ResNet50 models in both physical and logical access scenarios. They also showed that integrating the block with Squeeze-and-excitation (SE), which produces SE-Res2Net blocks, leads to better performance. Figure 1 illustrates the architecture and structure of these blocks. Significant results were also obtained in both scenarios for the proposed SE-Res2Net50 network based on SE-Res2Net blocks and Constant-Q Transform (CQT) feature. The network proposed in this work has nearly 0.9 million parameters, which is relatively small compared to other architectures. However, the main drawback to the model is the high number of

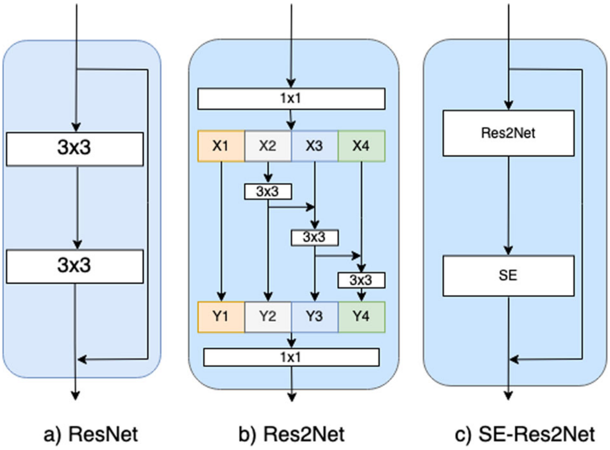

Fig. 1 ResNet, Res2Net, and SE-Res2Net blocks [20]

FLOPS, which leads to increased runtime in the inference phase due to the multiplicity of blocks and the structure of SE-Res2Net.

Zhang et al. focused on logical attacks in their work [36], explaining the lack of model generalization against unseen attacks as caused by the formulation of the spoof detection problem as a binary classification. The difficulty with using a binary classifier can be interpreted in terms of the distribution of training and test data for spoof and bonafide samples as not being the same. More specifically, samples in the test set generated by new systems or conditions not found in training data cause differences in distribution; which, however, is not the case for bonafide samples. The problem was, therefore, redefined as a one-class classification problem, where the distribution of a target class for a specific problem should be the same in both training and test datasets, irrespective of whether other classes have similar distributions or not. In such cases, the primary objective is to obtain the bonafide distribution and define a rigid decision boundary around it so that unseen samples from other classes cannot cross that decision boundary. To this aim, a one-class softmax loss function was incorporated for learning a feature space that can map bonafide samples in a dense space, while maintaining a good margin with spoofing samples. Finally, by means of the ResNet-18 network and the Linear-filter Frequency Cepstral coefficient (LFCC) features, the authors succeeded in attaining top-performing results for logical access attacks.

3 Proposed Model

3.1 Network Architecture

The overall architecture of the proposed network was designed with three main objectives in mind: a)- that the architecture be small enough to explicate an appropriate number of parameters, b)-maintaining an acceptable runtime in order to achieve satisfactory performance in most ASV applications; c)- the interpretability of the designed

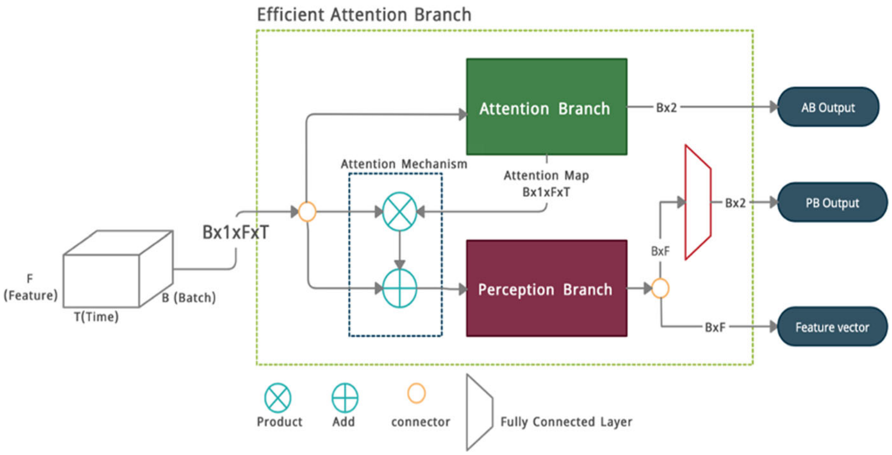

Fig. 2 Proposed Efficient Attention Branch Network architecture

architecture by humans. To put differently, the architecture was required to somehow express what discriminates bonafide speech from speech made in a spoof attack as a means to improve systems in the future; lastly, the model was configured to emulate comparable performance to relevant classifiers used for this purpose.

To achieve all these goals, the Efficient Attention Branch Network (EABN) was proposed in this study. The intended framework adopts a well-performed Attention Branch Network [8] in computer vision as the main idea for the EABN architecture. As shown in Fig. 2, this network consists of two branches of attention and perception. The attention branch seeks to improve the performance of the perception branch by means of producing an attention mask, which is then applied to make the discriminative parts of the input feature map more. In addition, in order to improve the performance of the perception branch, masks produced by the attention branch are also interpretable from a human point of view. The primary work load is performed in the perceptual branch, where the probability output of each class is produced.

3.2 Attention Branch

The attention branch itself comprises of two main parts, as shown in Fig. 3. As can be observed, the input feature map is initially fed into the attention branch, which uses four consecutive basic blocks to extract the appropriate features and to convert the input features to 16-feature maps. The blocks consist of two convolution layers with 3×3 kernels, which are then linked to the batch normalization layer. In addition to feature extraction, the first convolution layer also increases the feature map size, while the second convolution layer exclusively handles the feature extraction process. The obtained feature maps are eventually transformed from a 16-size map to a single feature using a convolution layer with a 1×1 kernel, which then goes through a softmax layer to yield the final output attention mask.

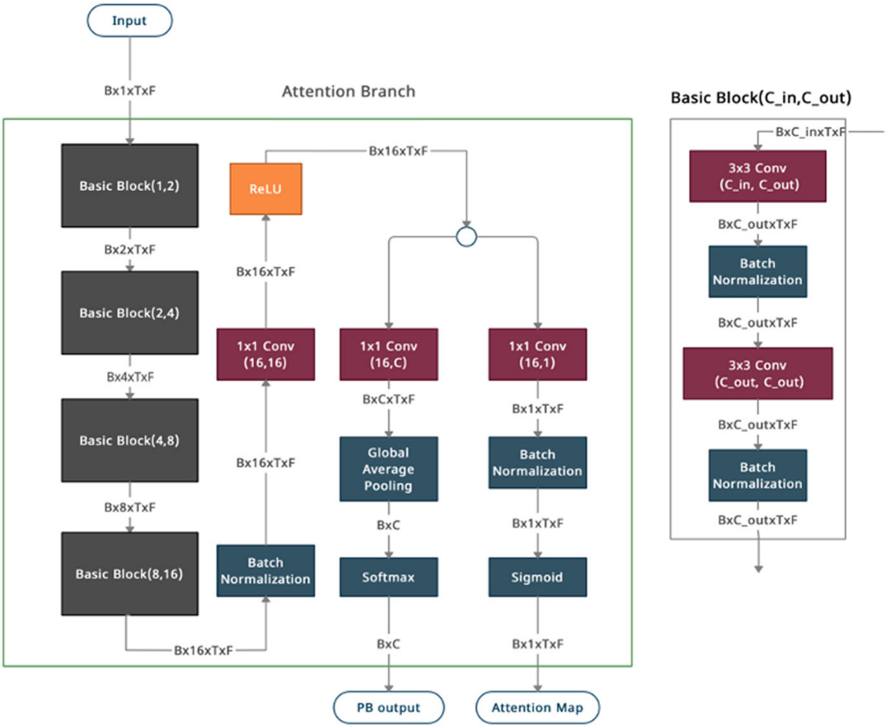

Fig. 3 Proposed architecture for Attention branch

The other branch produces a human-interpretable attention mask. This is carried out by using a convolution operation to transform the 16 feature maps into maps with the same size as the number of classes for the problem which in this study includes the bonafide and spoof classes. Then, using a global average pooling layer, these two feature maps are converted to a 2×1 tensor. Finally, by applying softmax, the probability of a feature map belonging to each class is obtained. These probabilities are later used in the optimization process for the proposed combined loss function. Through the process of optimization, feature maps are generated so that in addition to help the perception branch, they can be used for classification and are interpretable from a human perspective.

3.3 Perception Branch

The perception branch can be constructed by almost any classifier. However, as the primary objectives of this study call for low number of parameters, low runtime, and good performance in network design the EfficientNet architecture is employed. The EfficientNet has been used as a high performing model in image classification tasks [28] and speech processing tasks such as speech recognition [21] and keyword spotting [25]. The fundamental architecture of the EfficientNet family is called EfficientNet-

B0, which has about 4 million parameters. This number of parameters is not suitable for the target applications of this study. Alternatively, the approach introduced in the EfficientNet-Absolute Zero (EfficientNet-A0) work [25], which applies the reverse of the compound scaling method, was used. The scaling method (S) is designed to shrink a base model (M) by decreasing the depth (α), width (β), and resolution of the input image (γ), simultaneously. A formulation of this method is given below as an optimization problem, in which the goal is to satisfy the intended conditions so that the final model has the best performance.

$$\begin{aligned} &\max_{d,r,w} \text{Accuracy}(S(M,\ d,\ r,\ w)) \\ &\quad \text{s.t.} \\ &\qquad \frac{1}{20} \leq \alpha \cdot \beta^2 \cdot \gamma^2 \leq \frac{1}{16}, \\ &\qquad 0.2 \leq \alpha,\ \beta \leq 0.6,\ \gamma = 2 \end{aligned} \tag{1}$$

The two parameters α and β are set by applying a grid search on intervals $[0.2-0.6]$ with steps of 0.005. Eventually, 19 models were evaluated with a small subset of samples, with parameters α and β set at values 0.2 and 0.25, respectively. γ was also set at ≈ 2, given the input image size (513×400) and EfficieNet-B0 input-size of 256×256. Figure 4 illustrates the final model obtained for the perception branch with 95,000 parameters. The input to this branch is $m(x_i)$, where x_i is the input image for the i^{th} sample and is calculated from the following equation:

$$m\left(x_i\right) = \left(1 + g\left(x_i\right)\right) \times x_i \tag{2}$$

where $g(x_i)$ is the attention mask produced for the i^{th} sample by attention branch. The output of this network is a vector of length 256, which represents the embedded vector of the input image and is applied for two scenarios. The first one uses the vector along with a fully connected layer and the softmax layer to yield probabilities for each individual sample. The second scenario uses the vector as input to a loss function. Thus, samples are embedded in a 256-dimensional space in a most distinctive way.

3.4 Loss Function

To train model parameters, a combined loss function (equation 3) was used to account for all study objectives.

$$L_{\text{total}} = L_{\text{PB}} + \lambda_{\text{AB}} L_{\text{AB}} \tag{3}$$

To train an attention branch capable of producing interpretable masks, the AB_{output} was used as input to a weighted Cross-Entropy (CE) loss function (L_{AB} in equation 3). It should be noted that by introducing the proposed loss function with coefficient λ_{AB}, values in equation 3 are altered. Proceeding forward, the Triplet Center Loss (TCL) (L_{tc}) function is used to train the embedding vectors. TCL works in the same way as the triplet loss function, except that it no longer needs to mine triplets for training, and

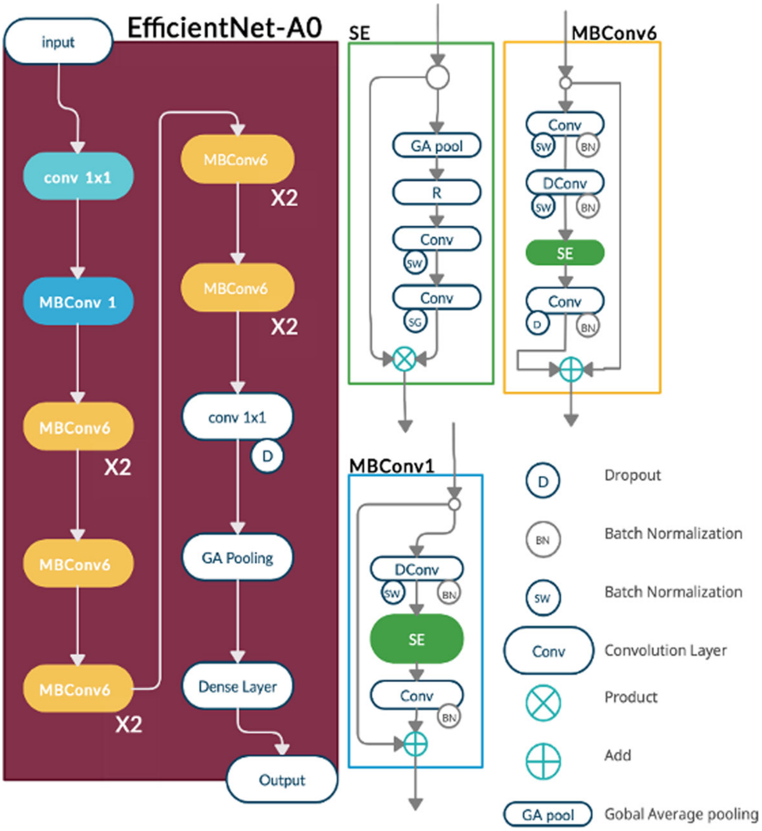

Fig. 4 Proposed architecture for perception branch operating via the reverse compound scaling method

this difference makes the training process faster and more stable. This loss function considers center points for each class in the problem, which are initially assigned random values. The loss function causes the samples of each class to move closer to the centers of their classes and away from the centers of other classes. The two centers used in this study to represent spoof and bonafide samples are C_{spoof} and C_{bonafide}, respectively. The goal here was to ensure that bonafide samples are close to the center of their respective target class, C_{bonafide}, and away from C_{spoof}. As a result, samples of a specific class in a dense space are closer to each other; representing feature vectors embedded for each sample in the desired space. L_{tc} can be obtained for the x_i sample as follows:

$$L_{tc}\left(x_i\right) = \begin{cases} \max\left(D\left(f_i, C_{\text{spoof}}\right) + m - D\left(f_i, C_{\text{bonafide}}\right), 0\right) \times w_{\text{spoof}} & \text{if} \quad x_i \in \{\text{ spoof samples }\} \\ \max\left(D\left(f_i, C_{\text{bonafide}}\right) + m - D\left(f_i, C_{\text{spoof}}\right), 0\right) \times w_{\text{bonafide}} & \text{if} \quad x_i \in \{\text{ bonafide samples }\} \end{cases} \tag{4}$$

where f_i represents the feature vector obtained from input x_i, measured in distance. w represents weights considered for each class with respect to the unbalanced number

Table 1 Summary of the ASVspoof2019 dataset

Partition	PA		LA	
	# Spoof	# Bonafide	# Spoof	# Bonafide
Train	48600	5400	22800	2580
Dev	24300	5400	22296	2548
Eval	116640	18090	63882	7355

Table 2 t-DCF hyperparameters value

Attack Type	Probabilities			ASV costs		Countermeasure costs	
	π_{tar}	π_{non}	π_{spoof}	$C_{\text{fa}}^{\text{asv}}$	$C_{\text{miss}}^{\text{asv}}$	$C_{\text{fa}}^{\text{cm}}$	$C_{\text{miss}}^{\text{cm}}$
PA	0.9405	0.0095	0.05	10	1	10	1
LA	0.9405	0.0095	0.05	10	1	10	1

of instances of the classes. m represents the margin that causes the distance of samples of the same class to be at least m less than the samples of the opposite class. The cost function is further augmented with a cross-entropy function to improve the final results and maintain the stability of the optimization. Given that spoof samples have different difficulties, the focal loss obtained from equation below is used instead of cross-entropy.

$$L_{\text{focal}}\ (p_t) = -\alpha_t\ (1 - p_t)^{0.005} \log\ (p_t) \tag{5}$$

finally the cost function of the perception branch is calculated from equation 6.

$$L_{PB}\ (x_i) = L_{tc}\ (x_i) + \lambda_{\text{focal}}\ L_{\text{focal}}\ (x_i) \tag{6}$$

4 Experimental Configuration

4.1 Dataset and Evaluation Metrics

The proposed method was evaluated using the ASVspoof 2019 and 2021 dataset, which includes two scenarios: physical access (PA) and logical access (LA). Details of this dataset are shown in Table 1. Furthermore, considering that one of the objectives of this research is the simultaneous use of countermeasure and ASV system, the tandem-detection cost function (t-DCF) and the equal error rate (EER) metrics are used. This metric was introduced as the primary evaluation metric of the 2019 challenge, which is calculated as:

$$\text{t} - \text{DCF}(s) = C_1 P_{\text{miss}}^{\text{cm}}(s) + C_2 P_{\text{fa}}^{\text{cm}}(s) \tag{7}$$

where $P_{\mathrm{fa}}^{\mathrm{cm}}(s)$ and $P_{\mathrm{miss}}^{\mathrm{cm}}$ are the false acceptance error rate and the false rejection error rate of the countermeasure, respectively. Considering the threshold, s, values for the two error rates can be obtained as follows:

$$P_{\mathrm{miss}}^{\mathrm{cm}}(s) = \frac{\#\{\text{ bona fide trials with CM score } \leq s\}}{\#\{\text{ Total bona fide trials }\}} \tag{8}$$

$$P_{\mathrm{fa}}^{\mathrm{cm}}(s) = \frac{\#\{\text{ spoof trials with CM score } > s\}}{\#\{\text{ Total spoof trials }\}} \tag{9}$$

The two constants C_1 and C_2 represent the predefined cost for the errors, which are determined based on prior probabilities as shown below:

$$\begin{cases} C_1 = \pi_{\mathrm{tar}}\left(C_{\mathrm{miss}}^{\mathrm{cm}} - C_{\mathrm{miss}}^{\mathrm{asv}} P_{\mathrm{miss}}^{\mathrm{asv}}\right) - \pi_{\mathrm{non}} C_{\mathrm{fa}}^{\mathrm{asv}} P_{\mathrm{fa}}^{\mathrm{asv}} \\ C_2 = C_{\mathrm{fa}}^{\mathrm{cm}} \pi_{\mathrm{spoof}}\left(1 - P_{\mathrm{miss,spoof}}^{\mathrm{asv}}\right) \end{cases} \tag{10}$$

Here, $C_{\mathrm{miss}}^{\mathrm{asv}}$ represents the cost incurred by the error of the ASV system for the false rejection error rate of the genuine person, and $C_{\mathrm{fa}}^{\mathrm{asv}}$ represents the false acceptance error rate when ASV authorizes the wrong person. Each countermeasure error also corresponds to two costs; $C_{\mathrm{miss}}^{\mathrm{cm}}$, which indicates the cost in recognizing a bonafide sample as a spoof, and $C_{\mathrm{fa}}^{\mathrm{cm}}$, which indicates a mistake in accepting a sample produced by a spoof system as bonafide. In addition, the probability of occurrence of any class of genuine (π_{tar}), non-target or imposter (π_{non}) and spoof attack (π_{spoof}) are also considered with the condition $\pi_{\mathrm{tar}} + \pi_{\mathrm{non}} + \pi_{\mathrm{spoof}} = 1$. Cost and probability values are calculated as in Table 2.

4.2 Feature Extraction and Engineering

Based on past researches and works, a single acoustic feature is considered for each of the attacks. For the PA scenario, we use the logarithm of power spectrm (logPowSpec) with 25 ms frames, 10ms step size with 1024 samples (with zero padding applied if needed), using Hamming window. All the samples are first transformed into 4 s voice segments. To do this, samples that are less than 4 s are repeated to achieve a 4 s segment. Longer samples are also divided into 4 s segments with no overlap, and each segment is considered as an individual utterance. The final input form consists of a logPowSpec with 513×400 dimensions (512 logarithms of spectrum magnitudes with 1 being the DC component). For the LA scenario, the LFCC feature is extracted according to the procedure used in the base model presented in the ASVspoof 2019 challenge. Here, 20ms frames with 512 point Fast Fourier Transform are used along with first and second derivatives of 20 LFCC features. Finally, a two-dimensional tensor with dimensions of 60×400 was obtained.

As a further step, specAug technique [22] is applied for better training and generalization. This method works well for most of speech processing tasks, such as speaker verification, speech recognition, and keyword potting [25]. The method is implemented by applying zero masks on the time and frequency axis for each training sample with a probability of 0.25. The size of SpecAug Mask (band) is randomly selected between 20 and 80 frames on the time axis and 25 and 100 on the frequency axis. Also, for LFCC coefficients (with their derivatives), the size of zero mask on horizontal axis is between 20 and 80 frames and between 5 and 20 on vertical axis.

4.3 Perception Branch Models

In addition to the proposed EfficientNet-A0 model used in the perception branch, a SE-ResNet50 model was also used, which achieved significant results. The models were then compared in terms of both efficiency and performance, and the EABN modularity idea was evaluated accordingly.

4.4 Training Procedure

The final results obtained from experiments on small subsets of the ASVspoof 2019 dataset yielded values of 0.1, 0.005, and 32 for λ_{AB}, λ_{focal} and m, respectively. To optimize the loss function with assigned values, configurations for the SE-ResNet50 architecture were adopted. In the case of Adam optimization, β_1, β_2, and learning rate were obtained at 0.9, 0.98, and 10^{-9}, respectively. The learning rate initially drops linearly for the first 1000 steps and then decreases in proportion to the inverse of the square root of the number of steps. All models were trained with 40 epochs and the model with the lowest EER on the development set of the dataset was selected as the optimal choice. Batch-sizes were set at 64 and 128 when using EfficienNet-A0 as the perception branch module with LFCC and logPowSpec respectively. Due to the relatively greater number of parameters for the SE-Res2Net50 model compared to EfficientNet-A0, a batch-size of 8 was used for LFCC and logPowSpec features. The models were implemented on a GTX-1080ti GPU on Linux OS. The source code of our implementations based on Python and Pytorch is publicly available.[1]

5 Results

5.1 Evaluation of Perception Branch's Models

This section evaluates the overall architectural EABN and the EfficientNet-A0 network as a classifier for spoof detection. To investigate EABN performance, the EfficientNet-A0 and SE-ResNet50 architectures were used for the perception branch, which have the lowest EER as a single model to the best of our knowledge. The results for both attacks are shown in Table 3. In the PA scenario, EfficientNet-A0 shows a better performance than SE-ResNet50 and has nearly ten times fewer parameters and seven times fewer FLOPS. This can be explained in terms of the enhanced performance of the EfficientNet-A0 model in extracting features from the LogPowerSpec. On the other hand, the SE-ResNet50 model performs better when LFCC feature are used for the LA scenario.

[1] https://github.com/AmirmohammadRostami/ASV-anti-spoofing-with-EABN

Table 3 Result of models used in perception branch and input features on ASVspoof 2019 evaluation dataset for PA and LA scenarios. K, M, and G represent Kilo, Mega, and Giga, respectively

#	Perception branch model	Input feature	#Parameters	#Flops	PA		LA	
					EER(%)	t-DCF	EER(%)	t-DCF
1	EfficientNet-A0	LFCC	95k	198 M	–	–	3.68	0.0931
2	EfficientNet-A0	LogPowSepc		1.696 G	0.86	0.0239	–	–
3	SE-Res2Net50	LFCC	964k	1.519 G	–	–	1.89	0.0597
4	SE-Res2Net50	LogPowSepc		12.929 G	0.98	0.2769	–	–

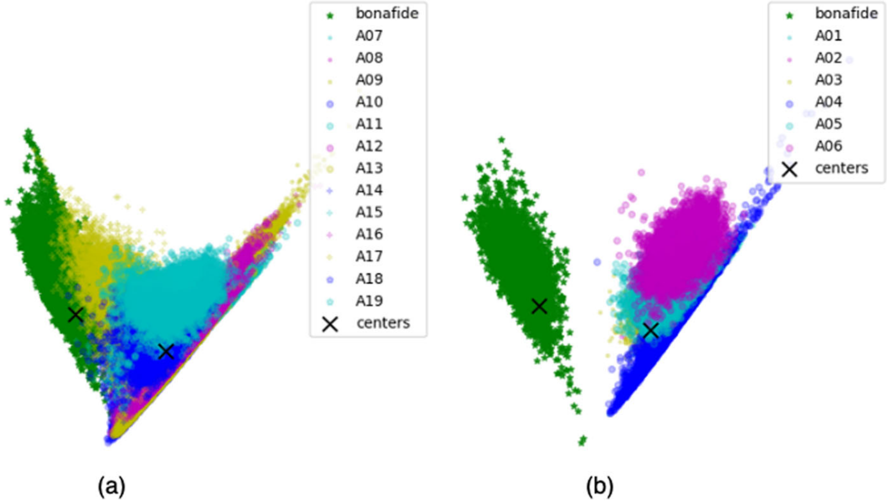

Fig. 5 Feature embedding visualization of our proposed loss function for evaluation (a) and training (b) sets of the ASVspoof 2019 LA attack. Features were reduced to 2-D space using PCA

5.2 Loss Function

The proposed combined loss function was used for the first time in this work to achieve a discriminative vector space to distinguish spoof samples from bonafide samples. More precisely, the triplet center loss was used to map input samples to a discriminative space. As shown in Fig. 5, the training samples mapping space is suitable for the classification problem. Examining test samples that include unseen attacks also demonstrates that the resulting space is reasonably discriminative. It can therefore be said that the model shows good generalization for unseen attacks. The best value for margin 32 was obtained in this study by testing three values of 16, 32, and 64 (for this margin please see the last paragraph of related works section).

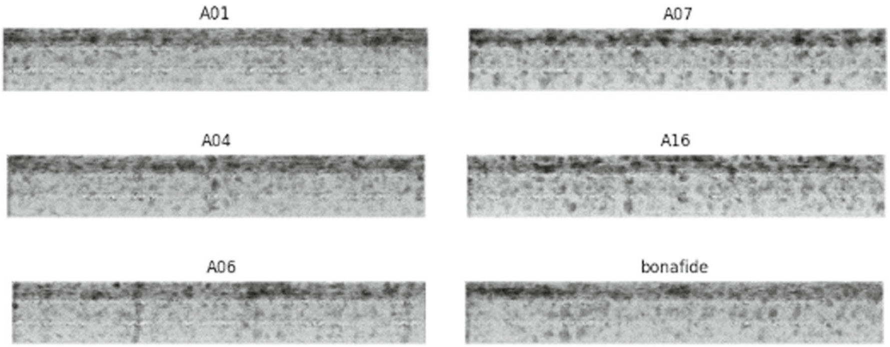

Fig. 6 Average of produced LFCC attention masks for some spoof attacks in ASVspoof 2019 evaluation

5.3 Attention Masks

One of the main concerns about the proposed architecture is to obtain attention masks that can be interpreted from a human point of view. This was investigated for the LFCC feature, with the averaged masks generated for all samples in the evaluation set shown in Fig. 6. Examining the LFCC feature masks obtained for different attack systems reveals that information corresponding to the second derivatives of LFCC coefficients are very effective in detecting spoof patterns.

For the logPowSpec mask, a few of samples from the evaluation set of the PA attack are shown in Fig. 7. The raw input features and results of the applied mask on them, which is input to the perception branch, are also shown in this figure. The masks blur or dominate some values at different frequencies. The mask points with lower values decrease the impacts of LogPowSpec values at frequencies that show lower capacity to discriminate spoof attacks from bonafide samples and vice versa. By examining the masks produced for physical access attacks, it can be understood that there is a lot of emphasis on silent parts. This is because it is easier and clearer to recognize the effects of recording and playback when there is no speech. In this regard, it can be said that paying attention to silence intervals and feature values corresponding to special frequency bands may lead to better detection of physical access attacks

5.4 Comparison with Other Single Models

The proposed models have been compared with some of the single models and the baseline models according to the presented objectives. Some of the top-performing models used for relevant purposes are shown and compared with the proposed model in Table 4. For the LA attack, the LFCC+SEResABNet+CombLoss model achieves an EER=1.89% and t-DCF=0.507, which outperforms the baseline model LFCC-GMM. The proposed model also outperforms its corresponding base model (LFCC+SEResNet50+CE) for approximately 0.98%. Also, by comparing the results obtained with other works, it can be seen that this model outperforms LFCC+ResNet18+OCS, which to the best of our knowledge, shows state-of-the-art

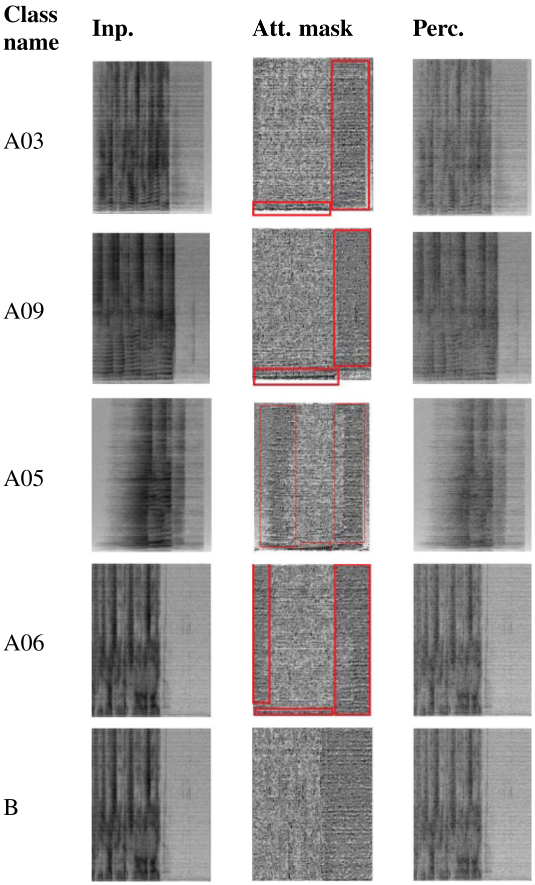

Fig. 7 Input feature (Inp.), produced attention mask (Att.), and final input feature for perception branch (Perc.) of some samples in the evaluation set for logPowSpec feature. B is bonafide class. Red boxes are parts of input features that the attention branch emphasizes and are interpretable from human's point of view

performance. For physical access attacks, the LogPowSpec+EABN+CombLoss model achieved EER=0.86% and t-DCF=0.0239. This result is significantly better than the base models. Compared to results reported in the 2019 challenge, the proposed model also appears to outperform 90% of methods which use fusion models. These results, and other favorable features such as fewer parameters and shorter runtime compared to other models, prove the efficiency of the proposed EABN model. Finally, we evaluated the best-proposed models obtained on the ASVspoof 2019 dataset on the 2021 version as shown in Table 5. The results show that these models perform better than all the base models. These results indicate that the model presents good performance on LA attacks.

Table 4 Performance comparison of the proposed systems with known single systems tested on the ASVspoof 2019 PA and LA evaluation set. Models are named base on their input feature, the classification model, and the loss function

Input feature + Classifier + Loss function	PA		LA	
	EER(%)	t-DCF	EER(%)	t-DCF
(Baseline) CQCC+GMM+EM [30]	11.04	0.2454	9.57	0.2366
(Baseline) LFCC+GMM+EM [30]	13.54	0.3017	8.09	0.2116
Spect+ResNet+CE [2]	3.81	0.9940	9.68	0.2741
MFCC+ResNet+CE [2]	–	–	9.33	0.2042
Spect+ResNet+CE [17]	1.29	0.0360	11.75	0.2160
Joint-gram+ResNet+CE [3]	1.23	0.0305	–	–
LFCC+LCNN+A-softmax [18]	4.60	0.1053	5.06	0.1000
Spect+LCNN+A-softmax [18]	–	–	4.53	0.1028
FG-CQT+LCNN+CE [31]	–	–	4.07	0.1020
Spect+LCGRNN+GKDE-softmax [9]	1.06	0.0222	3.77	0.0842
Spect+LCGRNN+triplet	0.92	0.0198	–	–
Fbank&CQT+ResNeWt+CE [4]	0.52	0.0134	–	–
CQTMGD+ResNeWt+CE [4]	0.94	0.0250	–	–
Spect+SE-Res2Net50+CE [20]	0.74	0.0207	8.73	0.2237
LFCC+SE-Res2Net50+CE [20]	1.46	0.434	2.87	0.0786
CQT+SE-Res2Net50+CE [20]	**0.46**	**0.0116**	2.50	0.0743
Raw signal+SincNet+CE [35]	–	–	20.11	0.3563
logCQT&powSpect+VGG+CE [35]	2.11	0.527	–	–
LFCC+ResNet18+OCS[36]	–	–	2.19	0.0590
Proposed: LFCC+SE-ResABNet+CombLoss	–	–	**1.89**	**0.0507**
Proposed: LogPowSpec+EABNet+CombLoss	0.86	0.0239	–	–

The bold values indicate the best results compared to the others

Table 5 Performance comparison of the proposed systems with ASVspoof 2021 baseline systems tested on the ASVspoof 2021 PA and LA evaluation set

Model	PA		LA	
	EER(%)	**t-DCF**	**EER(%)**	**t-DCF**
Baseline 01 [34]	38.07	0.9434	15.62	0.4974
Baseline 02 [34]	39.54	0.9724	19.30	0.5758
Baseline 03 [34]	44.77	0.998	9.26	0.3445
Baseline 04 [34]	48.60	0.9997	9.50	0.4257
Proposed: LFCC+SE-ResABNet+CombLoss	–	–	**5.62**	**0.2745**
Proposed: LogPowSpec+EABNet+CombLoss	**31.04**	**0.7812**	–	–

The bold values indicate the best results compared to the others

6 Conclusion

Spoof detection is considered as a major security concern in authentication systems, particularly the ASV system, demonstrating a clear need for solutions to combat spoof attacks. There are generally two approaches to detect spoofing attacks on ASV systems: the first is to develop an appropriate classifier targeted specifically at detecting attacks, while the second approach is conducted as a preliminary step for extracting discriminative features. In the case of the former, most classifiers fail to consider the issue of optimality in terms of number of parameters and runtime. On the other hand, most proposed models are not interpretable from a human point of view, and features are chosen according to expert's knowledge, and therefore lack generalization to unseen attacks. However, a modular architecture based on branches of attention and perception gives the system the ability to easily utilize any classifier or method to produce an interpretable attention mask and improve classification task. To this end, the proposed combined loss function, particularly the triplet center loss, succeeded in yielding a discriminative feature space that can help achieve a more generalized model for unseen attacks.

The proposed model and loss function were evaluated on ASVspoof 2019 data. Using LogPowSpec and LFCC features, along with the first-time use of the EfficientNet-A0 architecture and the efficient SE-Res2Net50, this study provides a novel method for detecting spoofs. The findings show that the LFCC+SEResNet50+CE model runs with an EER of 1.89% and t-DCF of 0.507 in the logical access scenario, which to the best of our knowledge, outperforms all state-of-the-art methods. The EABN+CombLoss also obtained an EER of 0.86% and t-DCF of 0.0239 for the physical access scenario, which is better than 90% of the models presented for the ASVspoof 2019 challenge. It is worth noting that the EfficientNet-A0 consists of only 95,000 parameters. The findings also shed light on certain special cases observed for the produced attention masks. For example, LFCC features outperformed MFCCs in detecting logical access attacks. Alternatively, to detect replay attacks, focusing more on silent segments and some frequency ranges in the human speech frequency range can improve the performance.

In this research, we were able to achieve the goals defined for a suitable countermeasure system. The first one was to provide a generalize system against unseen attacks. To achieve this goal, we proposed modular EABN architecture along with the combined loss function. In addition, providing a system that has a suitable (few) numbers of parameters and FLOPS is another main goal. We optimized EfficientNet-A0 and use it in the perception branch. This model has a few parameters and FLOPS as well as achieves comparable results. For future steps, considering that the proposed method has a modular architecture, other methods and models can be used in branches and their performance can be investigated. We can fuse branches in a multi branch network where each branch can use a specific architecture or a specific input feature.

Finally, it is possible to examine the effect of using several branches of perception which can be trained together or separately.

Data Availibility ASVspoof 2019 and 2021 datasets are available in the following link: https://www.asvspoof.org Our implementation of the proposed method is available in the following Github repository: https://github.com/AmirmohammadRostami/ASV-anti-spoofing-with-EABN

Declarations

Conflicts of Interest The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication

References

1. M.J. Alam, G. Bhattacharya, P. Kenny, Boosting the performance of spoofing detection systems on replay attacks using q-logarithm domain feature normalization. Odyssey **2018**, 393–398 (2018)
2. M. Alzantot, Z. Wang, M. B. Srivastava, Deep residual neural networks for audio spoofing detection. pp. 1078–1082 (2019) https://doi.org/10.21437/Interspeech.2019-3174
3. W. Cai, H. Wu, D. Cai, *et al.* The DKU replay detection system for the asvspoof 2019 challenge: on data augmentation, feature representation, classification, and fusion. pp. 1023–1027 (2019) https://doi.org/10.21437/Interspeech.2019-1230
4. X. Cheng, M. Xu, T. F. Zheng, Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019. in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 540–545. IEEE (2019)
5. B. Chettri, T. Kinnunen, E. Benetos, Deep generative variational autoencoding for replay spoof detection in automatic speaker verification. Comput. Speech Lang. **63**, 101092 (2020)
6. H. Delgado, M. Todisco, M. Sahidullah, *et al.*, ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. in *Odyssey 2018-The Speaker and Language Recognition Workshop* (2018)
7. S. K. Ergünay, E. Khoury, A. Lazaridis, *et al.* , On the vulnerability of speaker verification to realistic voice spoofing. in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6. IEEE (2015)
8. H. Fukui, T. Hirakawa, T. Yamashita, *et al.* Attention branch network: Learning of attention mechanism for visual explanation. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10705–10714 (2019)
9. A. Gomez-Alanis, J.A. Gonzalez-Lopez, A.M. Peinado, A kernel density estimation based loss function and its application to asv-spoofing detection. IEEE Access **8**, 108530–108543 (2020)
10. A. Hadid, N. Evans, S. Marcel et al., Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. IEEE Signal Process. Mag **32**(5), 20–30 (2015)
11. J.H. Hansen, T. Hasan, Speaker recognition by machines and humans: a tutorial review. IEEE Signal process. mag. **32**(6), 74–99 (2015)
12. L. Huang, C.-M. Pun, Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced denseNet-BiLSTM network. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1813–1825 (2020)
13. M.R. Kamble, H. Tak, H.A. Patil, Amplitude and frequency modulation-based features for detection of replay spoof speech. Speech Commun. **125**, 114–127 (2020)
14. T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. (2017)
15. P. Korshunov, S. Marcel, Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations. IEEE J. Sel. Top. Signal Process. **11**(4), 695–705 (2017)
16. C.-I. Lai, A. Abad, K. Richmond, *et al.*, Attentive filtering networks for audio replay attack detection. in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6316–6320. IEEE (2019)

17. C.-I. Lai, N. Chen, J. Villalba, *et al.* “ASSERT: Anti-spoofing with squeeze-excitation and residual networks.” arXiv preprint arXiv:1904.01120 (2019)
18. G. Lavrentyeva, S. Novoselov, A. Tseren, *et al.*, STC Antispoofing Systems for the ASVspoof2019 Challenge. pp. 1033–1037 (2019) https://doi.org/10.21437/Interspeech.2019-1768
19. D. Li, L. Wang, J. Dang, *et al.*, Multiple phase information combination for replay attacks detection. in *INTERSPEECH*, pp. 656–660 (2018)
20. X. Li, N. Li, C. Weng, *et al.*, Replay and synthetic speech detection with res2net architecture. in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6354–6358. IEEE (2021)
21. Q. Lu, Y. Li, Z. Qin, *et al.*, Speech Recognition Using EfficientNet. Association for Computing Machinery, New York, NY, USA (2020) https://doi.org/10.1145/3404716.3404717
22. D. Park, W. Chan, Y. Zhang, *et al.* SpecAugment: a simple data augmentation method for automatic speech recognition. pp. 2613–2617, 09 (2019)
23. Y. Qian, N. Chen, K. Yu, Deep features for automatic spoofing detection. Speech Commun. **85**, 43–52 (2016)
24. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process **3**(1), 72–83 (1995)
25. A. M. Rostami, A. Karimi, M. A. Akhaee, Keyword spotting in continuous speech using convolutional neural network. Speech Commun. (2022)
26. M. Sahidullah, H. Delgado, M. Todisco, *et al.*, Introduction to voice presentation attack detection and recent advances. in *Handbook of Biometric Anti-Spoofing*, pp. 321–361. Springer (2019)
27. G. Suthokumar, V. Sethu, C. Wijenayake, *et al.*, Modulation Dynamic Features for the Detection of Replay Attacks. in *Interspeech*, pp. 691–695 (2018)
28. M. Tan and Q. V. Le., EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.*CoRR*, volume abs/1905.11946 (2019) arXiv:http://arxiv.org/abs/1905.11946
29. M. Todisco, H. Delgado, N.W. Evans, A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. Odyssey **2016**, 283–290 (2016)
30. M. Todisco, X. Wang, V. Vestman, *et al.*, ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. in *Proc. Interspeech 2019*, pp. 1008–1012 (2019) https://doi.org/10.21437/Interspeech.2019-2249
31. Z. Wu, R. K. Das, J. Yang, *et al.* , Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. pp. 1101–1105 (2020) https://doi.org/10.21437/Interspeech.2020-1810
32. Z. Wu, T. Kinnunen, N. Evans, *et al.*, ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. in *Sixteenth Annual Conference of the International Speech Communication Association* (2015)
33. X. Xiao, X. Tian, S. Du, *et al.*, Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. in *Sixteenth Annual Conference of the International Speech Communication Association* (2015)
34. J. Yamagishi, X. Wang, M. Todisco, *et al.*, ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. pp. 47–54. 09 (2021)
35. H. Zeinali, T. Stafylakis, G. Athanasopoulou, *et al.*, Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge. pp. 1073–1077 (2019) https://doi.org/10.21437/Interspeech.2019-2892
36. Y. Zhang, F. Jiang, Z. Duan, One-class learning towards synthetic voice spoofing detection. IEEE Signal Process. Lett. **28**, 937–941 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.