IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Data augmentation techniques for Physical Access in voice anti-spoofing

Jose Carlos Sanchez, Antonio M. Peinado, Angel M. Gomez

Dpt. Signal Processing, Telematics and Communications - CITIC, Universidad de Granada, Spain

svjosecarlos@ugr.es, amp@ugr.es, amgg@ugr.es

Abstract

In this paper, we explore how data augmentation (DA) techniques can improve spoofed audio detection. Specifically, we will focus on replay attacks, where a genuine voice is surreptitiously captured and then played back through a loudspeaker to the voice biometric system. We propose several approaches to handle with reverberation variability, different types of additive noise, and unseen spoofing attacks, which have all been proven to reduce the performance of countermeasure systems. In order to test the effectiveness and generalization capability of these DA techniques, out-of-domain experiments are carried out on the PA ASVspoof 2021 dataset as well as on the ASVspoof 2019 Real corpus, employing a LCNN classifier fed with STFT features and trained over an augmented version of the ASVspoof 2019 corpus. Four DA methodologies are explored: time masking, noise addition, Room Impulse Response filtering and data mixup. The experimental results show that meaningful improvements can be achieved when the DA procedures are suitably selected.

Index Terms: anti-spoofing, data augmentation, physical access

1. Introduction

Automatic Speaker Verification (ASV) aims to verify the claimed identity of users in biometric systems by analyzing their voice patterns. As these systems enable secure access to sensitive information, ASV is widely utilized in practical scenarios and plays a significant role in our daily lives, such as bank authentication and voice assistants. Despite recent advancements [1], ASV systems remain susceptible to malicious attacks, which can be classified into Logical Access (LA) attacks, such as text-to-speech (TTS) and voice conversion (VC), and Physical Access (PA), including replay and impersonation attacks [2]. This paper addresses the issue of replay attacks. In this case, the attacker does not need any technical knowledge since the attack can be simply carried out by recording the original user's voice and presenting it to the ASV system via a loudspeaker [3].

Recent studies have focused on the development of anti-spoofing strategies, also referred to as countermeasures (CM), with the goal of detecting whether the audio voice presented to the ASV system is genuine (bonafide) or whether it originates from a spoofing attack (spoofed). Many of these CM systems rely on deep-learning based approaches, as they have been proven to be the most effective [4, 5, 6, 7]. While enhancing network architectures can lead to improved performance, these methods often suffer from overfitting and poor generalization capability mainly due to the lack of suitable training data [8, 9]. These issues are common to many classification systems [10, 11, 12] and have been commonly tackled by means of data augmentation (DA) and regularization techniques [13, 14, 15].

The aim of DA is to increase the diversity and amount of training data without collecting new samples [16, 17]. By applying various transformations, new data is created enabling the model to generalise to a wider range of scenarios. This process helps to mitigate overfitting, improve model robustness and enhance overall performance, particularly when the available dataset is limited.

In the case of audio systems, data augmentation can be performed either directly on signal samples (waveform) or on time-frequency domain features, including spectrogram, mel-spectrogram, and mel-frequency cepstral coefficients. For the waveform, there are several DA methods that are known for preserving the relevant information in speech audio as noise addition, speaking speed changing and perturbation [18], time shifting, among others.

In the spectral domain, methods intended for Automatic Speech Recognition (ASR) as in [8], have introduced strategies such as time warping, frequency masking, and time masking, which involve altering specific frequency bands or time frames to make the model more robust by forcing it to learn from incomplete or partially obscured data.

Additionally, since time-frequency features (spectrograms) can be treated as two-dimensional images, some DA strategies from computer vision can also be applied. An example is the Mixed Sample Data Augmentation (MSDA) techniques, also known as mixup [19], which combines the spectral features of two audios and their corresponding labels by varying a random parameter.

In recent years, the scientific community has explored the effectiveness of DA techniques to improve the generalization of voice anti-spoofing models. While significant advancements have been achieved in Logical Access (LA) and DeepFake (DF) detection [20, 21, 22], Physical Access (PA) has seen limited progress, with only small improvements emerging from small-scale studies [23]. In this paper, we address the existing knowledge gap in PA and determine which DA techniques yield the best results for anti-spoofing models, especially in the context of replay attacks. The techniques used in this study include time masking and noise addition, as well as more innovative methods to provide greater acoustic diversity. Specifically, we consider filtering with a new Room Impulse Response (RIR) to introduce varying levels of reverberation, and the mixup method for generating new spoofed and bonafide spectrogram instances.

In order to check the generalization capability of the studied DA techniques, out-of-domain tests are conducted using a Lightweight Convolutional Neural Network (LCNN) with log-magnitude STFT input features, which is a widely known baseline model for spoofing detection. The results, measured in

10.21437/IberSPEECH.2024-1

terms of Equal Error Rate (EER), are promising and indicate significant progress in Physical Access (PA) attack detection.

The rest of this paper is organized as follows. The next section presents the databases used for training, development, and evaluation of PA countermeasures, along the LCNN baseline model. In Section 3, we describe the DA techniques employed. Then, Section 4 shows the experiments conducted to check the effectiveness of these techniques and their results. Finally, we discuss and summarize our research in Section 5.

2. Databases and experimental setup

In this section, we describe the datasets used, an overview of our model architecture to assess the performance of the tested techniques and the evaluation metrics employed.

2.1. Speech datasets

In this study, we will use the training and validation subsets of ASVspoof 2019 PA [24] to train our model, and then test the effectiveness of the different DA methods considered against ASVspoof 2021 [1] and ASVspoof 2019 Real [25] evaluation sets.

The training and validation subsets of ASVspoof 2019 PA contain genuine voice recordings and replay attack samples generated through simulation with utterances taken from the VCTK corpus [26]. The simulation involves filtering with pre-established RIRs in order to simulate room's acoustics, and the application of the non-linearities corresponding to the loudspeakers used for replay. No environmental noise is considered. The number of speakers and audio samples in each subset is presented in Table 1.

Table 1: *Structure of ASVspoof 2019 PA data corpus.*

	Speakers		**Utterances**	
Subset	**Male**	**Female**	**Bonafide**	**Spoofed**
Training	8	12	5,400	48,600
Development	8	12	5,400	24,300
Evaluation	21	27	18,090	116,640
Total	37	51	28,890	189,540

The ASVspoof 2019 Real evaluation dataset was released with the aim of providing an extra, small collection of real spoofed recordings (captured and replayed in three distinct acoustic conditions) [25]. Unlike the simulated database, this dataset features authentic acoustic distortions, including noise and reverberation. It consists of 2,700 audio files in total, with 540 bonafide audios and 2,160 spoofed ones.

Unlike ASVspoof 2019 PA, the ASVspoof 2021 data [1] was created under a rigorously controlled setup across a diverse range of real acoustic environments, featuring various levels of reverberation and additive noise. Furthermore, recordings are performed using different playback and recording devices, making it more realistic for a replay attack scenario than its 2019 counterpart. The dataset consists solely of an evaluation partition, totaling 943,110 audio files, with 126,630 bonafide voices and 816,480 spoofed.

2.2. Classification model

The classification model used in our experiments is an alternative implementation of the LCNN architecture proposed in [27], which is fed with the log-magnitude spectra obtained via Short-Time Fourier Transform (STFT). In particular, the STFT provides a spectrogram $S[t, f]$ consisting of $N = 400$ frames and $M = 256$ frequency bins, derived from 512 FFT points and excluding $f = 0$. The transform is performed with a Blackman window of 25 ms frames and 10 ms shift, which, given a sampling rate of 16,000 Hz, restricts the audio to 4-second segments.

This network consists of five layers, each incorporating 2D convolutions followed by max pooling. Batch normalization is also employed to improve the stability and convergence of gradient descent. As illustrated in Table 2, the output is then fed into a Fully-Connected layer (FC1) to produce a 64-dimensional utterance-level spoofing identity vector. This vector is then passed through a second Fully-Connected layer (FC2), which generates a final score vector with two components, indicating whether the utterance is bonafide or spoofed.

Table 2: *The architecture of our LCNN model.*

Layer	**Type**	**Filter/Stride**	**Output Channels**
Layer 1	Conv2D	5x5/1x1	8
	BatchNorm2D	-	8
	MaxPool	2x2/2x2	8
Layer 3	Conv2D	1x1/1x1	8
	BatchNorm2D	-	8
	Conv2D	3x3/1x1	16
	MaxPool	2x2/2x2	16
Layer 4	Conv2D	1x1/1x1	16
	BatchNorm2D	-	16
	Conv2D	3x3/1x1	16
	MaxPool	2x2/2x2	16
Layer 5	Conv2D	1x1/1x1	16
	BatchNorm2D	-	16
	Conv2D	3x3/1x1	16
	MaxPool	2x2/2x2	16
-	FC1	-	64
-	BatchNorm1D	-	64
-	FC2	-	2

It is worth noticing that the two-dimensional convolutions are implemented using Max-Feature-Map (MFM) operations [27], which result in a reduction in the total number of parameters (lightweight). Additionally, a dropout rate of 0.7 was applied before the first Fully-Connected layer to prevent overfitting.

2.3. Evaluation metrics

We use the Equal Error Rate (EER) [28] as the metric to assess the model's overall accuracy. This metric is particularly useful in anti-spoofing systems because it indicates the trade-off between incorrectly accepting spoofed audio (false positives) and failing to detect bonafide audio (false negatives). It is important for the EER to be as low as possible, as it demonstrates that the model is achieving the most effective balance between these two types of errors.

3. Data augmentation techniques

This section details the DA techniques considered during the training of the proposed model, consisting of time and frequency masking, noise addition, Room Impulse Response filtering, and data mixup.

3.1. Time and frequency masking

Time masking is a method from the SpecAugment framework in [8], which enhances the model robustness by masking contiguous time segments within the spectrogram $S[t, f]$.

In time masking, a segment of Δt consecutive time steps, spanning from $[t_0, t_0 + \Delta t)$, is masked. The duration Δt is sampled from a uniform distribution in an interval $[0, T]$, while the starting point t_0 is chosen from the range $[0, N - \Delta t)$, where N is the total length of the time dimension in the spectrogram. This approach forces the model to employ variable time intervals rather than relying on specific time-localized patterns, which can be particularly advantageous for detecting spoofed speech.

Therefore, the masked spectrogram $\tilde{S}[t, f]$, spoofed or bonafide, is then defined as,

$$\tilde{S}[t, f] = \begin{cases} S[t, f] & \text{if} \quad t < t_0 \text{ or } t \geq t_0 + \Delta t, \\ 0 & \text{if} \quad t_0 \leq t < t_0 + \Delta t. \end{cases} \tag{1}$$

Frequency masking is similar to time masking, but applied to the frequency domain. In this case, a segment of Δf consecutive frequency bins, $[f_0, f_0 + \Delta f)$, is masked. Analogous to time masking, the duration Δf is sampled from a uniform distribution in $[0, F]$, and f_0 is chosen from $[0, M - \Delta f)$, where M is the total number of frequency bins.

The masked spectrogram $\tilde{S}[t, f]$ with frequency masking is given by,

$$\tilde{S}[t, f] = \begin{cases} S[t, f] & \text{if} \quad f < f_0 \text{ or } f \geq f_0 + \Delta f, \\ 0 & \text{if} \quad f_0 \leq f < f_0 + \Delta f. \end{cases} \tag{2}$$

This technique introduces variability in the spectral domain by simulating different frequency distortions.

3.2. Additive noise

Unlike time masking and frequency masking techniques, additive noise is applied directly to the temporal samples of the audio. This method aims to artificially introduce noise into the voice sequences to simulate various real-world conditions and enhance the model's robustness.

While white gaussian noise is traditionally used for augmenting audio samples [29], other noise types, such as background conversations or environmental sounds, can also be introduced to diversify the training data. Since our focus is on PA and, specifically, replay attacks, we will use our own database of realistic noises, such as car, bus noises, crowd chatter, street environment, and similar sounds that might be encountered in practice. Our DA technique involves, for each audio sequence $x[n]$ in a training batch, the random selection of a noise excerpt $v[n]$ from our noise database and a signal-to-noise ratio value (SNR, in dB) chosen from a uniform distribution between $[SNR_{min}, SNR_{min} + 10]$.

Consequently, the noisy audio signal $\tilde{x}[n]$ is

$$\tilde{x}[n] = x[n] + \alpha \cdot v[n + \tau], \tag{3}$$

where α is the scaling factor applied to the noise signal to achieve the desired SNR and can be calculated as

$$\alpha = \sqrt{\frac{P_x}{P_v \cdot SNR}}, \tag{4}$$

with P_x and P_v denoting the power of the audio and noise signals, respectively.

3.3. Room Impulse Response filtering

Like additive noise, filtering with a new RIR is a data augmentation technique applied directly to the temporal samples of spoofed and bonafide audios. Our innovative approach in PA aims to introduce acoustic variability to the training data, specifically in terms of reverberation time (RT60).

To achieve this, we use the Roomsimove simulation program [30], which generates RIRs based on the image-source method [31]. The program takes as input the acoustic parameters characterizing each room. To ensure realistic values, we use data from The Ace Challenge 2015 study [32], which provides experimental RT60 measurements for each frequency octave (125 Hz, 250 Hz, ..., 8000 Hz) in seven different rooms: two offices, a building lobby, two meeting rooms, and two conference rooms. Table 3 shows the averaged RT60 across frequencies for each case.

Table 3: *Size and averaged RT60 for each room in [32].*

Room name	Size (m)	Av. RT60 (s)
Office1	4.83 x 3.32 x 2.95	0.380
Office2	5.10 x 3.22 x 2.94	0.430
BuildingLobby	5.13 x 4.47 x 3.18	0.715
MeetingRoom1	5.11 x 6.61 x 2.95	0.480
MeetingRoom2	9.07 x 10.32 x 2.63	0.415
LectureRoom1	9.73 x 6.93 x 3.00	0.675
LectureRoom2	9.29 x 13.56 x 2.94	1.200

The radial distance between the sound source and the receiver is randomly selected, following a uniform distribution within three reverberation categories (low L, neutral N, and high H) as shown in Table 4. These parameters (room size, RT60, source and receiver positions, etc.) are used as inputs to the Roomsimove software [30], which generates the corresponding impulse responses. In total, 4050 unique RIRs were simulated.

Table 4: *Distance categories for the RIRs simulation.*

	L	N	H
Distance (cm)	[10, 50]	[50, 100]	[100, 150]

Next, the DA technique is performed by carrying out the discrete-time convolution operation between the audio, $x[n]$, and an impulse response, $h[n]$, randomly selected from the RIR dataset generated.

3.4. Data mixup

Mixup, as detailed in [19], is a recent DA technique for creating new synthetic features from the existing data in a training batch. Thus, this approach performs a linear interpolation between two random spectrograms and their labels within a batch. Specifically, given two spectrograms and their corresponding labels, $(S_i[t, f], y_i)$ and $(S_j[t, f], y_j)$, mixup generates a new spectrogram, $\tilde{S}[t, f]$ and its label, $\tilde{y}$, as follows,

$$\begin{aligned} \tilde{S}[t, f] &= \lambda S_i[t, f] + (1 - \lambda) S_j[t, f], \\ \tilde{y} &= \lambda y_i + (1 - \lambda) y_j, \end{aligned} \tag{5}$$

where λ is a mixing coefficient drawn from a Beta distribution, $\lambda \sim \text{Beta}(\alpha, \alpha)$, with $\alpha \in (0, \infty)$.

Spectrograms are randomly selected within the batch, and the parameter α can vary freely. A smaller value of α results

Table 5: *EER results for the different PA evaluation datasets after applying varied configurations of the DA techniques presented. Values representing a relative improvement equal or higher than 5% are highlighted in bold.*

		Training							
		2019 PA							
		No DA	**Time masking**	**Frequency masking**	**Additive noise**		**RIR filtering**	**Mixup**	
					30 dB	**15 dB**		$\boldsymbol{\alpha = 0.2}$	$\boldsymbol{\alpha = 0.1}$
Eval.	**2019 PA**	3.03%	**2.58%**	5.26%	11.12%	6.23%	3.64%	**2.55%**	**2.55%**
	2019 Real	30.12%	**26.78%**	29.44%	44.68%	36.91%	**26.67%**	**28.46%**	28.64%
	2021 PA	44.06%	43.30%	43.98%	44.73%	**36.33%**	42.50%	43.26%	43.69%

Table 6: *EER results for the different PA evaluation datasets using the combination of our best DA techniques.*

		Training			
		2019 PA			
		No DA	**Time masking + Freq. masking**	**Additive noise 15 dB + RIR filtering**	**RIR filtering + Time masking**
Eval.	**2019 PA**	3.03%	4.68%	5.12%	5.55%
	2019 Real	30.12%	36.78%	42.10%	30.52%
	2021 PA	44.06%	44.16%	**40.04%**	43.95%

in greater differentiation between the two classes in the newly created features, which is advantageous for generating clear examples of both spoofed and bonafide. Conversely, a very large value of α may result in an overmixing of features, potentially causing the new examples to become too similar and reducing the model's ability to distinguish between the two classes.

When an appropriate value for α is chosen, there is a balance between differentiation and overfitting and the model can generalize to new, unseen audio data.

4. Experiments

For each DA experiment, we performed three training runs and evaluate the databases across the three resulting models, ultimately reporting the corresponding averaged EER. In total, 33 training runs and 99 evaluation tests were conducted. In every training step, when a DA technique is applied to a batch of data, a new same-size batch is recreated, thereby doubling the size of the effective batch during training, or tripling it when pairs of DA techniques are combined.

We use the standard Adam optimizer for minimizing the weighted cross-entropy (WCE) loss function, with a learning rate of $3 \cdot 10^{-4}$, a weight decay of 10^{-4}, and an initial batch size of 144. All training processes use a maximum of 100 epochs and an early-stopping criterion of 15, except for the mixup technique, which requires the full 100 epochs to ensure model convergence. The total number of parameters in the neural network is 832,946.

4.1. Results

Table 5 shows the averaged EER for each DA technique evaluated. Values showing a relative improvement equal or higher than 5% when comparing to the no-DA baseline, are highlighted in bold.

For the experiments involving time masking and frequency masking, the selected width parameters (after some preliminary experiments) are $T = 80$ and $F = 20$, respectively. Noise addition DA was assessed at two different minimum SNR levels: a lower noise level with $SNR_{min} = 30$ dB, and a higher noise level with $SNR_{min} = 15$ dB. Additionally, the mixup DA approach was explored using two α values: $\alpha = 0.2$ and $\alpha = 0.1$, where a smaller α indicates greater class differentiation in the new generated data.

As shown, the additive noise data augmentation (DA) technique with noise levels between 30 dB and 40 dB provides the best results on the ASVspoof 2021 PA evaluation set, achieving a 36.33% EER (44.06% without DA). This improvement is mainly due to the fact that the ASVspoof 2021 PA dataset includes various levels and types of additive noise, making this method particularly effective for DA when testing on this database.

For the ASVspoof 2019 Real dataset, the best DA results are obtained using RIR filtering, resulting in a 26.67% EER compared to the original 30.12%. In this case, the ASVspoof 2019 Real data encompasses different real acoustic conditions, so applying new reverberation environments to the training data introduces new acoustic features that aid the LCNN's learning process. Additionally, the EER values obtained for time masking and mixup with $\alpha = 0.2$ are promising. These techniques introduce greater variability during training, preventing the neural network from learning specific temporal patterns that may be present in the spectrograms of the training data.

We attempted to combine the most effective DA techniques in hopes of improving results, but this was not the case, as the best improvement achieved was an EER of 40.04% for the ASVspoof 2021 evaluation dataset. This limitation is likely due to reaching the network's generalization capacity, constrained by its relatively small number of parameters.

5. Conclusions

In this paper, we have explored the enhancement of replay attack detection systems, associated to voice biometrics, by means of DA techniques. We evaluated four methodologies across various datasets to improve CM system robustness.

The results obtained show that DA can effectively improve the detection performance and, also, that the choice of a suitable DA technique depends on the context where the system is to be used (i.e., the characteristics of the test data). This suggests that adapting the detector model to the test context would be an interesting approach for future research.

6. Acknowledgements

This paper is part of the project PID2022-138711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

7. References

[1] J. Yamagishi, X. Wang, M. Todisco *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *ArXiv*, vol. abs/2109.00537, 2021.

[2] Z. Wu, P. De Leon, C. Demiroglu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 4, pp. 768 – 783, Apr. 2016.

[3] A. Khan, K. M. Malik, J. Ryan *et al.*, "Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward," *ArXiv*, vol. abs/2210.00417, 2022.

[4] A. G. Alanís, A. M. Peinado, J. A. González *et al.*, "A light convolutional gru-rnn deep feature extractor for asv spoofing detection," in *Interspeech*, 2019.

[5] A. Gomez-Alanis, A. M. Peinado, J. A. González *et al.*, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1985–1999, 2019.

[6] A. Tomilov, A. F. Svishchev, M. Volkova *et al.*, "Stc antispoofing systems for the asvspoof2021 challenge," *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[7] A. Gomez-Alanis, J. A. González-López, and A. M. Peinado, "Ganba: Generative adversarial network for biometric anti-spoofing," *Applied Sciences*, 2022.

[8] D. S. Park, W. Chan, Y. Zhang *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:121321299

[9] C.-C. Chiu, T. N. Sainath, Y. Wu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:206742954

[10] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:88499738

[11] C. F. G. dos Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1 – 25, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:245837898

[12] S. Salman and X. Liu, "Overfitting mechanism and avoidance in deep neural networks," *ArXiv*, vol. abs/1901.06566, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:58981513

[13] A. Sapru, "Using data augmentation and consistency regularization to improve semi-supervised speech recognition," in *Interspeech*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252337156

[14] Y. Zhou, C. Xiong, and R. Socher, "Improved regularization techniques for end-to-end speech recognition," *ArXiv*, vol. abs/1712.07108, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3483379

[15] D. Oneaţă and H. Cucu, "Improving multimodal speech recognition by data augmentation and speech representations," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4578–4587, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248427206

[16] S. Yang, W.-T. Xiao, M. Zhang *et al.*, "Image data augmentation for deep learning: A survey," *ArXiv*, vol. abs/2204.08610, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248240105

[17] K. Alomar, H. I. Aysel, and X. Cai, "Data augmentation in classification and segmentation: A survey and new strategies," *Journal of Imaging*, vol. 9, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257006426

[18] T. Ko, V. Peddinti, D. Povey *et al.*, "Audio augmentation for speech recognition," in *Interspeech*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:7360763

[19] H. Zhang, M. Cissé, Y. Dauphin *et al.*, "mixup: Beyond empirical risk minimization," *ArXiv*, vol. abs/1710.09412, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3162051

[20] H. Tak, M. R. Kamble, J. Patino *et al.*, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6382–6386, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:243847663

[21] A. Cohen, I. Rimon, E. Aflalo *et al.*, "A study on data augmentation in voice anti-spoofing," *Speech Commun.*, vol. 141, pp. 56–67, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239050551

[22] E. Rosello, A. Gomez-Alanis, A. M. Gómez *et al.*, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Interspeech*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260910522

[23] R. K. Das, "Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: Asvspoof 2021," *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:240510603

[24] M. Todisco, X. Wang, V. Vestman *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech*, 2019.

[25] M. Todisco, X. Wang, J. Yamagishi *et al.* (2019) Real PA database. Slides used during Interspeech 2019. [Online]. Available: https://www.asvspoof.org/interspeech2019_slides.pdf

[26] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:213060286

[27] X. Wu, R. He, Z. Sun *et al.*, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 2884–2896, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:5351802

[28] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *ArXiv*, vol. abs/1304.2865, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:14392885

[29] S. Huq, P. Xi, R. Goubran *et al.*, "Data augmentation and deep learning in audio classification problems: Alignment between training and test environments," *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 140–146, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:267765012

[30] E. Vincent. (2008) Roomsimove. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

[31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1976. [Online]. Available: https://api.semanticscholar.org/CorpusID:10721495

[32] A. Outman. (2016) The Ace Challenge 2015. [Online]. Available: https://dx.doi.org/10.5072/FK2NV9G85X