INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Group GMM-ResNet for Detection of Synthetic Speech Attacks

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

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China
zhenchun.lei@hotmail.com, wenyan@jxnu.edu.cn, yyg1999@sina.com, liuch@jxnu.edu.cn, sljsmml@163.com

Abstract

The CNN-based models have achieved a remarkable success for speaker recognition and spoofing speech detection. We propose the group GMM-ResNet for synthesis speech detection. The grouping technique is used to improve classification accuracy by exposing the group cardinality while reducing both the number of parameters and the training time. The grouping technique allows the model to jointly attend to information from different representation subspaces. We propose two grouping methods, which are based on the Gaussian components in GMM. And the GMM is trained using binary splitting method. On the ASVspoof 2021 LA task, the group GMM-ResNet achieves a minimum t-DCF of 0.2450 and an EER of 2.53%, which relatively reduces by 28.9% and 72.7% compared with the LFCC-LCNN baseline. On the ASVspoof 2021 DF task, the group GMM-ResNet achieves an EER of 15.96%, which relatively reduces by 28.7% compared with the RawNet2 baseline.

Index Terms: Group GMM-ResNet, synthetic speech detection, anti-spoofing

1. Introduction

Speech synthesis and voice conversion (VC) techniques can be used to cheat the automatic speaker verification (ASV) system. In recent literatures [1], the state-of-the-art Text to Speech (TTS) and VC systems achieve a high level of naturalness that are comparable with real human speech. Even humans can hardly distinguish between synthetic speech and real speech. Therefore, the synthetic speech detection is an important task for ASV. The task of synthetic speech detection is to design anti-spoofing system which distinguishes the spoofed speech from the real ones.

In recent years, the deep neural networks (DNNs) have also shown great success in speech classification or recognition tasks, such as speech recognition [2], speaker recognition [3, 4], speech emotion recognition [5], speech anti-spoofing [6, 7]. The deep learning methods based on convolutional neural network (CNN) are widely used in speech anti-spoofing. Light convolution neural network (LCNN) [6] and residual convolutional neural network (ResNet) [8] are generally used to learn deep speech representation. LCNN with max feature map (MFM) activation function achieves the best performance in the ASVspoof 2017 Challenge [9]. Chen et al. [10] trained ResNet18 based systems for spoofing detection and achieved very competitive results on the ASVspoof 2021 LA task. ECAPA-TDNN [4] introduces the 1-dimensional Res2Net modules with impactful skip connections and Squeeze-and-Excitation blocks to explicitly model channel interdependencies. RawNet2 [7] is an end-to-end network architecture which directly takes raw speech waveform as inputs and uses six residual blocks in the embedding extractor. AASIST [11] uses the RawNet2-based encoder for extracting high-level feature maps from raw input waveforms and proposes a variant of the graph attention layer. Tomilov et al. [12] proposed a weighted score-level ensemble system which contains LCNN9, ResNet18, and RawNet2.

Group convolution refers to dividing the input channels into distinct groups and performing a regular convolution over each group separately. It is first proposed in AlexNet [13] for distributed computing of convolutions in CNN over multiple GPUs. It was shown that the group convolution is very effective on reducing both the number of parameters and the training time of CNN, and could also be used to improve classification accuracy. Specifically, we can increase the accuracy by exposing a new higher dimension through grouped convolution cardinality (the size of set of transformations). Tianyan Zhou et al. [14] investigated the effectiveness of ResNeXt for speaker verification, and the second convolutional layer in the ResNeXt block is a multi-branch transformation with different cardinalities. In the Transformer [15] model, the multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions, and this is also a grouping technique. On the other hand, the grouping technique is also considered as an ensemble learning method that implements the idea of training sub-models on feature space subsets.

In our previous works [16, 17], the ResNet-based models were constructed with the GMM for spoofing speech detection. In this paper, we proposed two grouping methods, which are based on the Gaussian components trained using binary splitting method. The paper is organized as follows: Section 2 explain the architecture of group GMM-ResNet model. The experiments are described in section 3. Finally, the conclusion is given in section 4.

2. Group GMM-ResNet Model

The architecture of the proposed model is shown in Figure 1. The Linear Frequency Cepstrum Coefficients (LFCC) feature is used as the input of the GMM, and the GMM extract the Log Gaussian Probability (LGP) feature. Then, the LGP feature is divided into G groups, and the ResNets followed by a Adaptive-MaxPooling module extract the sub-embeddings respectively. After that, we concatenate all sub-embeddings into a vector, and the fully connected layer is used for spoofing speech detection.

2.1. Log Gaussian probability feature

In the previous works [16, 17], the ResNet-based models with the LGP feature achieve the state-of-art performance for spoofing speech detection. The GMM takes raw feature as input and outputs the log probability feature provided by each Gaussian

10.21437/Interspeech.2023-1249

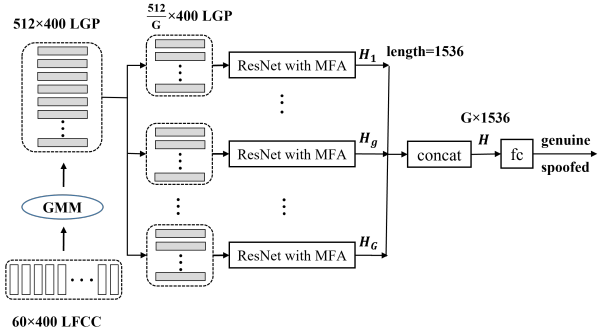


Figure 1: *The architecture of group GMM-ResNet.* ***G****: the group number, and it is set to 1, 2, 4, 8, or 16.*

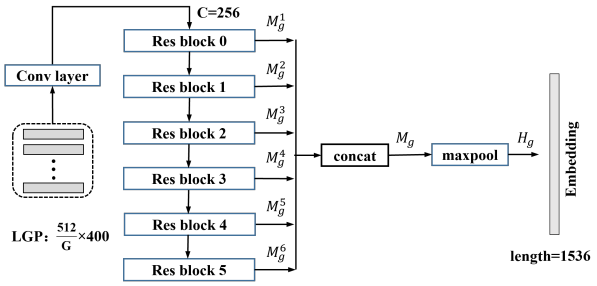


Figure 2: *The architecture of the g-th ResNet with MFA.* ***C****: the channel number.* ***G****: the group number.*

component. For a raw feature x (LFCC in our experiments), the element y_i of the LGP feature y is defined as:

$$y_i = -\frac{1}{2}x'\Sigma_i^{-1}x + x'\Sigma_i^{-1}\mu_i \tag{1}$$

where μ_i and Σ_i are the mean vector and covariance matrix of the i-th component in GMM. After that, the mean and variance normalization is used.

2.2. GMM-ResNet with multi-scale feature aggregation

The architecture of GMM-ResNet used in our system is shown in Figure 2, and the parameters are shown in Table 1. The number of GMM-ResNet modules is equal to the group number. The grouped LGP features are fed into the 1-d ResNet modules to extract discriminative embeddings. The ResNet is composed of 6 residual blocks which has 2 convolutional layers and skip residual connection. The Squeeze-and-Excitation is not used because it achieves worse performance in our experiments. After that, the embedding is extracted by applying the adaptive max-pooling operation.

In previous works [4, 7, 18, 19, 20] for speaker recognition and anti-spoofing, the low-level feature maps can also contribute towards the accurate embedding extraction. So, we apply the multi-scale feature aggregation (MFA) method to the GMM-ResNet model for performance improvement. We concatenate the output feature maps from all ResNet blocks before the max pooling layer.

$$M_g = Concat(M_g^1, M_g^2, ..., M_g^B) \tag{2}$$

$$H_g = AdaptiveMaxPooling(M_g) \tag{3}$$

where B denotes the number of ResNet blocks, M_g^b is the feature map extracted from the b-th residual block of the g-th group module, and H_g is the g-th sub-embedding. The max-pooling operation is applied to the vectors across temporal dimension. The sub-embeddings extracted by all GMM-ResNet modules are concatenated to obtain the final representation vector H:

$$H = Concat(H_1, H_2, ..., H_G) \tag{4}$$

where G denotes the group number. The length of embedding vector H is $C \times B \times G$, where C refers to the channel number in ResNet block. Finally, the embedding vector is fed into a fully connected layer with a softmax activation function to compute probabilites for genuine and spoofing speech classification.

Table 1: *ResNet based embedding extractor. Numbers denoted in Conv1d refer to kernel size, stride, and number of filters.*

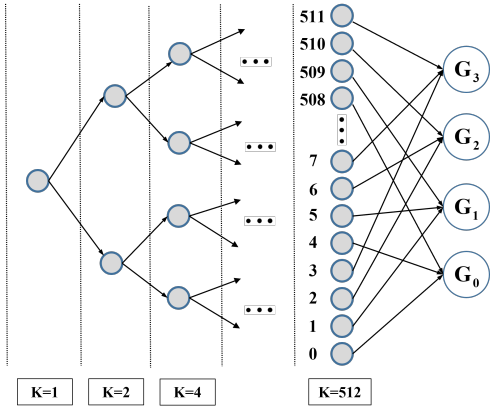
Layer	Input: LGP	Output shape
Conv layer	Conv1d(1,1,256) BN ReLU	(256, 400)
Res blocks	$\begin{pmatrix} Conv1d(3,1,256) \\ BN \\ ReLU \\ Conv1d(3,1,256) \\ BN \\ ReLU \end{pmatrix} \times 6$	(256, 400)
Max pool	AdaptiveMaxPool1d	(1536)

2.3. Grouping method

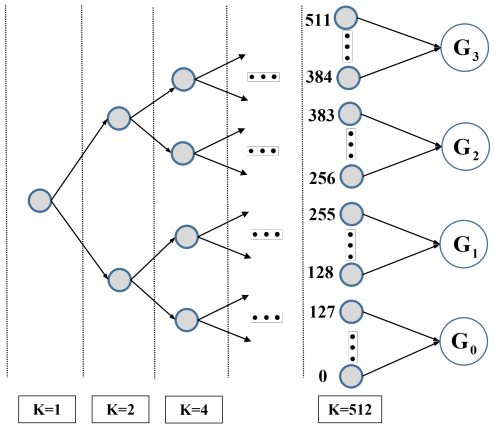
Grouping technology is widely used in CNN and Transformer models for its better performance or computational efficiency. We also group the LGP feature to achieve better performance while keeping the model size manageable. The LGP feature is based on the weighted distance to all centers of GMM components, so we can also group the Gaussian components to construct the feature subspaces. A simple method is random grouping, which randomly assigns a GMM component to a group. But this method does not consider the relationship between Gaussian components, and considers that all components are independent of each other.

We proposed two new grouping methods according to the GMM training procedure. In our experiments, the GMM is trained using binary splitting and expectation-maximization (EM). The binary splitting procedure is used to boot up the GMM from a single component to K components. After each split the GMM is re-estimated several times using the EM algorithm. The global mean and variance serves as the initial parameters of the 1-component GMM. The parameters are updated via the classical EM algorithm with maximum likelihood criterion.

The proposed grouping method is shown in Figure 3, in which all Gaussian components are divided into four groups. In Figure 3 (a), the components in each group come from different splitting branches. In this way, a feature subspace is represented by a split part of the Gaussian components. Since the lower components split by the same upper component have approximate parameters, we think there is a relationship between them, and they also can be input into different sub-modules. Therefore, we split the LGP feature into several groups of the same size along the channel axis. The size of grouped LGP features is $\frac{K}{G} \times L$, where K, G and L refer to the number of components,



(a) The grouped components come from different previous branches.



(b) The grouped components come from the same previous branch.

Figure 3: *Schematic diagram of grouping methods, in which 512 Gaussian components are divided into four groups, and each group contains 128 components.* ***K****: the number of Gaussian components.* G_i*: the i-th Gaussian component group.*

the number of groups and the time length respectively. Another grouping method is shown in Figure 3 (b). The Gaussian components come from the same previous branch are assigned to the same group.

After grouping, the LGP features are input into multiple ResNet (with MFA) modules respectively.

2.4. Data augmentation

The data augmentation methods are effective to improve the robustness of the spoofing countermeasures [10, 12, 20, 21, 22]. To avoid overfitting, the RawBoost proposed in [23] is used to enhance the variation of the training data. New speeches are generated using linear and non-linear (LnL) convolution noise, impulsive signal-dependent (ISD) additive noise, and stationary signal-independent (SSI) additive noise.

In addition, we employ two of the popular augmentation methods in speech processing for comparison, additive noise and room impulse response (RIR) simulation [24]. We use the audio clips from the MUSAN corpus [25]as the additive noise.

Table 2: *Performance of the group GMM-ResNet models on the ASVspoof 2021 LA task in terms of minimum t-DCF and EER (%).* ***DA****: Data augmentation.* ***a****: Grouping method a.* ***b****: Grouping method b.* ***r****: Random grouping method.*

Model	DA	t-DCF	EER(%)
LFCC-LCNN[26]	-	0.3445	9.26
RawNet2[26]	-	0.4257	9.50
GMM-ResNet	-	0.3621	7.83
Group GMM-ResNet(a)	-	0.3465	7.49
RawNet2[23]	Rawboost	0.3099	5.31
GMM-ResNet	Rawboost	0.2480	2.77
Group GMM-ResNet(a)	Rawboost	**0.2450**	**2.53**
Group GMM-ResNet(b)	Rawboost	0.2452	2.61
Group GMM-ResNet(r)	Rawboost	0.2479	2.66
Group GMM-ResNet(a)	RIR+noise	0.2686	4.06

3. Experiments

3.1. Experiment settings

The proposed models are evaluated on the ASVspoof 2021 [26] logical access (LA) and DeepFake (DF) tasks . According to the evaluation plan, all models are trained using ASVspoof 2019 [27] LA training data, which include 25380 utterances. The evaluation sets of ASVspoof 2021 LA and DF tasks include 181566 and 611829 utterances, respectively. The primary evaluate metric is minimum tandem detection cost function (t-DCF) [28] and the second is the equal error rate (EER).

The LFCC is used as acoustic feature in all experiments. The LFCC is extracted following the ASVspoof 2021 baseline configuration [26], using a 20 ms window with a 10 ms shift, a 1024-point Fourier transform, and comprising 19 static cepstra plus energy, delta and delta-delta coefficients. The extracted LFCC feature are turned to the fixed length of 400 by truncating or repeating. We train the GMM with 512 components and 30 EM iterations using the MSR Identity Toolbox [27] implementation on the ASVspoof 2019 training dataset. The size of LGP feature of each utterance is 512×400, which is input into the neural networks.

The proposed models are implemented using PyTorch framework and trained on GeForce RTX 3090 GPU. The Cross-entropy loss is adopted as the loss criterion, and the Adam optimizer with learning rate of 0.0001 is used during the training phase. The learning rate is adjusted by the ReduceLROnPlateau scheduler. No weight decay is used. The batch size is set to 32, and each model is trained for 100 epochs. The two-step training strategy [16] is also used in all experiments. Our source codes are publicly available on https://github.com/leizhenchun/interspeech2023, and all results in this paper are reproducible.

3.2. Results on ASVspoof 2021 LA task

The LA task contains bona fide speech and spoofed speech data generated by different TTS and VC systems with various coding and transmission effects. Table 2 shows the results on the ASVspoof 2021 LA task.

The proposed group GMM-ResNet model outperform the baseline system obviously. Compared with the LFCC-LCNN baseline, the group GMM-Resnet model can relatively reduce minimum t-DCF and EER by 28.9% and 72.7% on the evaluation dataset when using Rawboost method. The grouping

Table 3: *Performance of the group GMM-ResNet models with different parameters on the ASVspoof 2021 LA task in terms of minimum t-DCF and EER (%).* ***G****: the group number.* ***C****: the channel number in ResNet block.*

Parameter	t-DCF		EER(%)	
	best	avg	best	avg
G=1, C=512	0.2480	0.2526	2.77	2.90
G=2, C=256	0.2490	0.2505	2.76	2.83
G=4, C=256	0.2448	0.2496	2.56	2.78
G=8, C=256	0.2459	0.2481	2.59	2.67
G=16, C=256	**0.2450**	0.2470	**2.53**	2.59

Table 4: *Performance comparison between the group GMM-ResNet and other models on the ASVspoof 2021 LA task.*

Model	t-DCF	EER(%)
CQCC-GMM(baseline)[26]	0.4974	15.62
LFCC-GMM(baseline)[26]	0.5758	19.30
LFCC-LCNN(baseline)[26]	0.3445	9.26
RawNet2(baseline)[26]	0.4257	9.50
ECAPA-TDNN [29]	0.3094	5.46
RawNet2+RawBoost [23]	0.3099	5.31
GMM+LCNN [21]	0.2672	3.62
ResNet [10]	0.2608	3.21
T23 [26]	0.2176	1.32
Group GMM-ResNet	**0.2450**	**2.53**
XLS-128 [20]	-	3.54
wav2vec 2.0 + AASIST [22]	0.2066	0.82

method in which the Gaussian components come from different previous branches is better than the method in which the components come from the same previous branch, and they are better than the random grouping method. The data augmentation method can significantly improve the system performance. The Rawboost augmentation method is better than the RIR simulation and additive noise method, and it reduces EER from 7.49% to 2.53%.

Table 3 shows the performance of the group GMM-ResNet model with different group numbers. When the group number is 1, the group GMM-ResNet is the same as the GMM-ResNet model. The data augmentation method is the combination of LnL, ISD and SSI in series, and the results are obtained from five experiments for each configuration. We can see that the mean values of minimum t-DCF and EER on ASVspoof 2021 LA task continue to decrease with the increase of group number. The experiment with $C > 16$ is not run because computational cost is too high.

Table 4 compares the group GMM-ResNet with the state-of-the-art methods and four baseline systems (LFCC-GMM, CQCC-GMM, LFCC-LCNN, and RawNet2) provided by the ASVspoof 2021 challenge organizers on LA task. We can see that T23[26] which combines LCNN, ResNet, and RawNet2 using Mel-spectrogram, achieves the best performance on LA task with no externel training data. When considering more relaxed data policy, the best performance is 0.82% using wav2vec 2.0 front-end. The wav2vec 2.0 front-end is a wav2vec 2.0 XLS-R model which has 300M parameter and pretrained on a total of 436K hours of publicly available data.

Table 5: *Performance comparison between the group GMM-ResNet and other models on the ASVspoof 2021 DF task.*

Model	EER(%)
CQCC-GMM(baseline)[26]	25.56
LFCC-GMM(baseline)[26]	25.25
LFCC-LCNN(baseline)[26]	23.48
RawNet2(baseline)[26]	22.38
ECAPA-TDNN [29]	20.33
T06 [26]	19.01
GMM+LCNN [21]	18.30
M-GMM-MobileNet [17]	16.86
ResNet [10]	16.05
T23 [26]	15.64
Group GMM-ResNet	**15.96**
XLS-128 [20]	4.98
wav2vec 2.0 + AASIST [22]	2.85

3.3. Results on ASVspoof 2021 DF task

Evaluation data for the ASVspoof 2021 DF task is a collection of bona fide and spoofed speech utterances processed with different lossy codecs used typically for media storage. Different from the previous model structure, the GMM-ResNet only contain 3 residual blocks for better performance. Table 5 compares the proposed group GMM-ResNet with the state-of-the-art methods and four baseline systems. The group GMM-ResNet achieves an EER of 15.96%, which relatively reduce by 28.7% compared with the RawNet2 baseline system. Moreover, the performance of group GMM-ResNet is second only to T23 system on the leaderboard of the ASVspoof 2021 DF task.

The proposed model obtains very competitive results compared with other state-of-the-art systems that are compliant with the ASVspoof training protocol. But, the use of self-supervised learning (SSL) front-ends, such as wav2vec [20, 22], can substantially improve the performance when considering more relaxed data policy.

4. Conclusions

We propose the group GMM-ResNet architecture for synthetic speech detection in this paper. Two grouping mehods are proposed, and they are based on the relationship between each Gaussian component trained using binary splitting. In each group, the speech sub-embedding is extracted using ResNet with MFA. All sub-embeddings are concatenated and inputted into the fully connected layer for spoofing speech detection. The proposed group GMM-ResNet shows competitive performance on the ASVSpoof 2021 LA and DF tasks. In the future, we will further improve the network structure, and the grouping method will also be researched. We are also considering the multi-scale log Gaussian probability feature fusion method, which merges all features at different scales and provides a more effective feature representation. These models will also be applied to speaker recognition.

5. Acknowledgements

This work is supported by National Natural Science Foundation of P.R.China (62067004), and by Educational Commission of Jiangxi Province of P.R.China (GJJ2200331).

6. References

[1] Y. Ren, C. Hu, X. Tan *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[2] Q. Zhu, J. Zhang, Z. Zhang *et al.*, "A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3174–3178.

[3] N. Vaessen and D. A. Van L., "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7967–7971.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[5] E. Morais, R. Hoory, W. Zhu *et al.*, "Speech emotion recognition using self-supervised features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6922–6926.

[6] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[7] H. Tak, J. Patino, M. Todisco *et al.*, "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.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] G. Lavrentyeva, S. Novoselov, E. Malykh *et al.*, "Audio Replay Attack Detection with Deep Learning Frameworks," in *Proc. Interspeech 2017*, 2017, pp. 82–86.

[10] 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.

[11] J. Jung, H. Heo, H. Tak *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[12] A. Tomilov, A. Svishchev, M. Volkova *et al.*, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, December 3-6, 2012, Lake Tahoe, Nevada, United States*, 2012, pp. 1106–1114.

[14] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 301–307.

[15] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[16] Z. Lei, H. Yan, C. Liu *et al.*, "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] Y. Wen, Z. Lei, Y. Yang *et al.*, "Multi-Path GMM-MobileNet Based on Attack Algorithms and Codecs for Synthetic Speech and Deepfake Detection," in *Proc. Interspeech 2022*, 2022, pp. 4795–4799.

[18] J. Jung, Y. Kim, H. Heo *et al.*, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech 2022*, 2022, pp. 2228–2232.

[19] Y. Zhang, Z. Lv, H. Wu *et al.*, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," in *Proc. Interspeech 2022*, 2022, pp. 306–310.

[20] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9241–9245.

[21] 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.

[22] H. Tak, M. Todisco, X. Wang *et al.*, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[23] H. Tak, M. Kamble, J. Patino *et al.*, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6382–6386.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[25] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *Computer Science*, 2015.

[26] J. Yamagishi, X. Wang, M. Todisco *et al.*, "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.

[27] M. Todisco, X. Wang, V. Vestman *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[28] T. Kinnunen, H. Delgado, N. Evans *et al.*, "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.

[29] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 75–82.