Deepfake Algorithm Recognition System with Augmented Data for ADD 2023 Challenge

Xiao-Min Zeng, Jian-Tao Zhang, Kang Li, Zhuo-Li Liu, Wei-Lin Xie and Yan Song*

National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.

Abstract

In this paper, we describe our submitted systems to the ADD2023 Challenge Track 3–Deepfake algorithm recognition (AR). This task requires not only identifying known deepfake algorithms in closed-set but also distinguishing unknown algorithms. By closed-set classification experiments, we select the output of the pre-trained wav2vec2.0-base model as acoustic features. Then, we apply the ECAPA-TDNN model to recognize different deepfake algorithms and determine whether the samples belong to the unknown algorithms by threshold. Besides, we adopt data augmentation to improve the generalization and robustness of our model. We evaluate our system on the ADD2023 Challenge Track 3 and achieve a 75.41% F1-score. Our submission ranked third in the deepfake algorithm recognition track of the ADD2023 Challenge.

Keywords

deepfake algorithm recognition, open-set recognition, data augmentation

1. Introduction

Speech synthesis and voice conversion technologies [1] are evolving rapidly, thanks to the development of deep neural networks. While these advanced technologies provide convenience for application, they also threaten the identity of speaker [2]. For the safety of speakers, audio deepfake detection systems are required to develop, and these defense technologies have attracted lots of research. A classic example is the ASVspoof Challenges [3, 4, 5, 6] that focus on the development of anti-spoofing countermeasures for verification systems. Besides, the first Audio Deepfake Detection Challenge (ADD 2022) [7] introduces more challenging attacking situations in realistic scenarios.

In past work, several handcrafted features are used for deepfake audio detection, including MFCC, iMFCC, LFCC, DCT-DFTspec, log-linear filterbank, and CQT features [8, 9, 10, 11, 12, 13, 14]. Previous works indicate that LFCC or CQT is more suitable for deepfake audio detection. In addition, some works [15, 16, 17] adopt raw waveforms as the input to construct end-to-end deepfake detection systems. Moreover, various data augmentation technologies are also applied to develop detection systems. Codec augmentation are used in [8, 18, 19, 20, 21]. Noise augmentation, reverberation, and cutmix are also essential technologies in anti-spoof systems [20, 21, 22]. It demonstrates that improving the generalization and robustness of the model is one of the key objectives for deepfake detection systems.

However, there are many algorithms for generating deepfake audio. While detecting spoof audio, we would like to know by which algorithms this fake audio is generated. Recently, the second Audio Deepfake Detection Challenge (ADD 2023) [23] is launched, aiming at spurring researchers around the world to build new innovative technologies that can further accelerate and foster research on detecting and analyzing deepfake speech utterances. In ADD2023 Challenge Track 3 [23], the objective of this task is to recognize the algorithms of deepfake utterances. The deepfake algorithms are diverse, and it is impossible to cover all algorithms in the training set. Therefore, the testing process is actually an open-set recognition [24], which requires not only correctly classifying the types of algorithms that appear in the training set but also distinguishing the unknown algorithms.

In the field of open-set recognition, advanced methods have been proposed, *e.g.* OpenMax [25], OpenGAN [26] and OpenHybrid [27]. These methods specifically design a novel module for identifying the open-set samples. Besides, distinguishing the open-set samples based on the comparison of threshold and posterior probabilities is one of the most general methods. It is also interesting to note that S. Vaze et al. [28] proves a positive correlation between closed-set accuracy and open-set recognition performance.

In this paper, we describe our submission for ADD2023 Challenge Track 3 in detail. First, motivated by [28], we conduct classification experiments on closed training set for selecting the optimal input acoustic features. Finally, we choose the output of wav2vec2.0 [29] for the experiment of open-set recognition. The pre-trained wav2vec2.0 model extracts powerful acoustic features,

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R

*Corresponding author.

zxmin115@mail.ustc.edu.cn (X. Zeng); songy@ustc.edu.cn (Y. Song)

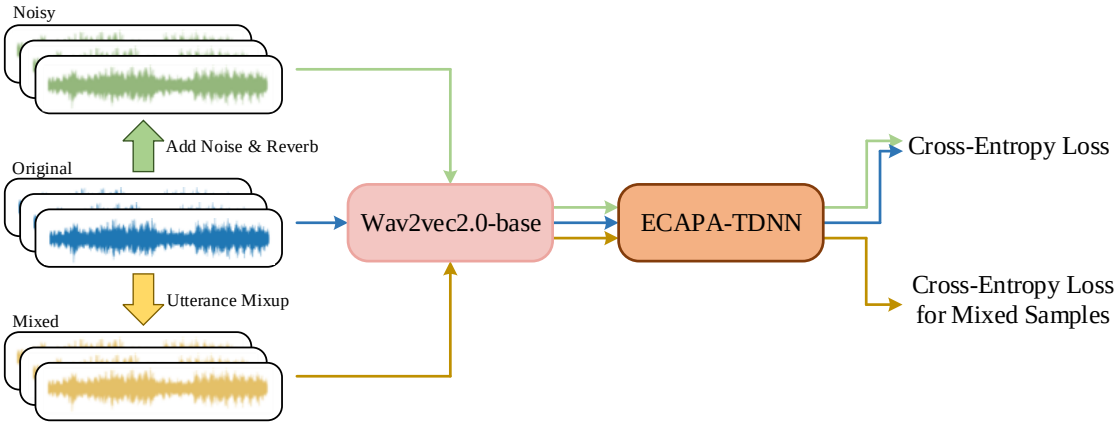

Figure 1: An overview of our deepfake algorithms recognition system. We apply noise and reverb to generate noisy samples. The mixed samples are also obtained by the utterance mixup. All augmented samples, together with the original samples, are used to train the model. Note that the wav2vec2.0-base is frozen during training.

and then these features are fed into ECAPA-TDNN [30] to classify deepfake algorithms. In addition, we exploit several data augmentation, including noise [31], reverberation [32], and utterance mixup [33], to generate augmented samples. In our experiments, all augmented and original samples form a larger training batch for improving the generalization and robustness of our model. Ultimately, we achieve a 75.41% F1-score, which ranks 3rd in ADD2023 Challenge Track 3.

2. Methods

Our deepfake algorithms recognition system is illustrated in Fig. 1. We apply several data augmentation (add noise & reverb [31, 32] and utterance mixup [33]) to obtain augmented samples. These samples are used along with the original samples for training. They are fed into the fixed wav2vec2.0-base model to extract acoustic features and then are classified by ECAPA-TDNN. In this section, we introduce the model's architecture, data augmentation methods, and open-set recognition approaches we used.

2.1. The Architecture of Model

2.1.1. Feature Extractor

Wav2vec2.0 can learn powerful representations from speech audio [29]. After pre-training with a large dataset, it has been demonstrated the feasibility of speech recognition with limited amounts of labeled data.

The raw speech signal is first fed into a feature encoder which consists of several convolution layers (CNN), where the kernel widths and strides behave as the window length and hop of the Short-Time Fourier Transform (STFT). Following wav2vec2.0, every 25ms of audio is extracted as a 512-dimensional vector representation, and the stride is 20ms. Then, the vector representations are processed by a series of transformer blocks. With the help of self-attention module and feed-forward network, the context and content information is shown in representations, which is conducive to learning more informative representations.

The wav2vec2.0 is pre-trained in a self-supervised manner using contrastive loss. The objective is to predict the true quantized latent speech representation for a masked time step within a set of distractors. The output of pre-trained wav2vec2.0 can be applied to downstream supervised learning tasks associated with speech.

In our deepfake algorithm recognition system, we adopt wav2vec2.0-base as the pre-trained feature extractor, aiming at extracting more helpful information from the waveform.

2.1.2. Classification Model

ECAPA-TDNN [30] is a classical model in the field of speaker recognition and verification. As front-end systems of speaker recognition, deepfake detection systems are also required to process speaker-related information. Therefore, we think that ECAPA-TDNN is also suitable for deepfake algorithm recognition systems.

After extracting the representations from wav2vec2.0, we adopt ECAPA-TDNN to aggregate the features further. The SE-Res2Block achieves channel attention that uses a global context incorporated in the frame layers. Then, attentive statistic pooling is used to aggregate utterance-level embeddings about the deepfake algorithm. After the embedding of input speech is obtained, it is fed into an algorithm classifier. Simple, this classifier is achieved by a fully-connected layer.

Table 1
The relationship between k and different data augmentations, where T denotes various thresholds.

random number k	data augmentation methods
$[T_{ori}, 1]$	no augmentation
$[T_{rir}, T_{ori}]$	add noise
$[0, T_{rir}]$	reverb

2.2. Data Augmentation

In this work, we employ several data augmentation methods in order to improve the generalization and robustness of the learned representations.

2.2.1. Add Noise and Reverb

We use the MUSAN dataset [31] and the RIR dataset [32] to randomly add noise or reverb to the input audio. We get a random number k that follows a uniform distribution between 0 to 1 and choose different data augmentation methods depending on k. Specifically, the relationship between k and different data augmentations is shown in Table 1. In our work, the augmented samples and the original samples form a larger batch that is utilized together to optimize the network.

2.2.2. Utterance Mixup

Mixup [33] is one of the common methods of data augmentation in the field of computer vision. We introduce utterance mixup to speech deepfake algorithm recognition. The utterance mixup aims to simulate multiple deepfake algorithm audio and improve model generalization.

To generate the mixed audio from multiple deepfake algorithms, we randomly select two utterances x_1, x_2 from each training batch and mix them according to the following equation,

$$x_m = \lambda x_1 + (1 - \lambda) x_2 \tag{1}$$

where mixing coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$. Obviously, the mixed sample contains two different deepfake algorithms, which means the mixed samples belong to two different algorithmic categories. Therefore, the label of mixed sample is modified as follows,

$$y_m = \lambda y_1 + (1 - \lambda) y_2 \tag{2}$$

Then, the loss function is also modified to,

$$\mathcal{L} = \lambda \mathcal{L}_1 + (1 - \lambda) \mathcal{L}_2 \tag{3}$$

where $\mathcal{L}_n$ represents the loss calculated with the label y_n corresponding to sample x_n.

2.3. Unknown Algorithm Recognition

In the training procedure, we follow the closed-set setting, and the model is optimized to perform well under the multi-classes classification task. The objective is to accurately distinguish the deepfake algorithms in the training set so that a compact category distribution is obtained.

During the testing phase, we have to solve an open-set recognition task. We first calculate the prototypes p_c of known algorithm categories in the training set,

$$p_c = \frac{1}{|\mathcal{S}_c|} \sum_{x_i \in \mathcal{S}_c} f_\theta(x_i) \tag{4}$$

where $\mathcal{S}_c$ is a set of all training samples labeled with deepfake algorithm c and $|\mathcal{S}_c|$ is its cardinality. f_θ is the model we use with parameter θ.

After obtaining prototypes, we calculate the cosine-similarity between test embedding and prototypes p_c. The category corresponding to the maximum similarity $simi_{max}$ is the classification result we determine. For unknown algorithm recognition, if the maximum similarity is lower than a pre-defined threshold T_{open}, this sample is regarded as an open-set category, *i.e.* unknown deepfake algorithm. This process can be expressed as,

$$\textbf{\textit{result}} = \begin{cases} \textbf{\textit{known class}} & \textit{if } simi_{max} \geq T_{open} \\ \textbf{\textit{unknown class}} & \textit{if } simi_{max} \leq T_{open} \end{cases} \tag{5}$$

This is one of the most widely applied methods of open-set recognition. The method is simple and intuitive, so we adopt it to distinguish unknown deepfake algorithms. Meanwhile, this threshold determination method has to guarantee the accuracy of the closed-set categories.

3. Experimental Setup

3.1. Datasets

The ADD2023 Challenge Track 3 [23] aims to recognize the algorithms of deepfake utterances. The training dataset contains seven classes, including one real category and six deepfake algorithm generated categories. In addition, ADD2023 also released a development dataset

Table 2
The number of samples for each category, where 'sr' means audio sampling rate for the corresponding category.

	0	1	2	3	4	5	6
train	3200	3200	3197	3200	3200	3200	3200
dev	1200	1200	1200	1200	1200	1200	1200
sr	16k	16k	16k	24k	24k	16k	16k

Table 3
The accuracy of closed-set classification.

Feature	Accuracy
MFCC	90.56
LFCC	**99.48**
iMFCC	93.62
wav2vec2.0	**99.63**

for model optimization and selection. The number of samples for each category in the training and development datasets is presented in Table 2. It can be found that there is a difference in the sampling rate. Therefore, we resample all the audio so that all sampling rates are 16kHz. We integrate all samples in the training and development datasets into a larger dataset for optimizing the model.

In the test set, there are 79490 test utterances in total. It contains eight categories, seven of which are the same as the training and development datasets. Besides, one additional deepfake algorithm category is added as an open-set class. Therefore, the ADD2023 Challenge Track 3 is a typical open-set recognition task.

3.2. Implementation Details

Our network consists of wav2vec2.0-base and ECAPA-TDNN. In our experiments, we employ an open-source, self-supervised pre-trained wav2vec2.0-base model. The output of wav2vec2.0-base is 768-dimensional features. Thus the input dimension of ECAPA-TDNN needs to be modified to match. In the training phase, we fix the parameters of the pre-trained wav2vec2.0-base model and optimize only the ECAPA-TDNN model.

For each training utterance, we randomly crop 4 seconds of audio to construct batches with a batch size of 64. An SGD algorithm is utilized with a momentum of 0.9 and weight decay of 1e-4 to optimize our model over 50 epochs. The initial learning rate of 0.1 is used for training during the first 25 epochs, declining to 0.01 between the 25th and 40th epochs and 0.001 for the remaining 10 epochs. Moreover, the hyper-parameter α in utterance mixup is set to 1.0 in our experiments, and we select the optimal T_{open} by grid searching.

Table 4
The precision rate (%), recall rate (%) and F1-score (%) for various input features and classification models.

Feature	Classifier	Precision	Recall	F1-score
LFCC	ResNet18	62.88	60.60	60.63
wav2vec2.0	ResNet18	**81.63**	57.21	63.15
wav2vec2.0	ECAPA-TDNN	77.57	**61.75**	**65.70**

3.3. Evaluation Metrics

For ADD2023 Challenge Track 3, we are required to recognize the known and unknown algorithms of the deepfake utterances. Therefore, the performance is evaluated by macro-average precision, recall, and F1-score. The final ranking of this challenge is based on the F1-score [23].

4. Results

4.1. Closed-set Experimental Results

According to the conclusion in [28], the more accurate the closed-set classification, the better the performance of open-set recognition. Therefore, in order to select the optimal input acoustic features, we first conduct closed-set classification experiments on the training set. We adopt different acoustic features as input and classify seven deepfake algorithms of the training set using ResNet18 with Time-Frequency Attention Pooling [34]. The results of closed-set classification are presented in Table 3. It is not hard to obverse that when LFCC and wav2vec2.0 are utilized, the closed-set can achieve satisfactory performances. We believe that these two features can reflect more information about deepfake algorithms, so we choose them for the following open-set recognition experiments.

4.2. Open-set Recognition Results

We compare the performance of various input features and classification models for open-set recognition. We adopt grid search for unknown algorithm thresholds T_{open} and present the optimal results in Table 4. The results indicate that the output of pre-trained wav2vec2.0 captures higher-level semantic information and more detailed information from waveform than LFCC. It is more beneficial for deepfake algorithm recognition. The ResNet18 exploits 2D convolution to process features

Table 5
The precision rate (%), recall rate (%) and F1-score (%) of ablation experiments with different data augmentation, where '*N & R*' represents add noise and reverb and '*UM*' defines as utterance mixup.

Data augmentation				Evaluation metrics		
N & R	*UM*	T_{ori}	T_{rir}	Precision	Recall	F1-score
✗	✗	-	-	77.57	61.75	65.70
✓	✗	0.75	0.25	76.60	71.60	72.59
✓	✗	0.9	0.25	78.93	72.29	73.81
✓	✗	1	0.25	76.63	73.71	74.28
✗	✓	-	-	78.04	70.61	72.44
✓	✓	1	0.25	**79.09**	**74.30**	**75.41**

and treats the input as images, which is not the most appropriate way for audio. Instead, the ECAPA-TDNN is designed specifically for speech tasks. The results also show that ECAPA-TDNN achieves better performance. Besides, we discover that when using ECAPA-TDNN as a classification model, the optimal unknown algorithm threshold T_{open} is much larger than that when ResNet18 is employed. This phenomenon suggests that ECAPA-TDNN models a more compact embedding space for closed-set samples, which results in a higher confidence level and better performance. This conclusion is consistent with [28].

To improve the generalization and robustness of the model, we applied various data augmentation in the training process. In Table 5, we study the influences of the different data augmentations on performance. After adding noise and reverbing, the performance is significantly improved. The larger value of T_{ori} means that more augmented samples caused by noise and reverberation are generated. We find that the model makes further progress as the number of augmented samples increases. Utilizing utterance mixup on the basis of noise addition and reverberation, we further improve the performance. With the help of several data augmentation, the optimal F1-score of 75.41% is achieved. This result ranks 3rd in the ADD2023 Challenge Track 3.

5. Conclusions

This paper has presented our proposed systems for the ADD2023 Challenge Deepfake Algorithm Recognition Track, which is based on pre-trained wav2vec2.0-base and ECAPA-TDNN models. According to the accuracy of closed-set algorithm classification experiments, we select the output of pre-trained wav2vec2.0-base as acoustic features. While the wav2vec2.0 is robust and general, we adapt the ECAPA-TDNN to deepfake algorithm recognition through data augmentation techniques. Our proposed systems show competitive results and rank third in Track 3 of the ADD2023 Challenge. In future work, more general data augmentation and more efficient open-set recognition approaches should be further considered.

References

[1] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2020) 132–157.

[2] C. B. Tan, M. H. A. Hijazi, N. Khamis, P. N. E. b. Nohuddin, Z. Zainol, F. Coenen, A. Gani, A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction, Multimedia Tools and Applications 80 (2021) 32725–32762.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Sixteenth annual conference of the international speech communication association, 2015.

[4] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K. A. Lee, The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection (2017).

[5] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. A. Lee, Asvspoof 2019: Future horizons in spoofed and fake audio detection, arXiv preprint arXiv:1904.05441 (2019).

[6] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, 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. doi:10.21437/ASVSPOOF.2021-8.

[7] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, et al., Add 2022: the first audio deep synthesis detection challenge, in: ICASSP2022, IEEE, 2022, pp. 9216–9220.

[8] W. H. Kang, J. Alam, 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. doi:10.21437/ASVSPOOF.2021-16.

[9] W. H. Kang, J. Alam, A. Fathan, Investigation on activation functions for robust end-to-end spoofing attack detection system, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 83–88. doi:10.21437/ASVSPOOF.2021-13.

[10] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, G. Lavrentyeva, STC Antispoofing Systems for the ASVspoof2021 Challenge, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 61–67. doi:10.21437/ASVSPOOF.2021-10.

[11] M. Sahidullah, T. Kinnunen, C. Hanilci, A comparison of features for synthetic speech detection, in: Proc. Interspeech 2015, 2015, pp. 2087–2091. doi:10.21437/Interspeech.2015-472.

[12] M. Todisco, H. Delgado, N. W. Evans, A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients., in: Odyssey, vol-

ume 2016, 2016, pp. 283–290.

[13] J. Yang, R. K. Das, H. Li, Extended constant-q cepstral coefficients for detection of spoofing attacks, in: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2018, pp. 1024–1029.

[14] R. K. Das, J. Yang, H. Li, Long Range Acoustic Features for Spoofed Speech Detection, in: Proc. Interspeech 2019, 2019, pp. 1058–1062. doi:10.21437/Interspeech.2019-1887.

[15] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with rawnet2, in: ICASSP2021, IEEE, 2021, pp. 6369–6373.

[16] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, N. Evans, End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 1–8. doi:10.21437/ASVSPOOF.2021-1.

[17] W. Ge, J. Patino, M. Todisco, N. Evans, Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 22–28. doi:10.21437/ASVSPOOF.2021-4.

[18] 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. doi:10.21437/ASVSPOOF.2021-5.

[19] T. Chen, E. Khoury, K. Phatak, 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. doi:10.21437/ASVSPOOF.2021-14.

[20] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.-M. Wang, H. Meng, Partially fake audio detection by self-attention-based fake span discovery, in: ICASSP2022, IEEE, 2022, pp. 9236–9240.

[21] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, X. Li, Audio deepfake detection system with neural stitching for add 2022, in: ICASSP2022, IEEE, 2022, pp. 9226–9230.

[22] Z. Lv, S. Zhang, K. Tang, P. Hu, Fake audio detection based on unsupervised pretraining models, in: ICASSP2022, IEEE, 2022, pp. 9231–9235.

[23] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, H. Li, Add 2023: the second audio deepfake detection challenge, in: IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), 2023.

[24] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 3614–3631.

[25] A. Bendale, T. E. Boult, Towards open set deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1563–1572.

[26] S. Kong, D. Ramanan, Opengan: Open-set recognition via open data generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 813–822.

[27] H. Zhang, A. Li, J. Guo, Y. Guo, Hybrid models for open set recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 102–117.

[28] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Open-set recognition: A good closed-set classifier is all you need?, in: International Conference on Learning Representations, 2022.

[29] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[30] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Proc. Interspeech 2020, 2020, pp. 3830–3834. doi:10.21437/Interspeech.2020-2650.

[31] D. Snyder, G. Chen, D. Povey, Musan: A music, speech, and noise corpus, arXiv preprint arXiv:1510.08484 (2015).

[32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, S. Khudanpur, A study on data augmentation of reverberant speech for robust speech recognition, in: ICASSP2017, IEEE, 2017, pp. 5220–5224.

[33] H. Zhang, M. Cissé, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: 6th International Conference on Learning Representations, 2018.

[34] H. Chen, Y. Song, L.-R. Dai, I. McLoughlin, L. Liu, Self-supervised representation learning for unsupervised anomalous sound detection under domain shift, in: ICASSP2022, IEEE, 2022, pp. 471–475.