

Frequency-mix Knowledge Distillation for Fake Speech Detection

Cunhang Fan¹, Shunbo Dong¹, Jun Xue¹, Yujie Chen¹, Jiangyan Yi², Zhao Lv¹

¹School of Computer Science and Technology, Anhui University, China

²Institute of Automation, Chinese Academy of Sciences, China

e22201030@stu.ahu.edu.cn, e22201148@stu.ahu.edu.cn

Abstract

In telephony scenarios, the fake speech detection (FSD) task to combat speech spoofing attacks is challenging. Data augmentation (DA) methods are considered effective means to address the FSD task in telephony scenarios, typically divided into time domain and frequency domain stages. While each has its advantages, both can result in information loss. To tackle this issue, we propose a novel DA method, Frequency-mix (Freqmix), and introduce the Freqmix knowledge distillation (FKD) to enhance model information extraction and generalization abilities. Specifically, we use Freqmix-enhanced data as input for the teacher model, while the student model's input undergoes time-domain DA method. We use a multi-level feature distillation approach to restore information and improve the model's generalization capabilities. Our approach achieves state-of-the-art results on ASVspoof 2021 LA dataset, showing a 31% improvement over baseline and performs competitively on ASVspoof 2021 DF dataset.

Index Terms: fake speech detection, data augmentation, knowledge distillation

1. Introduction

With the rapid development of text-to-speech (TTS) and voice conversion (VC), it has become increasingly challenging for the human ear to distinguish genuine speech from fake speech. Fake speech detection (FSD) task aims to devise effective counter measures (CM) that bolster the resilience of Automatic Speaker Verification (ASV) system against deceptive attacks. While previous studies [1, 2, 3, 4] have mainly focused on FSD task under controlled laboratory settings, ignoring the practical implications introduced by data encoding, compression, and transmission. FSD task in the logical access (LA) scenario of the ASVspoof 2021 challenge [5] involves training models in controlled laboratory conditions and subsequently deploying them for anti-spoofing tasks on real-world speech data. Addressing the generalization of the model becomes particularly important.

Data augmentation (DA) consistently stands out as an important technique for improving model generalization. Notably, Rawboost [6] is devised for the FSD task in scenarios involving communication-distorted and compressed-coded speech data. It operates directly on the raw waveform, causing signal distortion in training samples, thereby enhancing the model's generalization capability. SpecAugment [7] treats the log mel spectrogram as an image, applying masking on contiguous time steps and mel frequency segments to enhance model performance. Specmix [8] combines two training samples to create a new sample, with the labels of the combined corresponding samples serving as the label for the new sample, directly impacting fea-

tures in the time-frequency domain. MixSpeech [9] combines two distinct speech features through weighted blending to generate a novel feature for Automatic Speech Recognition (ASR). While these masking methods can indeed enhance the model's generalization ability, applying them directly to the input of a classification model may potentially erase artefacts directly.

In recent years, the knowledge distillation (KD) method has often been utilized for model compression [10, 11], condensing a network with thousands of layers into a smaller model. OCKD [12] combines one-class method with KD, reducing the number of layers in Wav2vec 2.0 [13, 14] to prevent overfitting caused by an excessive number of parameters. DKDSSD [15] is a dual-branch method using KD to address the issue of performance degradation of detecting spoofed speech in noisy environments. However, as the research progresses, researchers have discovered that a student model with the same amount of training parameters as the teacher model can obtain better results than the teacher model [16, 17]. [18] proposed a self-distillation method, using the deep network as the teacher model and the shallow network as the student model, using deep information to guide shallow information learning to improve fine-grained information recognition. While these works involve KD, we believe that transferring knowledge between completely identical input data may not extract sufficiently rich information.

In this study, we propose a method called Freqmix knowledge distillation (FKD) to improve the information extraction and generalisation capabilities of the model. Taking advantage of the successful performance of Rawboost for the FSD task in telephony scenarios, the input data of our student model is augmented by Rawboost. Freqmix divides the teacher model input into two parts: one with original audio and the other with spectrogram-masked speech samples. The diverse inputs in the teacher model guide the student model to learn varied knowledge. The original data helps to restore artefacts from distorted signals, thus improving the student model information extraction ability. Meanwhile, spectrogram-enhanced data enables the student model to combine the effects of spectrogram masking and signal distortion, thereby boosting the model's generalization capabilities. In conclusion, for the FSD task in telephony scenarios, our proposed FKD method innovatively combines knowledge distillation with different DA methods and original information, further improving the model's performance. Compared to the baseline, the performance improvement on the ASVspoof 2021 LA evaluation set is 31%. Competitive results are maintained on the ASVspoof 2021 DF dataset.

Rest of the paper is organized as follows: Section 2 describe our method. Experiments, results, and discussions are reported in Section 3. Finally, we conclude the paper in Section 4.

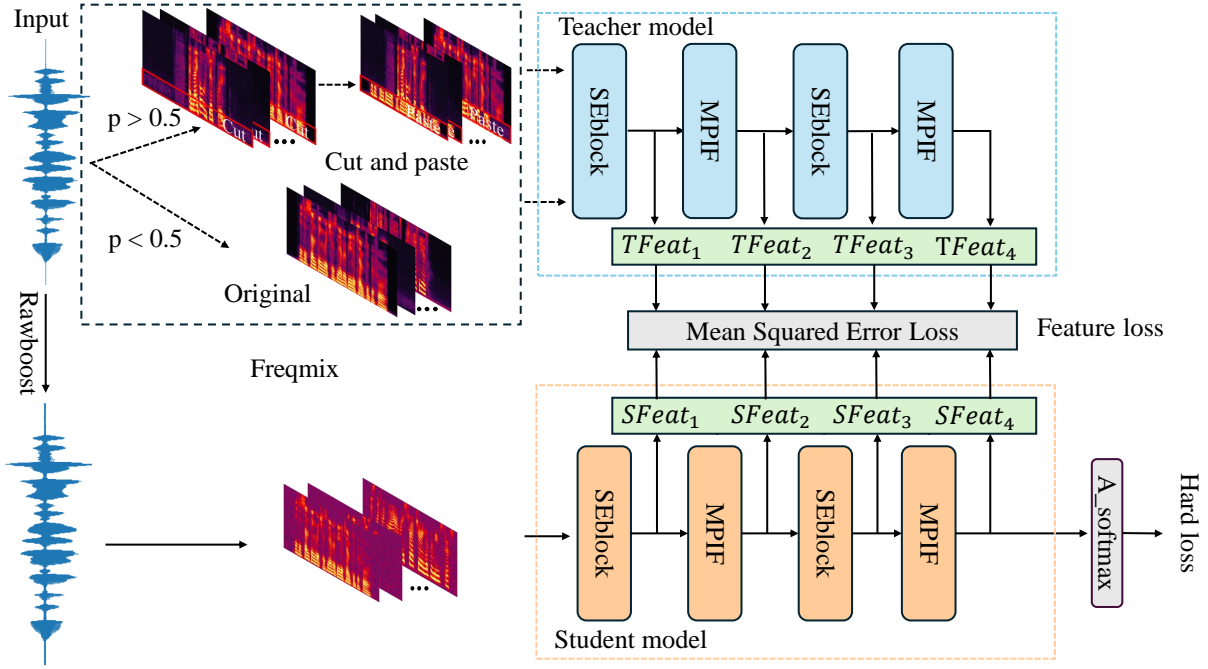


Figure 1: The illustration of our proposed Freqmix knowledge distillation (FKD) for FSD method in telephony scenarios. The student model and the teacher model both adopt the MPIF-Res2Net architecture an identical number of parameters. During the training of the student model, the parameters of the teacher model remain unchanged.

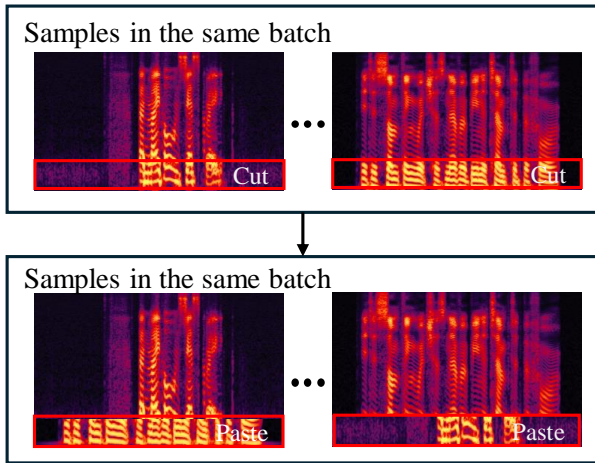


Figure 2: The illustration of the cut and paste operation among the samples in the same batch

2. Proposed Method

In this section, we show the detailed method of FKD, and its overall architecture is shown in Figure 1. First we introduce the process of Freqmix, and then we introduce the specific distillation strategy.

2.1. Freqmix Data Augmentation

Inspired by [8] [19] and our previous work [20], we further develop Specmix and introduce the Freqmix method, a fre-

quency domain DA method. The cut and paste operation is employed on spectrogram of samples within the same batch. Like Specmix, it does not introduce external data. In the cut and paste operation, the content of the identical frequency range ($f_0, f_0 + f$) in each sample is cut out, shuffled and then pasted back into the corresponding frequency range ($f_0, f_0 + f$) of the original samples, following the order established during the shuffle. The cut and paste operation is shown in Figure 2.

To ensure the student model effectively restores the original information and learns the enhancement effects of spectrogram masking, the teacher model must have both the raw audio information and the information after applying the DA method of spectrogram masking. Freqmix leaves part of the teacher model's input unchanged, while applying spectrogram masking to the other part of the data. This allows the student model to learn the original information, compensating for the information distortion introduced by Rawboost, the time-domain DA method, and also combines the DA effects of spectral masking with Rawboost to further improve the model's generalization ability. Therefore, before applying the spectral masking operation to the input of the teacher model, a decision must be made on whether or not to perform cut and paste operation. First, a random value, p , is generated. If $p > 0.5$, the input data for the teacher model is subjected to the cut and paste operation. After processing through the teacher model, the effects of spectrogram masking can be transferred to the student model. This enables the data of the student model to integrate the enhancement effects of Rawboost and spectrogram masking, the generalization ability of the student model is improved. On the other hand, when $p < 0.5$, the input data to the teacher model is the original data. This portion of the data is utilized to correct potential information loss in the student model caused by

Rawboost, artefacts can be restored, the information extraction ability of student model is enhanced.

2.2. Knowledge Distillation

MPIF-Res2Net[20], within the same channel group, merges information from different receptive fields through channel attention. This addresses the issue of information redundancy caused by a single receptive field in Res2Net, mitigating the problem of masking useful information. As shown in Figure 1, the teacher model and the student model constitute the two parts of our FKD method. The input data of the teacher model undergoes DA with Freqmix, in order to cope with the FSD task in real communication scenarios, the input of the student model is enhanced through Rawboost. The student model learns only the relevant knowledge of the output features of each layer from the teacher model, ignoring the prediction results of the teacher model. The purpose is to enable the student model’s feature, $SFeat_i$, to fully capture the features of the teacher model’s feature, $TFeat_i$, independently of the teacher model’s prediction results. This is advantageous for more effective information recovery and the integration of different DA effects. The feature loss function between the student model and the teacher model uses the Mean Squared Error (MSE) function. The MSE function is highly sensitive to the difference between two inputs, making it an appropriate choice for a feature loss function. The calculation method for the feature loss, denoted as L_{feat} , is as follows:

$$Loss_{feat} = \sum_{i=1}^4 MSE(TFeat_i, SFeat_i) \quad (1)$$

Here, $TFeat_i$ and $SFeat_i$ respectively denote the outputs of the i th feature layer for the teacher and student model. In the end, the feature loss $Loss_{feat}$ is obtained.

For the loss function $Loss_{hard}$ between the student model’s predictions and labels, we employ the $A_softmax$ [21]:

$$Loss_{hard} = A_softmax(predict, label) \quad (2)$$

Here, $predict$ represents the prediction outcome of the student model, and $label$ denotes the label of samples in the dataset. $A_softmax$ corresponds to our loss function.

Finally, the ultimate loss function $loss$ is obtained by taking the weighted sum of $Loss_{feat}$ and $Loss_{hard}$:

$$loss = \alpha Loss_{feat} + \beta Loss_{hard} \quad (3)$$

Where α and β are the weight of $Loss_{feat}$ and $Loss_{hard}$.

The training of the teacher model and the student model are independent. During the distillation stage, the parameters of the teacher model are frozen and do not participate in the update. In the pre-training stage, the teacher model only uses Rawboost for DA. The purpose of this is to be consistent with the student model. MPIF-Res2Net is selected as the classification model. According to our experience, the model based Res2Net can adapt well to the time-frequency features.

3. Experiments and Results

3.1. Datasets

Two datasets are used in our experiments. The progress subset of the ASVspoof 2021 LA dataset comprises 1,676 genuine speech samples and 14,788 synthesized speech samples. The evaluation subset of ASVspoof 2021 LA dataset has 14,816

bona fide samples and 133,360 spoof utterances, all these samples are transmitted through authentic telephony systems. ASVspoof 2021 LA task aims to design the spoofing countermeasures to enhance the capability of models to detect fake speech under various unknown channel variations. The primary metric for this task are Equal Error Rate (EER) and the minimum tandem detection cost function (t-DCF). The evaluation subset in the ASVspoof 2021 DF dataset consists of a greater number of samples. These samples have been subjected to a sequence of encoding and decoding steps, introducing distortions associated with the codec devices. The evaluation process specifically focuses on the EER as the primary metric.

3.2. Experimental Setup

Training a model on a clean dataset and achieving robust performance on an evaluation set with communication interference poses a significant challenge. During the pre-training stage, we employ Rawboost augmentation to the input of teacher model, introducing noise to the original waveform. The noise types include impulsive signal-dependent (ISD) additive noise and stationary signal-independent (SSI) additive noise.

The network takes subbands of the log power spectrum (LPS) in the range of 0-45Hz as input. In the Freqmix DA method, the maximum frequency span f during frequency masking does not exceed 10Hz, the value for deciding whether to apply masking enhancement to the input of the teacher model is set to 0.5. In the short-time Fourier transform, we utilize the Blackman window function, with a window length of 1728 and a hop length of 130. Inspired by [22], the frequency band of 0-45 dimension is retained, truncated, and flipped before concatenation, resulting in a fixed frame length of 600. The final feature input to the network is a 45×600 F0 sub-band feature map. We employ the MPIF-Res2Net network as the classifier, with Adam used as the optimizer, a batch size of 32, and a learning rate of 0.0001. The random seed is initialized to 1. In the final loss calculation, the weights α and β for feature loss and hard loss are set to 0.2 and 0.8, respectively.

Table 1: Results of ablation study of the proposed FKD system on the progress and evaluation subset of ASVspoof 2021 LA dataset. Tea means the teacher model, our baseline, before KD; Stu denotes the student model, after KD; The "F" and "R" in parentheses indicate applying the Freqmix and Rawboost DA method to the model’s input, respectively. "C" means the clean data. Our experimental results are highlighted in bold.

System	Evaluation set		Progress set	
	t-DCF	EER (%)	t-DCF	EER (%)
Tea(R)	0.2910	4.19	0.2860	4.77
Tea(F+R)	0.2547	3.31	0.2250	3.04
Tea(C)..Stu(R)	0.2520	3.10	0.2236	2.92
FKD (ours)	0.2460	2.88	0.2184	2.75

3.3. Results

3.3.1. Ablation Study

Table 1 presents the results of various experiments on the progress and evaluation subsets of ASVspoof 2021 LA dataset. Tea(R) serves as the baseline, employing Rawboost directly on MPIF-Res2Net inputs. The approach for Tea(F+R) involves

simultaneously applying Freqmix and Rawboost for MPIF-Res2Net inputs. Experiments Tea(C)_.Stu(R) and FKD show the results of the student model after distillation. In Tea(C)_.Stu(R), the teacher model employs the original dataset, while the student model is enhanced with Rawboost. In FKD, the teacher model’s input incorporates Freqmix, while the student model’s input uses Rawboost. Tea means the teacher model, Stu means the student model. Notably, the Tea(R) is our baseline, the inputs were augmented by Rawboost, lacking original speech restoration, attains an EER of 4.19% and a t-DCF of 0.2910 on the evaluation subset. The experiment Tea(C)_.Stu(R) shows when the teacher model uses the original data to perform information compensation on the Rawboost-enhanced input of the student model, the obtained EER and t-DCF are 3.10% and 0.2520 respectively. This confirms that DA may cause information loss and using original data for information recovery is a viable approach to adjusting the extent of DA, thereby enhancing model performance. On the other hand, in the Tea(F+R) experiment where Freqmix and Rawboost data enhancement were applied to the MPIF-Res2Net model at the same time, the EER was 3.31% and the t-DCF was 0.2547, which seems to be a good result. However, in the FKD experiment, where Freqmix data augmentation was used as input to the teacher model, the student model achieved EER and t-DCF results of 2.88% and 0.2460, respectively. We attribute this improvement to the fusion of DA methods and the compensation of information in the student model input. Specifically, Freqmix retains some information in the original data to compensate for the information loss caused by Rawboost enhancement. In addition, the teacher model imparts the masking enhancement effect in the two-dimensional spectrogram to the data in the student model, so that the masking enhancement effect of Freqmix and communication enhancement effect of Rawboost are integrated in the student model. Therefore, the performance of the model is further improved.

Table 2: Results Comparison with Fusion Systems on the Performance of progress and evaluation subset of ASVspoof 2021 LA Dataset. Our experimental results are highlighted in bold.

System	Evaluation set		Progress set	
	t-DCF	EER (%)	t-DCF	EER (%)
T06 [23]	0.2853	5.66	0.2476	5.61
Fusion systems [24]	0.2882	4.66	–	–
T36 [23]	0.2531	3.10	0.2373	3.69
T35 [23]	0.2480	2.77	0.2115	2.61
T23 [23]	0.2176	1.32	0.1816	0.89
FKD (ours)	0.2460	2.88	0.2184	2.75

3.3.2. Performance Comparison With Other Systems

Table 2 shows the results of the FKD method and other different fusion models on the eval subset and progress subset of ASVspoof 2021 LA dataset. The comparison shows that the performance of our method even exceeds some fusion systems. Although the results of T23 and T35 are better than our model, their fusion strategies are quite complex. For example, the T23 fusion system consists of multiple subsystems extracting different spectral features from various codec-augmented data to train different classifiers, and finally performs weighted average score fusion. Additionally, the t-DCF value of 0.2480 on the eval subset for T35 is slightly higher than the t-DCF value of

Table 3: Results Comparison with Fusion Systems on the Performance of progress and evaluation subset of ASVspoof2021 LA Dataset. Our experimental results are highlighted in bold.

System	min t-DCF	EER(%)
RawGAT-ST [25]	0.3782	6.92
M-GMM-MobileNet(C) [26]	0.3231	6.80
RawNet2 [25]	0.3099	5.31
AASIST [27]	0.3398	5.59
DFIM [28]	0.2601	3.05
FKD (ours)	0.2460	2.88

our FKD system. This indicates that the reliability of the fusion system T35 is slightly lower than our approach when working in tandem with the ASV system.

Table 3 shows the effect of our method on the ASVspoof 2021 LA evaluation subset compared with the FKD system and other single systems. It can be seen that our method has good performance compared with other single systems. Compared to the best-known experimental results to date, our method has shown improvements in both EER and t-DCF. This indicates that our method yields more accurate classification results and is more reliable when used in tandem with an ASV system.

Table 4: Results Comparison with single Systems on the Performance of ASVspoof2021 DF Dataset. Our experimental results are highlighted in bold.

System	21DF EER(%)
ARawNet2 [29]	19.03
T06 [23]	19.01
T22 [23]	19.22
CQT-LCNN(D3) [30]	18.31
T08 [23]	18.30
ResNet-S-LDE [31]	17.25
FKD (ours)	17.16

Table 4 presents a comparison of the results between FKD and other methods. It can be observed that our approach still maintains competitive results on the DF dataset.

4. Conclusion

In this article we propose an approach that combines the DA and KD methods. First, we introduce the Freqmix DA method. This method divides the input of the teacher model into two parts and imparts two types of knowledge to the teacher model: original information and the DA effect of spectrogram masking. In the KD method, the teacher model imparts both types of knowledge to the student model. Specifically, the original information in the teacher model helps to restore some artefacts in the student model caused by Rawboost, thereby improving the information extraction ability of the student model. At the same time, the DA effect of spectrogram masking in the teacher model is transferred to the student model and merges with the Rawboost in the student model, further improving the generalisation ability of the model. Our experimental results surpass the best results of all individual systems, demonstrating the effectiveness of combining DA and KD methods in the FSD task under telephony scenarios.

5. Acknowledgements

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No.62201002, No.62322120), Distinguished Youth Foundation of Anhui Scientific Committee (No. 2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province (No. 202203a07020008), Open Fund of Key Laboratory of Flight Techniques and Flight Safety, CACC (No. FZ2022KF15).

6. References

- [1] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, “Voice spoofing countermeasure for logical access attacks detection,” *IEEE Access*, vol. 9, pp. 162 857–162 868, 2021.
- [2] 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.
- [3] R. K. Das, J. Yang, and H. Li, “Long range acoustic and deep features perspective on asvspoof 2019,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1018–1025.
- [4] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, “Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.
- [5] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, “Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan,” *arXiv preprint arXiv:2109.00535*, 2021.
- [6] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, “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)*. IEEE, 2022, pp. 6382–6386.
- [7] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.
- [8] G. Kim, D. K. Han, and H. Ko, “SpecMix : A Mixed Sample Data Augmentation Method for Training with Time-Frequency Domain Features,” in *Proc. Interspeech 2021*, 2021, pp. 546–550.
- [9] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, “Mixspeech: Data augmentation for low-resource automatic speech recognition,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7008–7012.
- [10] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *stat*, vol. 1050, p. 9, 2015.
- [11] C. Fan, Y. Chen, J. Xue, Y. Kong, J. Tao, and Z. Lv, “Progressive distillation based on masked generation feature method for knowledge graph completion,” *arXiv preprint arXiv:2401.12997*, 2024.
- [12] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, “One-class knowledge distillation for spoofing speech detection,” *arXiv preprint arXiv:2309.08285*, 2023.
- [13] X. Wang and J. Yamagishi, “Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures,” in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 100–106.
- [14] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, “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.
- [15] C. Fan, M. Ding, J. Tao, R. Fu, J. Yi, Z. Wen, and Z. Lv, “Dual-branch knowledge distillation for noise-robust synthetic speech detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [16] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born again neural networks,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 1607–1616.
- [17] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep mutual learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4320–4328.
- [18] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, “Learning from yourself: A self-distillation method for fake speech detection,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [19] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.
- [20] S. Dong, J. Xue, C. Fan, K. Zhu, Y. Chen, and Z. Lv, “Multi-perspective information fusion res2net with randomspecmix for fake speech detection,” *arXiv e-prints*, pp. arXiv–2306, 2023.
- [21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.
- [22] C. Fan, J. Xue, J. Tao, J. Yi, C. Wang, C. Zheng, and Z. Lv, “Spatial reconstructed local attention res2net with f0 subband for fake speech detection,” *Neural Networks*, vol. 175, p. 106320, 2024.
- [23] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, “Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [24] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, “A study on data augmentation in voice anti-spoofing,” *Speech Communication*, vol. 141, pp. 56–67, 2022.
- [25] 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)*. IEEE, 2021, pp. 6369–6373.
- [26] Y. Wen, Z. Lei, Y. Yang, C. Liu, and M. Ma, “Multi-path gmm-mobilenet based on attack algorithms and codecs for synthetic speech and deepfake detection,” in *INTERSPEECH*, 2022, pp. 4795–4799.
- [27] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “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)*. IEEE, 2022, pp. 6367–6371.
- [28] B. Huang, S. Cui, J. Huang, and X. Kang, “Discriminative frequency information learning for end-to-end speech anti-spoofing,” *IEEE Signal Processing Letters*, vol. 30, pp. 185–189, 2023.
- [29] J. Li, Y. Long, Y. Li, and D. Xu, “Advanced RawNet2 with Attention-based Channel Masking for Synthetic Speech Detection,” in *Proc. INTERSPEECH 2023*, 2023, pp. 2788–2792.
- [30] R. K. Das, “Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: Asvspoof 2021,” *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 29–36, 2021.
- [31] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, “Pindrop labs’ submission to the asvspoof 2021 challenge,” *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 89–93, 2021.