Interspeech 2022
18-22 September 2022, Incheon, Korea

Multi-Path GMM-MobileNet Based on Attack Algorithms and Codecs for Synthetic Speech and Deepfake Detection

Yan Wen, Zhenchun Lei, Yingen Yang, Changhong Liu, Minglei Ma

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China

wenyan@jxnu.edu.cn, zhenchun.lei@hotmail.com, yyg1999@sina.com, liuch@jxnu.edu.cn, sljsmml@163.com

Abstract

The generalization ability of the speech spoofing detection system in real unseen sources is a great challenge. Spoofed speech from different attack algorithms or codecs has different feature distribution, which is the variant of the genuine speech. The conventional GMM describes the common distribution of all speech feature. But the GMM does not pay attention to the specificity of speech generated using an attack algorithm or codec, which may be useful to model the feature distribution of speech from unknown source. We propose the multi-path GMM-MobileNet model, which includes the GMMs trained on genuine and spoofed speech generated using various attack algorithms or codecs respectively. The 1-D variant of the MobileNet structure is used to extract embedding vector, and the multi-path structure is used to improve the generalization ability. On ASVspoof 2021 LA task, the M-GMM-MobileNet achieves a minimum t-DCF of 0.3231 and an EER of 6.80%, which relatively reduce by 6.2% and 26.6% compared with the LFCC-LCNN baseline. On the ASVspoof 2021 DF task, the M-GMM-MobileNet achieves an EER of 16.86%, which relatively reduce by 24.7% compared with the RawNet2 baseline. Compared with the systems on the ASVspoof 2021 DF leaderboard, our model is competitive.

Index Terms: anti-spoofing, speaker verification, GMM-MobileNet, multi-path architecture

1. Introduction

Automatic speaker verification (ASV) systems are widely used in our daily lives. At the same time, protecting the ASV systems from spoofing attacks is becoming increasingly important. There are four primary spoofing attacks, including impersonation, voice conversion (VC), text-to-speech synthesis (TTS) and replay. These attacks pose a significant threat to ASV systems. Therefore, the anti-spoofing research has attracted widespread attention in speaker verification field.

The automatic speaker verification spoofing and countermeasures (ASVspoof) challenges started in 2015 [1]. Main objective of this public initiative is to support the development of spoofing countermeasures for ASV systems [2]. The latest edition, ASVspoof 2021 [3], contains three tasks: logical access (LA), physical access (PA), and speech deepfake (DF). LA evaluation data includes a collection of genuine and spoofed utterances transmitted over a variety of telephony systems. DF dataset includes utterances processed with different lossy codecs used for media storage. Utterances in DF dataset are similar to LA data but not related to ASV. There is a large mismatch between the ASVspoof 2019 [4] training set and ASVspoof 2021 evaluation set. One of the objectives for the 2021 challenge is to improve the generalization of the spoofing countermeasures in more practical and realistic conditions.

The MobileNetV1 [5] is an efficient network architecture which introduces depthwise separable convolutions. The MobileNetV1 with depthwise separable convolutions is nearly as accurate as VGG16 [6] but has an effective reduction in computational cost on ImageNet [7]. The MobileNetV2 [8] is more efficient and smaller which introduced the linear bottleneck and inverted residual structure in order to make even more efficient layer structures. The MobileNetV3 [9] built upon MobileNetV1 and MobileNetV2, are upgraded with the squeeze-and-excitation layer and h-swish activation method. The MobileNetV3 is developed and applied to classification, detection and segmentation tasks to achieve high accuracy and low latency. Inspired by the effectiveness of the MobileNets across a wide range of applications, we applied the MobileNetV3 combined with Gaussian Mixture Model to the synthetic speech and deepfake detection.

It is a great challenge to build a spoofing detection system that is robust to the unknown conditions. The generalized system should perform well on detecting diverse and unknown attacks. Many strategies are founded to improve the generalization ability of attack detection systems, such as data augmentation [10, 11, 12], adversarial learning [13], and one-class learning [14]. Das [10] considered various known and unknown data augmentation methods to develop a robust spoofing countermeasure which is one of the top performing systems on ASVspoof 2021 challenge. Kang et al. [11] trained multiple systems on a training set augmented with various audio codecs to improve the generalization of the spoofing countermeasures. Zhang et al. [12] proposed several strategies to enhance the channel robustness of countermeasure systems by using the channel-shifted data. Wang et al. [13] proposed the dual-adversarial domain adaptation (DADA) framework to enable the fine-grained alignment of spoofed and genuine data separately by using two domain discriminators for improving spoofing detection performance. Zhang et al. [14] proposed a spoofing detection system which is robust to unknown attacks using one-class learning, which compact the genuine speech representation and inject an angular margin to separate the spoofing attacks in the embedding space.

In this work, we propose the GMM-MobileNet and multi-path GMM-MobileNet for spoofing deepfake detection. The multi-path structure is constructed according to the different attack algorithms and data augmentation methods. The contributions of this work include: 1) Firstly, we combine GMM and 1-D variant of MobileNetV3 for speech spoofing detection. 2) Secondly, we propose the multi-path architecture of GMM-MobileNet based on attack algorithms and codecs to improve the generalization ability.

10.21437/Interspeech.2022-10312

2. Multi-Path GMM-MobileNet

2.1. Log Gaussian probability feature

The GMM is a well-known classifier for speech spoofing detection. For the distributions of genuine and spoofed speeches are different in feature space, their score distributions on all Gaussian components are also different. Hence the score distribution information is useful for spoofing detection. In our previous works [15, 16], the GMM was used as the feature extractor. The GMM takes raw frame feature as input and outputs the log probability of each frame provided by each component, which is called Log Gaussian Probability (LGP) feature. For a D dimensional frame feature x (LFCC in our experiments), the element y_i of the LGP feature y is defined as:

$$\begin{aligned} y_i &= \log p_i(x) \\ &= -\frac{1}{2} x' \Sigma_i^{-1} x + x' \Sigma_i^{-1} \mu' + Const \end{aligned} \tag{1}$$

where $p_i(x)$ is the Gaussian density function of the i-th component in GMM, which is parameterized by a $D \times 1$ mean vector, μ_i and a $D \times D$ covariance matrix, Σ_i. The constant term $Const$ can be removed to reduce computational cost. After that, the mean and standard deviation of the features in training data are computed and used for mean and variance normalization for each feature of speech.

2.2. GMM-MobileNet

The MobileNetV3 is a popular light-weight model which applies h-swish and squeeze-and-excitation [17] method. The variant of MobileNets has been applied to many tasks such as image classification or objective detection. We propose the 1-D variant of MobileNetV3 with the LGP feature for speech anti-spoofing tasks. For an utterance, the LGP feature is a matrix, in which the horizontal axis is related to time while the vertical axis is related to the Gaussian components in GMM. There is no apparent dependence between any two Gaussian components, so the vertical axis can be regarded as channel. The convolutional layer in our model should be 1-D convolution along the time axis, and the simplified MobileNetV3 structure is used in GMM-MobileNet.

Figure 1 shows the architecture of the proposed GMM-MobileNet model. C denotes the number of channels and L refers to the frame number of the utterance. The variant of MobileNetV3 consists of one full convolutional block, six depthwise and pointwise convolution based blocks, an adaptive average pooling layer, and two fully connected layers.

The first convolutional block in the proposed MobileNet consists of a 1-D convolutional layer with kernel of size 3 and stride of size 1, followed by a batchnorm and ReLU nonlinearity. Each building block in this structure is a combination of depthwise convolution, squeeze-and-excitation block and pointwise convolution. The depthwise convolution layer is a 1-D full convolutional layer with kernel of size 3 and stride of size 1, followed by a batchnorm and ReLU nonlinearity. The 1×1 pointwise convolution layer is followed by a batchnorm and has no nonlinearity. As the SE block, the size of the squeeze-and-excite bottleneck was fixed to be 1/4 of the number of input channels. The number of input channels and output channels of all the convolution layers are fixed to be C. Since the input and output have the same number of channels, they are connected with a residual connection. The convolution output is squeezed to channel size by the adaptive average pooling. The last layer with softmax activation function is used to discriminate between genuine and spoofed speech.

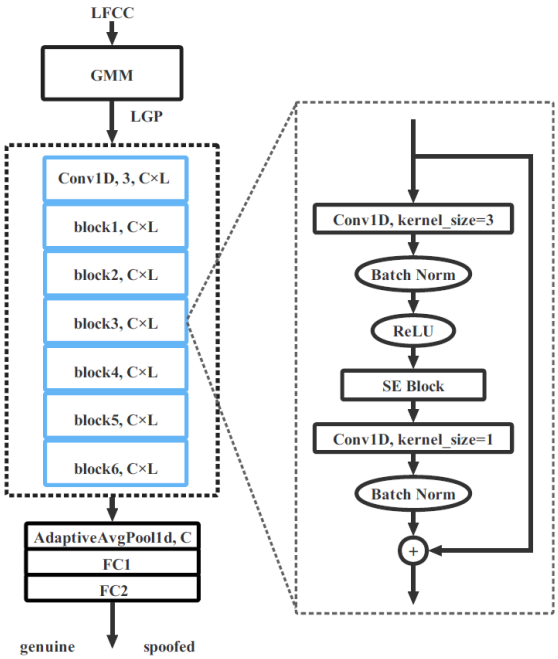

Figure 1: *The architecture of GMM-MobileNet.* ***LGP****:The log Gaussian Probability Feature.* $C \times L$*: The shape of the convolution output.* ***C****: The number of input channels of the convolutional layers which is set to be 128 or 256.* ***L****: The length of the input feature which is fixed to 400.*

The feature segmentation [18, 19] is used to create a unified feature map during the evaluation stage. It breaks a variable-length feature X of T frames into fixed-length segments of L frames by a sliding window. The sliding window is shifted by S-frames interval at each step. For the case $L \geq T$, the feature X is padded to make the length L. In the evaluation, the score of an utterance is computed by averaging the scores of all the segments from the utterance. In our experiments, L is set to 400 and S is set to 200.

2.3. Multi-path GMM-MobileNet based on attacks

Improving the generalization ability of the spoofing countermeasure system in unknown condition is a great challenge. The spoofed speech from different sources has different feature distribution, which is a variant of genuine speech. The conventional GMM is usually used to model the feature distribution of all speeches, which describes the common distribution of genuine and spoofed speech. But the GMM does not pay attention to the specificity of speech features generated using an attack algorithm or codec, which may be useful to model the feature distribution of speech from unknown source. Therefore, we propose a multi-path architecture with MobileNet (M-GMM-MobileNet) for spoofing detection, and each path is constructed according to one attack algorithm.

Figure 2 shows the M-GMM-MobileNet which contains seven GMM-MobileNets with same architecture. Each spoofed utterance in ASVspoof 2019 training set is generated using one of six attack algorithms, namely A01-A06, respectively. The LGP features are extracted by seven GMMs, which are sep-

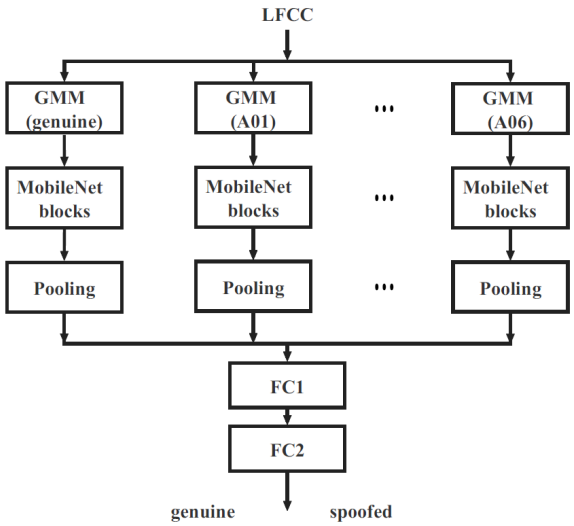

Figure 2: *The M-GMM-MobileNet model based on attacks. Seven GMMs trained on genuine utterances and spoofed utterances generated using six attack algorithms in training set.*

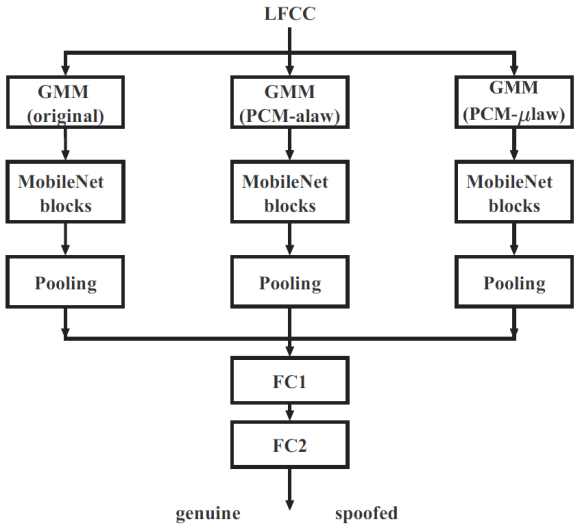

Figure 3: *The M-GMM-MobileNet model based on codecs. Three GMMs are trained on original training set and two augmented sets by different codecs.*

arately trained on genuine utterances and spoofed utterances generated using six algorithms in training set. The embedding vectors from all paths are concatenated and fed into the fully connected layers for spoofing detection.

2.4. Multi-path GMM-MobileNet based on codecs

In the ASVspoof 2021 LA task, the genuine and spoofed utterances generated using text-to-speech and voice conversion algorithms are communicated across telephony and VoIP networks with various coding and transmission effects. Previous works [10, 20] have shown that data augmentation could make effective progress on attack detection. Like the attack algorithms in previous subsection, the codec can also be used to construct the multi-path GMM-MobileNet. And the proposed model emphasizes the feature distribution of speech generated with different codecs.

Figure 3 shows the multi-path GMM-MobileNet based on codecs. Three GMMs are trained on original training set and two augmented sets generated with PCM-alaw and PCM-μlaw codecs separately. The size of each augmented set is identical to that of original training set. The MobileNet following each GMM takes LGP feature as input and outputs the embedding vector. The embedding vectors are concatenated and fed into the fully connected layer for spoofing detection.

3. Experiments

3.1. Experimental setup

The proposed models are evaluated on the ASVspoof 2021 challenge [3] LA and DF tasks. All models are trained only using the training data in ASVspoof 2019 LA task according to the requirement of the challenge. There are progress phase and evaluation phase for all the tasks in ASVspoof 2021. The progress phase is planned to observe the trend of the scores to the CodaLab submission platform and the evaluation phase determines the final results. The evaluation sets of LA and DF tasks of ASVspoof 2021 include 181566 and 611829 utterances respectively. The ASVspoof 2021 challenge organizers provide 4 baseline systems: LFCC-GMM [21], CQCC-GMM [22], LFCC-LCNN [23], and RawNet2 [24]. Performance is measured in terms of minimum tandem detection cost function (t-DCF) [25] and equal error rate (EER).

The LFCC is used as acoustic feature in all experiments, and the feature extractor is implementation of spoofing detection baseline system provided by the organizers with the default configuration. The extracted LFCC feature are turned to the fixed length of 400. The LGP features of size 128×400 or 256×400 are extracted by 128-component GMM or 256-component GMM respectively.

Our models are implemented using PyTorch framework and tested on workstation with a GeForce GTX 2080 TI graphics card. Cross-entropy loss is adopted as the loss criterion and Adam optimizer with learning rate of 0.0005 is used during the training process. The learning rate is adjusted by StepLR with step size of 15 and factor of 0.1. The weight decay of 0.0001 is used to prevent overfitting. The batch size is set to 64 in all experiments, and each model is trained for 100 epochs.

3.2. Results on ASVspoof 2021 LA task

Table 1 shows the evaluation results of four baseline systems, GMM-MobileNet, and M-GMM-MobileNet in ASVspoof 2021 LA progress phase and evaluation phase. Compared with the classical GMM systems, the GMM-MobileNet gets better performance. Moreover, we trained the single-path systems on both original training data and augmented data. The performance of GMM(128)-MobileNet and GMM(256)-MobileNet systems are improved significantly when trained on the augmented training set. It is also proved that data augmentation is effective and helpful to improve system's robustness.

As shown in table 1, all multi-path GMM-MobileNet systems perform better than baseline systems in both progress and evaluation phase. Compared with GMM(128)-MobileNet, M-GMM(128)-MobileNet(A) reduces t-DCF by 3.6% and EER by 10.4% without data augmentation. M-GMM(128)-

Table 1: *Performance of GMM-MobileNet and multi-path GMM-MobileNet models evaluated on ASVspoof 2021 logical access (LA) task in terms of min-tDCF and EER (%).* ***DA****: Data augmentation.* ***GMM****: The number of the components in GMM.* ***A****: Attacks based model.* ***C****: Codecs based model.*

Model	DA	GMM	Progress		Evaluation	
			t-DCF	EER	t-DCF	EER
CQCC-GMM [3]	-	-	0.4948	15.80	0.4974	15.62
LFCC-GMM [3]	-	-	0.5836	21.13	0.5758	19.30
LFCC-LCNN [3]	-	-	0.3152	8.90	0.3445	9.26
RawNet2 [3]	-	-	0.4152	9.49	0.4257	9.50
GMM-MobileNet	-	128	0.3177	8.12	0.3547	8.75
GMM-MobileNet	✓	128	0.3047	7.46	0.3477	7.97
GMM-MobileNet	-	256	0.3146	8.12	0.3496	8.94
GMM-MobileNet	✓	256	0.2826	5.84	0.3301	7.00
M-GMM-MobileNet(A)	-	128	0.3066	7.22	0.3421	7.84
M-GMM-MobileNet(A)	-	256	0.3045	6.56	0.3436	7.74
M-GMM-MobileNet(C)	✓	128	0.2874	6.26	0.3257	7.13
M-GMM-MobileNet(C)	✓	256	0.2883	6.32	**0.3231**	**6.80**
ResNet-L-LDE [26]	✓	-	0.2377	3.10	0.2720	3.68
CQT-LCNN(L7) [10]	✓	-	0.3001	5.67	0.3197	5.27
light TDNN Focal [20]	✓	-	0.3263	7.33	0.3645	7.51
MFM-thin-ASSERT34 [27]	-	-	0.349	12.41	0.676	17.35
MFM-ASSERT18 [27]	-	-	0.332	11.87	0.674	17.41

Table 2: *Performance of GMM-MobileNet and multi-path GMM-MobileNet models evaluated on ASVspoof 2021 deepfake (DF) task in terms of min-tDCF and EER (%).* ***DA****: Data augmentation.* ***GMM****: The number of the components in GMM.* ***A****: Attacks based model.* ***C****: Codecs based model.*

Model	DA	GMM	EER	
			Progress	Evaluation
CQCC-GMM [3]	-	-	17.63	25.56
LFCC-GMM [3]	-	-	21.01	25.25
LFCC-LCNN [3]	-	-	11.61	23.48
RawNet2 [3]	-	-	6.10	22.38
GMM-MobileNet	-	128	9.97	20.08
GMM-MobileNet	✓	128	7.94	18.39
GMM-MobileNet	-	256	11.03	20.11
GMM-MobileNet	✓	256	5.67	17.82
M-GMM-MobileNet(A)	-	128	11.08	20.12
M-GMM-MobileNet(A)	-	256	11.92	19.79
M-GMM-MobileNet(C)	✓	128	4.75	17.97
M-GMM-MobileNet(C)	✓	256	5.20	**16.86**
ResNet-L-FM [26]	✓	-	1.79	16.36
CQT-LCNN(D3) [10]	✓	-	3.87	18.31

MobileNet(C) and M-GMM(256)-MobileNet(C) are trained on enlarged training set with codec augmentation. They also obtain performance improvement compared with GMM(128)-MobileNet and GMM(256)-MobileNet.

The best result on ASVspoof 2021 LA task is demonstrated by M-GMM(256)-MobileNet(C). In the evaluation phase, M-GMM(256)-MobileNet(C) achieves a minimum t-DCF of 0.3231 and an EER of 6.80% and relatively reduces the t-DCF and EER by 6.2% and 26.6% compared with the LFCC-LCNN baseline system. It is observed that with lager number of GMM components, GMM-MobileNet systems can obtain better performance. This may benefit from the more information contained in the Gaussian probability feature.

Compared with other single state-of-the-art system, the M-GMM-MobileNet has better performance when the data augmentation method is not used. But the performance of the M-GMM-MobileNet is worse than that of other state-of-the-art systems which use the data augmentation. The reason may be that there are not enough data augmentation methods in our experiments. We will construct model structure with more paths based on more augmentation method, such as reverberation effect, background noise, and audio compression effect.

3.3. Results on ASVspoof 2021 DF task

All the GMM-MobileNet-based models outperform the baseline systems in DF task, as shown in table 2. The M-GMM-MobileNet(C) models obtain significant performance improvement compared with the single-path GMM-MobileNet models. The performance of M-GMM-MobileNet based on codecs is better than that of M-GMM-MobileNet based on attack algorithms. The M-GMM(256)-MobileNet(C) achieves an EER of 16.86%, which relatively reduce by 24.7% compared with the RawNet2 baseline system.

Our model obtains very competitive results compared with other state-of-the-art systems.The performance of M-GMM(256)-MobileNet(C) model is close to that of the single system ResNet-L-FM in [26]. The fusion result in [26] is an EER of 16.05% which ranked second on the DF task leaderboard. The top-performing single system of the third team [10] achieves an EER of 18.31%. Notably, compared with the top-performing systems [10, 26], the M-GMM(256)-MobileNet(C) model obtains a very competitive performance. There are only two data augmentation methods used in our experiments, and our model has the potential to obtain better performance by adding augmentation methods. On the other hand, we also observe that the M-GMM-MobileNet(C) model can get stable performance in both progress and evaluation phases.

4. Conclusions

In this paper, we proposed the GMM-MobileNet model which includes a GMM, six blocks consist of depthwise convolution layer, SE block and pointwise convolution layer for speech spoofing detection. Experiments on ASVspoof 2021 LA and DF tasks show the effectiveness of the MobileNets for detecting deepfake and synthesis speech attacks. The M-GMM-MobileNet model has also been proposed, which has multi-path GMM-MobileNet based on attack algorithms and codecs. The multi-path structure contributes a lot to improve the generalization ability. The M-GMM-MobileNet systems outperform the baseline systems on both ASVspoof 2021 LA and DF tasks, and are competitive compared with other state-of-the-art systems. The future work will be further improving the system's performance on logical access task and conducting experiments on replay attack detection. More data augmentation methods will also be added to the multi-path structure to improve the generalization ability.

5. Acknowledgements

This work is supported by National Natural Science Foundation of P.R.China (62067004).

6. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech 2015*, 2015, pp. 2037–2041.

[2] H. Delgado, N. Evans, T. Kinnunen, K. Aik Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," 2021.

[3] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[4] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, p. arXiv:1409.1556, Sep. 2014.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115(3), pp. 211–252, 2015.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[9] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[10] R. K. Das, "Known-unknown Data Augmentation Strategies for Detection of Logical Access, Physical Access and Speech Deepfake Attacks: ASVspoof 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 29–36.

[11] W. H. Kang, J. Alam, and A. Fathan, "CRIM's System Description for the ASVSpoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 100–106.

[12] Y. Zhang, G. Zhu, F. Jiang, and Z. Duan, "An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems," in *Proc. Interspeech 2021*, 2021, pp. 4309–4313.

[13] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Dual-Adversarial Domain Adaptation for Generalized Replay Attack Detection," in *Proc. Interspeech 2020*, 2020, pp. 1086–1090.

[14] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[15] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection," in *Proc. Interspeech 2020*, 2020, pp. 1116–1120.

[16] Z. Lei, H. Yan, C. Liu, M. Ma, and Y. Yang, "Two-path gmm-resnet and gmm-senet for asv spoofing detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6377–6381.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[18] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6316–6320.

[19] Q. Wang, K. A. Lee, and T. Koshinaka, "Using Multi-Resolution Feature Maps with Convolutional Neural Networks for Anti-Spoofing in ASV," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 138–142.

[20] J. Cáceres, R. Font, T. Grau, and J. Molina, "The Biometric Vox System for the ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 68–74.

[21] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech 2015*, 2015, pp. 2087–2091.

[22] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[23] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Proc. Interspeech 2021*, 2021, pp. 4259–4263.

[24] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[25] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[26] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 89–93.

[27] Z. Benhafid, S. A. Selouani, M. S. Yakoub, and A. Amrouche, "LARIHS ASSERT Reassessment for Logical Access ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 94–99.