"Enhancing Robustness in Audio Deepfake Detection for VR Applications using data augmentation and Mixup"

ABSTRACT

The rapid advancement of virtual reality (VR) technology has heightened the need for robust and reliable deepfake audio detection to ensure the authenticity and integrity of virtual interactions. Although current state-of-the-art models exhibit promising results, they are often overconfident, which can lead to poor generalization and reduced effectiveness against novel or slightly altered deepfake attacks. In this work, we investigate the application of data augmentation techniques and Mixup techniques to increase the diversity of training data and improve the generalization of deepfake audio detection models. Mixup creates new training examples by combining pairs of existing examples, promoting smoother and more robust decision boundaries, while data augmentation creates new training examples altering a sample with a given probability. Our results demonstrate that applying such techniques to the Wav2vec 2.0 model significantly improves its generalization ability, leading to more reliable deepfake detection in VR environments.

KEYWORDS

Deepfake Detection, Audio Classification, Machine Learning, Feature Abstraction, Mixup

ACM Reference Format:
. 2018. "Enhancing Robustness in Audio Deepfake Detection for VR Applications using data augmentation and Mixup". In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Virtual reality is a computer-simulated environment that allows users to interact with it using special glasses, replacing natural vision and hearing with artificial images and sounds, creating a deep immersion [5]. However, with the advances in audio-based artificial intelligence technologies, such as voice deepfakes, the immersive experience provided by virtual reality can become potentially dangerous.

Deepfakes are AI techniques that create extremely realistic falsified digital content, such as images, videos, and audio [8]. Utilizing deep learning models based on convolutional neural networks (CNNs) and generative adversarial networks (GANs), these models trained with large amounts of data can accurately replicate the visual and auditory characteristics of a specific person. In the case of voice deepfakes, recent voice conversion models [9] can mimic the timbre, intonation, and rhythm of a person's speech, creating audio that is almost indistinguishable from the real thing.

Unfortunately, these technologies have been exploited by criminals to carry out sophisticated scams and spread false news, causing serious disruptions [2, 3]. In a fully immersive virtual environment, where the user's senses are completely engaged and their ability to discern the real from the artificial is reduced, the risk of malicious use of these technologies increases significantly. Therefore, it is necessary to develop and implement tools capable of detecting deepfakes and protecting users within the virtual environment.

The rapid advancement of virtual reality (VR) technology has amplified the need for robust and reliable audio deepfake detection to maintain the authenticity and integrity of virtual interactions. However, current state-of-the-art models are usually trained using the Wav2vec 2.0 model on the ASVSpoof dataset [10], which might display overconfidence in their predictions, leading to vulnerabilities and diminished performance in real-world scenarios. Overconfident models struggle to generalize effectively [11], making them less capable against novel or slightly altered deepfake attacks, posing significant risks in VR applications where user trust and experience are crucial.

In this work, we explore the use of data augmentation and Mixup techniques to increase the diversity of training data and improve the generalization of deepfake audio detection models. Mixup creates new training examples by randomly combining pairs of existing examples, resulting in more varied and comprehensive training data. By integrating this technique, we aim to create a more resilient detection model that maintains high performance even when faced with diverse and sophisticated deepfake audio. Data augmentation uses a function that alters the audio creating a different sample, this process is done given a probability to each data augmentation function. Our work is organized as follows: Section 2 presents related work, methods are detailed in Section 3, results are presented in Section 4, and finally, Section 5 presents our conclusions.

2 RELATED WORKS

Although detecting deepfakes in immersive realities (VR) is an emerging field, recent studies in audio forgery detection provide a valuable foundation. In "Audio Deepfake Detection with Self-Supervised WavLM and Multi-Fusion Attentive Classifier" [4], the self-supervised WavLM model was combined with a Multi-Fusion Attentive (MFA) classifier to detect audio deepfakes, achieving state-of-the-art results on the ASVspoof 2021 DF dataset [12]. Similarly, "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks" [6] introduced the AASIST model, which uses graph attention networks to capture spectro-temporal artifacts, demonstrating remarkable efficacy on the ASVspoof 2019 LA dataset [10]. Additionally, "Experimental Study: Enhancing Voice Spoofing Detection Models with wav2vec 2.0" [7] explored the use of the wav2vec 2.0 model as an audio feature extractor,

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

fine-tuning the Transformer layers to improve forgery detection, achieving an equal error rate (EER) of 0.12% on the ASVspoof 2019 LA dataset, representing the current state of the art.

This research stands out for its use of the mix-up technique, which will be explained in detail in Section 3, to improve the model's robustness and its combination with classical data augmentation techniques.

3 METHODOLOGY

The ASV Spoof 2019 dataset is a consolidated dataset in the field of deepfake detection and is used as a benchmark by many researchers. It includes predefined sets of training, validation, and evaluation. The training and validation datasets contain fake audios generated using one group of text-to-speech (TTS) algorithms, while the evaluation dataset consists of audios produced by a separate, non-overlapping group of algorithms. This setup aims to validate the model's performance against fake audios generated by algorithms not seen during training.

The project commenced with the implementation of an audio pre-processing procedure. In this context, all audio files were standardized to a fixed duration of 3 seconds and converted to mono format, if necessary, by calculating the mean of the channels. Additionally, for audio files with durations shorter than the established threshold, zero-padding technique was employed, while for those exceeding the threshold, truncation technique was applied.

For conducting experiments, we fine-tuned the pretrained wav2vec-xls-r-300m[1] model using the predefined dataset splits. The experiments were made using Adam optimizer with a standard fixed learning rate of 3e-5, a batch size of 48 samples and 30 epochs. To ensure result reproducibility, we used the PyTorch seed generator with a seed value of 42. The experiments were conducted on a V100 GPU with 32GB of VRAM.

To enhance the robustness and generalization capabilities of our deep fake audio detection model, we employed a series of data augmentation techniques. These augmentations introduce variability to the dataset, thereby improving the model's ability to detect whether an audio sample is bonafide or fake.

The augmentation pipeline consists of the following steps:

- **Gaussian Noise Addition**: This augmentation simulates real-world scenarios by adding Gaussian noise to the audio samples, thereby mimicking background noise conditions. It helps the model become more resilient to variations in audio quality and environmental noise. In our implementation, Gaussian noise is added with a probability $p = 0.5$.
- **Low-Pass Filtering**: A low-pass filter is applied to allow frequencies below a certain threshold to pass while attenuating higher frequencies. This mimics the effect of audio compression or transmission over media that cut off higher frequencies, thus preparing the model to handle such distortions. The low-pass filter is applied with a probability $p = 0.5$.
- **Time Masking**: This technique involves masking a random section of the audio sample, effectively silencing it for a short duration. Time masking forces the model to recognize essential features of the audio even when parts are missing, improving its robustness. Time masking is performed with a probability $p = 0.5$ and masks frequencies in the range of 150 Hz to 7500 Hz.

To improve the robustness of our wav2vec-xls-300m model in detecting deepfakes, we employed the mix up technique with parameters $\alpha = \beta = 0.5$. Mix up is a powerful augmentation strategy that generates new training samples by combining pairs of examples and their corresponding labels. This technique helps in regularizing the model by introducing a smooth distribution of the input space, which in turn enhances the model's generalization capabilities. Given two input samples, one real and one fake, (x_i, y_i) and (x_j, y_j), the mix up technique creates a new training sample $(x_{\text{new}}, y_{\text{new}})$ as follows:

$$x_{\text{new}} = \lambda x_i + (1 - \lambda) x_j$$
$$y_{\text{new}} = \lambda y_i + (1 - \lambda) y_j$$

where λ is a mixing parameter drawn from a Beta distribution $\text{Beta}(\alpha, \beta)$. For our experiments, we set $\alpha = \beta = 0.5$, which provides a balanced mix of the two samples. We use this technique on every train steps, at sample level, meaning that all samples seeing during the train are new generated ones, with different degrees of real and fake labels.It is expected that this technique improves the robustness of the model against different kinds of spoof audios samples.

To evaluate the performance of our models, we employed the tandem detection cost function (t-DCF) as the primary metric, following the ASVspoof 2019 evaluation plan [1]. The t-DCF assesses the impact of both spoofing and countermeasures on the reliability of an automatic speaker verification (ASV) system. It is calculated as:

$$\text{t-DCF}_{\text{norm}}^{\text{min}} = \min_{s} \left\{ \beta P_{\text{miss}}^{\text{cm}}(s) + P_{\text{fa}}^{\text{cm}}(s) \right\}, \tag{1}$$

where β depends on application parameters (priors, costs) and ASV performance (miss, false alarm, and spoof miss rates). $P_{\text{miss}}^{\text{cm}}(s)$ and $P_{\text{fa}}^{\text{cm}}(s)$ are the countermeasure (CM) miss and false alarm rates at threshold s. The minimum in (1) is taken over all thresholds on given data (development or evaluation) using the ground truth to determine the optimal threshold.

In addition to t-DCF, we also used the equal error rate (EER) as a secondary metric. The EER corresponds to the operating point where the miss rate and the false alarm rate are equal. This metric was the primary metric in previous editions of ASVspoof and remains useful for assessing fake audio detection in the absence of an ASV system.

The ASV scores used in the computation of the t-DCF were provided by the organizers of the ASVspoof 2019 challenge, ensuring robustness and comparability of our results [1].

The experiments proposed by this study is presented in the following:

(1) Baseline;
(2) Data augmentation;
(3) Mix up;
(4) Data augmentation followed by mix up;
(5) Mix up (extended);
(6) Mix up and data augmentation;
(7) Mix up and data augmentation (extended);

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m

In our first experiment, we performed the fine-tuning of the pretrained model with no additional techniques as our baseline. In experiments 2 and 3, we investigate the use of data augmentation and mix up alone. In experiment 4, we tested the cascaded training, where the data augmentation training is followed by more 10 epochs of training using the mix up technique. In experiment 5, we extended the number of training epochs for mix up by an additional 10 epochs compared to experiment 3, for further comparison with experiment 4. In experiment 6 we tested using data augmentation along with mix up, and in experiment 7 we extended the number of training epochs of experiment 6 by an additional 10 epochs, similar to the extension in experiment 5.

4 RESULTS AND DISCUSSION

In this section, we discuss the impacts of the proposed techniques in terms of effectiveness for real-world applications of deepfake detection. Our findings are presented in Table 1.

Exp.	DA	Mixup	Cascade	Epochs	EER	min-tDCF
1				30	9.97%	0.22
2	X			30	3.11%	**0.06**
3		X		30	8.11%	0.20
4			X	30	**2.57**%	0.09
5		X		40	10.67%	0.30
6	X	X		30	4.55%	0.14
7	X	X		40	2.80%	0.08

Table 1: Experimental results for deepfake detection using different techniques. DA: data augmentation; EER: Equal Error Rate; min-tDCF: Minimum Detection Cost Function.

The experiment without additional techniques in the training process served as the baseline for comparing the effectiveness of the data augmentation and mixup techniques investigated. This baseline resulted in an EER of 9.97% and a min-tDCF of 0.22, indicating low effectiveness in detecting deepfakes, which could compromise user experience in virtual reality environments.

While the baseline did not achieved optimal results, the experiment with simple data augmentation techniques achieved an EER of 3.11% and a min-tDCF of 0.06, representing a significant improvement over the baseline.

The experiment using mix up alone achieved an EER of 8.11% and a min-tDCF of 0.20, slightly outperforming the baseline in terms of EER but still showing a relatively high error rate, which could be critical for sensitive applications. This results indicates that while Mixup is beneficial to improve robustness, the use of the technique alone might not lead to optimal results.

To further explore these techniques, we conducted experiments combining mixup and data augmentation and also a cascaded training experiment. Since both mixup and data augmentation can improve robustness, we are interested in validate if the combination of the two can improve the results. The experiment combining data augmentation and mix up altogether achieved an EER of 4.55% and a min-tDCF of 0.14, showing improvement over the baseline and experiment 2, with competitive performance for general use, but the best performance was achieved with the cascaded training experiment, where the model was initially trained with data augmentation for 30 epochs and then with mix up for 10 epochs more, resulting in an EER of 2.57% and a min-tDCF of 0.09. This result surpassed both the baseline and the other techniques tested, indicating optimal results for robust deepfake detection.

To validate that cascaded training indeed represents an improvement and it was not just because it was trained for more time, we conducted an additional experiment using data augmentation and mix up for the same number of epochs, resulting in an EER of 2.80% and a min-tDCF of 0.08, which also outperformed the baseline and demonstrated good effectiveness in resource-constrained environments, but with a slightly worse EER with compared to the cascade experiment.

These experiments demonstrate that while data augmentation alone achieves good results, combining it with other techniques such as mix up can further improve performance. The combination of mix up with data augmentation resulted in a 3.8 times improvement over the baseline in terms of EER, while cascaded training resulted in a 3.5 times improvement. Both approaches show promising for applications in virtual reality, where efficiency and accuracy are crucial for an immersive and secure user experience.

5 CONCLUSION

In this study, we demonstrated that the use of data augmentation and mix up techniques combined improves the fine-tuning results of a wav2vec model for deepfake classification. Our study also indicates the potential of a cascade training, where the mixup technique is trained after the data augmentation, achieving a 2.57% error rate and a 0.09 min-tDCF against the ASVSpoof 2019 evaluation dataset. Besides the promising results, we believe that it is necessary to conduct tests against data from other domains to validate its effectiveness as well as to investigate its confidence in a real-world VR scenario.

REFERENCES

[1] 2019. ASVspoof 2019: The Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf. [Online].

[2] Fatih Arslan. 2023. Deepfake Technology: A Criminological Literature Review. *The Sakarya Journal of Law (The SJL)* 11, 1 (2023), 701–720.

[3] Rebecca A. Delfino. 2023. Deepfakes em julgamento: uma chamada para expandir o papel de controle do juiz de julgamento para proteger os processos legais contra falsificação tecnológica. *Hastings Law Journal* 74 (2023), 293. https://repository.uclawsf.edu/hastings_law_journal/vol74/iss2/3

[4] Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. 2023. Audio Deepfake Detection with Self-Supervised WavLM and Multi-Fusion Attentive Classifier. *arXiv preprint arXiv:2312.08089* (2023). https://doi.org/10.48550/arXiv.2312.08089 arXiv:2312.08089 Submitted on 13 Dec 2023 (v1), last revised 10 Jan 2024 (this version, v2).

[5] Miao Hu, Xianzhuo Luo, Jiawen Chen, Young Choon Lee, Yipeng Zhou, and Di Wu. 2021. Virtual Reality: A Survey of Enabling Technologies and its Applications in IoT. *arXiv preprint arXiv:2103.06472* (2021). https://doi.org/10.48550/arXiv.2103.06472 arXiv:2103.06472 Submitted on 11 Mar 2021.

[6] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2021. AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks. *arXiv preprint arXiv:2110.01200* (2021). https://doi.org/10.48550/arXiv.2110.01200 arXiv:2110.01200

[7] Taein Kang, Soyul Han, Sunmook Choi, Jaejin Seo, Sanghyeok Chung, Seungeun Lee, Seungsang Oh, and Il-Youp Kwak. 2024. Experimental Study: Enhancing Voice Spoofing Detection Models with wav2vec 2.0. *arXiv preprint arXiv:2402.17127* (2024). https://doi.org/10.48550/arXiv.2402.17127 arXiv:2402.17127

[8] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. 2024. Deepfake Generation and Detection: A Benchmark and Survey. *arXiv*

preprint arXiv:2403.17881 (2024). https://doi.org/10.48550/arXiv.2403.17881 arXiv:2403.17881 Submitted on 26 Mar 2024 (v1), last revised 16 May 2024 (this version, v4).

[9] Tomasz Walczyna and Zbigniew Piotrowski. 2023. Overview of voice conversion methods based on deep learning. *Applied Sciences* 13, 5 (2023), 3100.

[10] X. Wang, J. Yamagishi, and et al. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language (CSL)* 64 (2020), 101114.

[11] Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Yuankun Xie, Yukun Liu, Xiaopeng Wang, Xuefei Liu, Yongwei Li, Jianhua Tao, Yi Lu, Xin Qi, and Shuchen Shi. 2024. Generalized Fake Audio Detection via Deep Stable Learning. *arXiv preprint arXiv:2406.03237* (2024). https://doi.org/10.48550/arXiv.2406.03237 arXiv:2406.03237 accepted by INTERSPEECH2024.

[12] Junichi Yamagishi, Xuechen Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuenan Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and et al. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop - Automatic Speaker Verification and Spoofing Countermeasures Challenge*.