RAWBOOST: A RAW DATA BOOSTING AND AUGMENTATION METHOD APPLIED TO AUTOMATIC SPEAKER VERIFICATION ANTI-SPOOFING

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco and Nicholas Evans

EURECOM, Sophia Antipolis, France

ABSTRACT

This paper introduces RawBoost, a data boosting and augmentation method for the design of more reliable spoofing detection solutions which operate directly upon raw waveform inputs. While RawBoost requires no additional data sources, e.g. noise recordings or impulse responses and is data, application and model agnostic, it is designed for telephony scenarios. Based upon the combination of linear and non-linear convolutive noise, impulsive signal-dependent additive noise and stationary signal-independent additive noise, RawBoost models nuisance variability stemming from, e.g., encoding, transmission, microphones and amplifiers, and both linear and non-linear distortion. Experiments performed using the ASVspoof 2021 logical access database show that RawBoost improves the performance of a state-of-the-art raw end-to-end baseline system by 27% relative and is only outperformed by solutions that either depend on external data or that require additional intervention at the model level.

***Index Terms*—** spoofing, presentation attack detection, automatic speaker verification, data augmentation

1. INTRODUCTION

The recent ASVspoof 2021 challenge [1] addressed the problem of spoofing or presentation attack detection (PAD) in a logical access (LA) scenario in which both bona fide and spoofed utterances are encoded and transmitted across telephony networks. The task was to learn reliable detection solutions using only the training and development partitions of the ASVspoof 2019 LA datasets which are without such encoding and transmission effects. There is hence an interest in data augmentation techniques to compensate for the lack of in-domain training and development data [1, 2].

Still with a broad range of text-to-speech (TTS) and voice conversion (VC) spoofing attacks, the challenge maintained the focus of previous editions upon the development of generalised countermeasures that perform reliably not only in the face of spoofing attacks generated with TTS and VC algorithms different to those seen in training and development data, but also unseen encoding and transmission conditions. While numerous data augmentation solutions have been proposed, e.g. SpecAugment [3] and SpecMix [4], they are suitable only for spoofing detection models which operate on two-dimensional front-end representations. End-to-end (E2E) spoofing detection solutions which operate on raw waveforms rather than two-dimensional representations are now gaining popularity [5–11]. The usual data augmentation techniques are then incompatible; they cannot be applied directly to raw waveform inputs. We have hence explored data augmentation techniques that are compatible with our RawNet2 [6] and RawGAT-ST [8] systems.

The first author is supported by the VoicePersonae project funded by the French Agence Nationale de la Recherche (ANR) and the Japan Science and Technology Agency (JST).

We introduce RawBoost, a data boosting and augmentation technique that can be applied directly to raw audio. The aim is to improve spoofing detection reliability in the face of nuisance variation stemming from unknown encoding and transmission conditions which typify the LA or telephony scenario. RawBoost is based upon well-known signal processing techniques and is computationally inexpensive with regard to the cost of learning on augmented data. Furthermore, unlike WavAugment [12], a popular approach to data augmentation through pitch modification, band reject filtering, time dropping or the addition of reverberation or noise, techniques which can all be applied to raw waveforms, RawBoost operates without the need for any additional data sources, e.g., noise recordings or impulse responses, and neither requires any intervention at the model level.

RawBoost is data, application and model agnostic. While we report its application to improve spoofing detection performance, it might have application to other related classification tasks where similar nuisance variability is expected, e.g., automatic speaker verification or automatic speech recognition.

2. DATA AUGMENTATION

Data augmentation (DA) is commonly applied in many machine learning tasks to generate new samples from a source database, here utterances, to augment the pool of data available for training. The use of additional augmented data which exhibits variability not contained in the source data can help to reduce overfitting and bias, and hence to improve classification performance. Nowadays, DA is an integral component of modern machine learning pipelines and has been applied successfully in a host of different machine learning fields, such as image processing [13], speech recognition [14, 15] and speaker verification [16]. Recent work has also demonstrated its use in anti-spoofing [10, 11, 17–21]. A number of approaches to DA have been proposed in the literature, e.g., random cropping, rotation and mirroring for image-related tasks [22]; speed perturbation, pitch shifting, time stretching, random frequency filtering, reverberation, text-to-speech data augmentation and vocal tract length transformations for speech-related tasks [12, 23, 24].

Knowing that evaluation data would contain both bona fide and spoofed utterances treated with a variety of unknown codecs, participants of the ASVspoof 2021 LA challenge used, e.g., speed perturbation [14], SpecAugment [3] and codec augmentation [15] to help improve performance. SpecAugment, a form of spectral-domain augmentation, is applied to mask random intervals or bands of the spectrum and/or temporal frames during training but cannot be applied easily at the waveform level. Interest in raw E2E techniques for spoofing detection is currently growing [6–11, 25]. There is hence a need for DA techniques that account for the variability expected in LA or telephony scenarios and, in particular, techniques that can also be applied at the raw waveform level.

3. RAWBOOST DATA BOOSTING AND AUGMENTATION

RawBoost[1] is a data boosting and augmentation method which operates at the raw waveform level. Signal boosting approaches in machine learning have been gaining ground recently. Data boosting can encode prior knowledge about data or task-specific invariances, act as a regulariser to prevent overfitting, and can improve model robustness [26]. RawBoost employs established linear and non-linear signal processing techniques to boost or distort a set of utterances in a training dataset and/or augment a dataset with additional training utterances. RawBoost is illustrated in Fig. 1 and comprises the three independent processes described below.

3.1. Linear and non-linear convolutive noise

Any channel involving some form of encoding, compression-decompression and transmission introduces stationary convolutive distortion. Most such channels will also introduce non-linear disturbances which are themselves also subject to stationary convolutive distortion, but of different characteristics (see [27], Fig. 6). In order to improve robustness to such nuisance variation, we have explored the combination of multi-band filtering and Hammerstein systems [28]. Hammerstein systems are proven, popular models of non-linear, dynamic systems in which non-linear static and linear dynamic subsystems are separated into different orders [28]. While Hammerstein models estimate multi-band filters from the response of non-linear systems, here we use the same idea to generate signal distortions.

Multi-band filters are designed to generate convolutive noise using time domain notch filtering. They are applied to a single utterance at a time and with a set of N_{notch} notch filters, each with randomly chosen center frequencies f_c and filter widths Δf. A single finite impulse response (FIR) filter with randomly chosen gain value g_j^{cn} is then defined using a window-based filter design method [29], resulting in a filter with the desired frequency response using a randomly chosen number of filter coefficients N_{fir}. The higher the number of coefficients, the more abrupt the frequency response; filters with fewer coefficients will exhibit passband ripple or distortion in addition to smoother cut-in and cut-off responses. An example filter frequency response is illustrated in Fig. 2. It has $N_{\text{notch}} = 3$ notch filters, each with different center frequencies, stop-band widths and number of filter coefficients.

Hammerstein systems generate higher-order harmonics whereby a component f_0 in the input to a non-linear system is supplemented at the output by $N_{\text{f}} - 1$ new components at $2f_0, 3f_0, ..., N_{\text{f}}f_0$, leading to non-linear harmonic distortion. The frequency and amplitude of each higher-order harmonic are dependent upon those of the original component and the characteristics of the non-linear system. Convolutive noise y_{cn}, denoted ① in Fig. 1, is generated according to:

$$y_{\text{cn}}[n] = \sum_{j=1}^{N_{\text{f}}} g_j^{\text{cn}} \sum_{i=0}^{N_{\text{fir}_j}} b_{i_j} \cdot x^j[n-i] \tag{1}$$

where $x \in [-1, 1]^{l \times 1}$ denotes a raw waveform of l samples, $j \in [1, N_{\text{f}}]$ is the order of the (non-)linearity ($N_{\text{f}} = 1$ refers to the filter applied to the linear component x), b_{i_j} denotes the coefficients of the j^{th} multi-band filter.

[1] https://github.com/TakHemlata/RawBoost-antispoofing

3.2. Impulsive signal-dependent additive noise

Impulsive signal-dependent noise is commonly introduced through data-acquisition, resulting from, e.g., clipping, non-optimal device operation (microphones and amplifiers), synchronization and overflow issues, or as a result of insufficient computational power. It is typically orders of magnitude lower in amplitude than signal-independent noise [30]. We model such nuisance variability as non-stationary impulsive disturbances (see ② in Fig. 1) consisting of instantaneous or impulse-like amplitude variations. The disturbance z_{sd} is applied to a maximum of $P \leq l$ uniformly distributed samples $\{p_1, p_2, ..., p_P\}$ in x to obtain y_{sd} according to:

$$y_{\text{sd}}[n] = x[n] + z_{\text{sd}}[n] \tag{2}$$

where

$$z_{\text{sd}}[n] = \begin{cases} g^{\text{sd}} \cdot D_R\{-1,1\}[n] \cdot x[n], & \text{if } n = \{p_1, p_2, ..., p_P\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

is a *signal-dependent additive noise* component, $g^{\text{sd}} > 0$ is a simple gain parameter and where $D_R\{-1, 1\}[n]$ denotes P values randomly chosen from the distribution:

$$f_R(r) = \begin{cases} -log(r), & 0 < r \leq 1 \\ -log(-r), & -1 \leq r < 0 \end{cases} \tag{4}$$

For convenience, the maximum number of samples P is chosen relatively as $P_{\text{rel}} = P/l$.

3.3. Stationary signal-independent additive noise

The use of signal-independent, additive noise is one of the most popular forms of data augmentation and has been applied in a wide variety of applications, including speech recognition [31], speaker recognition [32], speech emotion recognition [33], as well as audio forgery [34] and spoofing detection [19]. Signal-independent additive noise can result from loose or poorly joined cable connections, transmission channels effects, electromagnetic interference or thermal noise. In contrast to impulsive noise, a stationary white noise w (see ③ in Figure 1) is coloured using a FIR filter designed in the same way as described in Section 3.1, before being added to the entire utterance:

$$y_{\text{si}}[n] = x[n] + g_{snr}^{\text{si}} \cdot z_{\text{si}}[n] \tag{5}$$

where

$$g_{snr}^{\text{si}} = \frac{10^{\frac{SNR}{20}}}{\|z_{\text{si}}\|^2 \cdot \|x\|^2} \tag{6}$$

is a gain parameter corresponding to a randomly chosen SNR and where z_{si} is the result of white noise w coloured by the FIR filter.

4. EXPERIMENTS AND RESULTS

Described in this section are the database, evaluation metric, baseline system and our results.

4.1. Dataset, protocols and metrics

The ASVspoof 2021 logical access (LA) task focuses on the development of spoofing and deepfake detection solutions that are robust to encoding and transmission channel variability. Spoofed speech data are generated with the same text-to-speech (TTS), voice conversion (VC) and hybrid algorithms (VC with TTS-generated inputs)

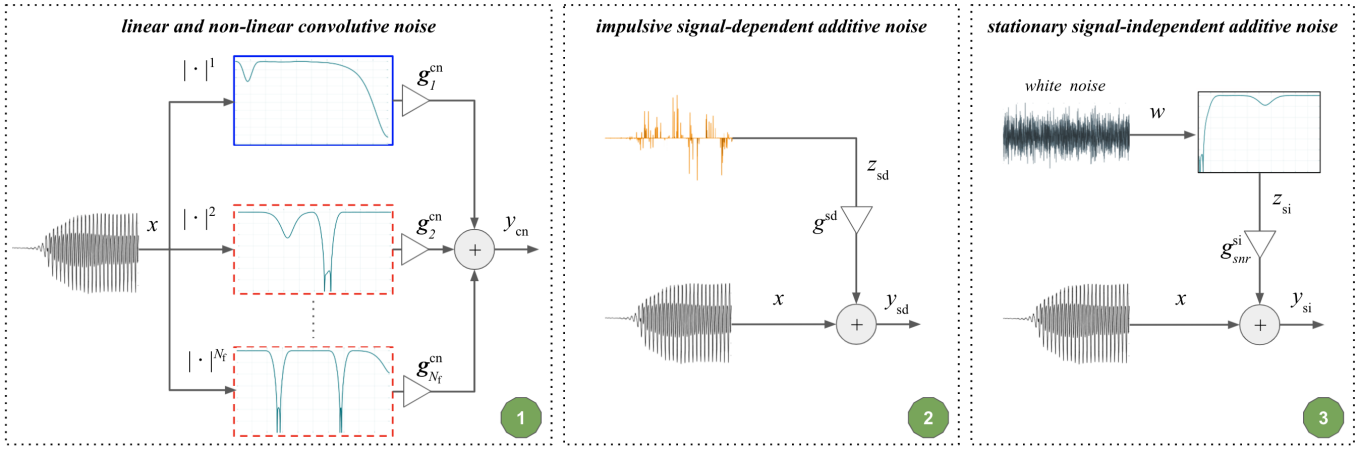

Fig. 1. Proposed RawBoost data augmentation framework including: (1) linear and non-linear convolutive noise; (2) impulsive signal-dependent additive noise; (3) stationary signal-independent additive noise. In (1), the profile in each rectangular box shows the frequency response for first harmonic (linear, solid blue box) and higher order harmonics (non-linear, dashed red boxes).

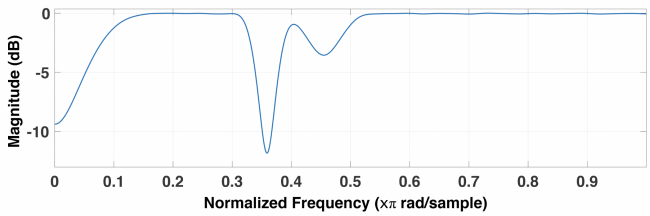

Fig. 2. Magnitude response of a multi-band filter with $N_{\text{notch}} = 3$ notch filters centered at normalised frequencies of 0.01, 0.35 and 0.45, bandwidths 0.06, 0.03 and 0.02 and number of filter coefficients 30, 94 and 52.

Table 1. RawBoost parameter values. Values within expressed ranges are selected at random (uniform distributions).

Param	N_{notch}	N_{fir}	N_f	f_c [Hz]	Δf [Hz]	g_1^{cn} [dB]	$g_{2\text{-}N_f}^{\text{cn}}$ [dB]	P_{rel} [%]	g^{sd}	SNR [dB]
①	5	[10,100]	5	[20,8k]	[100,1k]	[0,0]	[-5,-20]	-	-	-
②	-	-	-	-	-	-	-	[0,10]	2	-
③	5	[10,100]	1	[20,8k]	[100,1k]	-	-	-	-	[10,40]

used for the 2019 challenge. In contrast to the ASVspoof 2019 LA training and development data, all ASVspoof 2021 evaluation data is transmitted across some form of communications network, e.g., a public switched telephone network (PSTN) or a voice over Internet Protocol (VoIP) network using one of a number of different, popular telephony codecs, e.g., A-law and G.722 codecs, though other, unknown or unannounced codecs were also used (codec and other meta data was withheld from participants), giving the following conditions: **C1**: no encoding/transmission, **C2**: A-law, VoIP; **C3**: PSTN; **C4**: G.722, VoIP; **C5-C7**: unknown. As per the 2021 challenge rules, we used only the ASVspoof 2019 LA training and development partitions in optimising our spoofing countermeasure. We used the default minimum normalised tandem detection cost function (t-DCF) [35] as a primary metric but also report results in terms of the pooled equal error rate (EER).

4.2. Baseline

The baseline is an end-to-end RawNet2 system [6]. It is among the best-performing single systems and all results are fully reproducible using open source software.[2] The same system was adopted as one of four baselines for the ASVspoof 2021 challenge [1, 2]. The first sinc layer is initialised with a bank of 20 mel-scaled filters. Each filter has an impulse response of 1025 samples (64 ms duration) which is convolved with the raw waveform. The latter are truncated or concatenated to give segments of 4 seconds duration (64,600 samples). The sinc layer is followed by a residual network and a gated recurrent unit (GRU) to predict whether the input audio is bona fide or spoofed. We used the Adam optimiser with a mini-batch size of 128 and a fixed learning rate of 0.0001 and train for 100 epochs. Full details of the baseline system are available in [6].

4.3. RawBoost configurations

RawBoost parameters are generated according to the configuration options illustrated in Table 1 for each of the three techniques. Values expressed within ranges are drawn from the corresponding uniform distributions. Each technique is applied alone as well as in different combinations and in both series and parallel.[3] For series combinations, the output of one technique is used as the input to the next. For parallel combinations, an original input utterance is treated independently with each technique before the resulting distortions are combined. Output waveforms are normalised to prevent overflow. In our experiments, we used RawBoost to add nuisance variability on-the-fly to *existing* training data, instead of to generate *additional* data. Since the ASVspoof 2019 LA development data exhibits neither encoding nor transmission variability, and in order to respect properly the ASVspoof 2021 protocols and evaluation rules, we applied RawBoost also to the development data. RawBoost parameters and ranges illustrated in Table 1 were then selected based on the results of experimentation involving boosted and augmented training and development data only.

[2] https://github.com/asvspoof-challenge/2021/tree/main/LA/Baseline-RawNet2

[3] Due to space limitation, only a selection of best results is reported.

Table 2. ASVspoof 2021 LA RawNet2 results in terms of min t-DCF for each codec, pooled min t-DCF (P1) and pooled EER (P2).

Augmentation	Method	C1	C2	C3	C4	C5	C6	C7	P1	P2
none	-	0.4629	0.5594	0.7886	0.4954	0.5582	0.6774	0.5727	0.4257	9.50
RawBoost	(1) linear and non-linear convolutive noise	0.4531	0.5077	0.6160	0.4731	0.5019	0.5819	0.5317	0.3527	7.22
	(2) impulsive signal dependent noise	0.4373	0.5015	0.5041	0.4751	0.4920	0.5385	0.5099	0.3260	6.09
	(3) stationary signal independent noise	0.4544	0.5094	0.5349	0.4811	0.5036	0.5289	0.4964	0.3372	7.85
	***series*: (1)+(2)**	**0.4449**	**0.4806**	**0.5046**	**0.4635**	**0.4616**	**0.5025**	**0.4776**	**0.3099**	**5.31**
	parallel: (1)+(2)	0.4471	0.5094	0.5507	0.4724	0.5032	0.5585	0.5243	0.3261	5.57
	series: (1)+(3)	0.4569	0.5203	0.5576	0.4765	0.5057	0.5442	0.5134	0.3361	6.27
	series: (2)+(3)	0.4640	0.5056	0.5100	0.4910	0.5060	0.5240	0.5171	0.3329	6.58
	series: (1)+(2)+(3)	0.4437	0.4910	0.4986	0.4576	0.4937	0.5037	0.4858	0.3192	5.39
WavAugment	(1) time-drop	0.4582	0.5049	0.5133	0.4598	0.5094	0.5296	0.4739	0.3490	8.72
	(2) band-reject	0.4763	0.5417	0.5912	0.4957	0.5387	0.5628	0.5174	0.3692	8.86
	(3) additive-noise	0.5508	0.6721	0.7014	0.5531	0.6649	0.6549	0.5660	0.4819	13.38
	series: (1)+(2)+(3)	0.4652	0.4897	0.5172	0.4736	0.4802	0.5163	0.4990	0.3435	7.32
	series: (pitch)+(reverberation)+(1)+(3)	0.6130	0.7013	0.7351	0.6138	0.7153	0.7229	0.6307	0.5414	15.66
SpecAugment	(1) frequency-masking	0.4579	0.5292	0.7171	0.4894	0.5399	0.6642	0.5335	0.4214	9.80
	(2) time-masking	0.4581	0.5049	0.5134	0.4598	0.5094	0.5295	0.4739	0.3491	8.72
	series: (1)+(2)	0.4668	0.4985	0.5032	0.4927	0.4918	0.5162	0.4822	0.3418	8.25

4.4. Results

Results are illustrated in Table 2 for the baseline system (row 1), and for the same system trained using one of the three approaches to data augmentation: RawBoost; WavAugment; SpecAugment. For SpecAugment experiments,[4] frequency (channel) masking is applied at the sinc filterbank level to mask random contiguous sinc channels during training. In each case, results are shown for separate augmentation techniques and a selection of combinations (column 2). Columns 3-9 show results for each evaluation condition (C1-C7). Columns 10 and 11 show the pooled min t-DCF (P1) and pooled EER (P2). All RawBoost DA strategies lead to better performance than the baseline for all 7 evaluation conditions. The baseline pooled min t-DCF of 0.4257 drops to 0.3527 when using linear and non-linear convolutive noise (1), to 0.3260 using impulsive signal dependent additive noise (2) and to 0.3372 using stationary signal-independent additive noise (3). The best result is obtained using the RawBoost (1)+(2) system for which the min t-DCF is 0.3099 (27% relative reduction over the baseline) and the EER is 5.31% (44% relative reduction). The addition of stationary noise, while beneficial on its own, does not lead to any improvements in performance when combined with other techniques. This is not entirely surprising given that ASVspoof LA data does not contain any ambient noise. The technique may yet prove beneficial for other tasks, e.g., the physical access (PA) scenario, that *do* contain ambient noise.

4.5. Comparison to competing systems

Illustrated in Table 3 is a comparison of RawBoost performance to that of competing systems reported in the literature. To focus upon the benefits of data augmentation, the comparison is restricted to single systems.[5] The RawNet2 system with (1)+(2) RawBoost DA gives the third best result. Among the top three systems, only

[4]SpecAugment is not applied to raw waveforms but at the filterbank output instead. Results are included nonetheless for comparison. Only augmentation techniques applied at the raw waveform level support learning of the filterbank layer using augmented data.

[5]While some ensemble systems outperform those considered here, they are substantially more complex and their inclusion would compound the difficulty in assessing *data augmentation*. Unlike the comparisons made in Table 2, differences in Table 3 stem from differences in data augmentation as well as the underlying models/classifiers.

Table 3. A performance comparison for the ASVspoof 2021 evaluation partition in terms of pooled min t-DCF and pooled EER for different state-of-the-art single systems.

system	front-end	DA approach	min t-DCF	EER
LCNN [36]	Mel STFT	RS Mixup and FIR	0.2430	2.21
ResNet-L-LDE [37]	LFB	Frequency masking	0.2720	3.68
Ours:RawNet2	Raw	RawBoost (1)+(2)	0.3099	5.31
SE-ResNet18 [38]	LFCC	codecs	0.3129	6.62
RawNet2 [11]	Raw	codecs	0.3168	6.36
LCNN [20]	CQT	codecs	0.3197	5.27

RawNet2 operates upon raw waveform inputs. The ResNet-L-LDE system [37], which uses SpecAugment data augmentation in the form of frequency masking, uses external data contained within the MUSAN database. In contrast, RawBoost requires no such external data. The top-performing LCNN system [36] uses random square (RS) mixup [39] and FIR filtering DA. The FIR filtering approach aims to emulate the application of different telephony codecs and is conceptually similar to our use of FIR filtering. While applied at the data level, RS mixup is accompanied with modifications at the model level (the loss function in [36]). RawBoost requires no such intervention at the model level. The remaining systems included in Table 3 all augment data using speech codecs. RawBoost is competitive with all these approaches while not requiring the use of any additional codec implementations.

5. CONCLUSIONS

RawBoost can be used to boost or augment the pool of data available for training by generating new utterances which exhibit the variability expected in telephony scenarios. New raw waveforms are generated by perturbing a set of source utterances using linear and non-linear convolutive, and both impulsive and stationary additive noise. Our results show that RawBoost improves the performance of a raw end-to-end baseline spoofing detection solution by up to 27% relative. RawBoost is also data, application and model agnostic; it operates upon an existing source database without the need for any additional external data, nor intervention at the model level. It might hence have application to other related audio classification tasks.

6. REFERENCES

[1] H. Delgado, N. Evans et al., "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *arXiv preprint arXiv:2109.00535*, 2021.

[2] J. Yamagishi, X. Wang et al., "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof 2021 Workshop*, 2021.

[3] D. S. Park, W. Chan et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, 2019.

[4] G. Kim, D. K. Han et al., "SpecMix: A mixed sample data augmentation method for training withtime-frequency domain features," in *Proc. INTERSPEECH*, 2021.

[5] J.-w. Jung, S.-b. Kim et al., "Improved RawNet with Filter-wise Rescaling for Text-independent Speaker Verification using Raw Waveforms," in *Proc. INTERSPEECH*, 2020.

[6] H. Tak, J. Patino et al., "End-to-end anti-spoofing with rawnet2," in *Proc. ICASSP*, 2021.

[7] Y. Ma, Z. Ren et al., "RW-Resnet: A novel speech anti-spoofing model using raw waveform," in *Proc. INTERSPEECH*, 2021.

[8] H. Tak, J. Jung et al., "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. ASVspoof 2021 workshop*, 2021.

[9] W. Ge, J. Patino et al., "Raw differentiable architecture search for speech deepfake and spoofing detection," in *Proc. ASVspoof 2021 workshop*, 2021.

[10] X. Chen, Y. Zhang et al., "UR channel-robust synthetic speech detection system for ASVspoof 2021," in *Proc. ASVspoof 2021 Workshop*, 2021.

[11] J. Cáceres, R. Font et al., "The Biometric Vox system for the ASVspoof 2021 challenge," in *Proc. ASVspoof2021 Workshop*.

[12] E. Kharitonov, M. Rivière et al., "Data augmenting contrastive learning of speech representations in the time domain," in *Proc. IEEE SLT*, 2021.

[13] J. Wang, L. Perez et al., "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, 2017.

[14] T. Ko, V. Peddinti et al., "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015.

[15] T.-L. Vu, Z. Zeng et al., "Audio codec simulation based data augmentation for telephony speech recognition," in *Proc. APSIPA ASC*, 2019.

[16] C. Zhang, S. Ranjan et al., "An analysis of transfer learning for domain mismatched text-independent speaker verification," in *Proc. Odyssey Workshop*, 2018.

[17] T. Chen, A. Kumar et al., "Generalization of audio deepfake detection," in *Proc. Odyssey Workshop*, 2020.

[18] Y. Zhao, R. Togneri et al., "Replay anti-spoofing countermeasure based on data augmentation with post selection," *Computer Speech & Language*, vol. 64, 2020.

[19] R. K. Das, J. Yang et al., "Data augmentation with signal companding for detection of logical access attacks," in *Proc. ICASSP*, 2021.

[20] R. K. Das, "Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021," in *Proc. ASVspoof2021 Workshop*, 2021.

[21] Y. Zhang, G. Zhu et al., "An empirical study on channel effects for synthetic voice spoofing countermeasure systems," in *Proc. INTERSPEECH*, 2021.

[22] A. Krizhevsky, I. Sutskever et al., "ImageNet classification with deep convolutional neural networks," *NeurIPS*, vol. 25, 2012.

[23] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML WDLASL*, 2013.

[24] X. Cui, V. Goel et al., "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM TASLP*, vol. 23, 2015.

[25] H. Dinkel, N. Chen et al., "End-to-end spoofing detection with raw waveform CLDNNS," in *Proc. ICASSP*, 2017.

[26] H. Guo and H. L. Viktor, "Boosting with data generation: improving the classification of hard to learn examples," in *Proc. IEA/AIE*, 2004.

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

[28] A. Y. Kibangou and G. Favier, "Wiener-hammerstein systems modeling using diagonal volterra kernels coefficients," *IEEE signal processing letters*, vol. 13, 2006.

[29] A. V. Oppenheim and R. W. Schafe, "Discrete-time signal processing (3rd Ed.)," 2011.

[30] C.-W. Kok and T. Q. Nguyen, "Multirate filter banks and transform coding gain," *IEEE transactions on signal processing*, vol. 46, 1998.

[31] S. Yin, C. Liu et al., "Noisy training for deep neural networks in speech recognition," *EURASIP JASM*, vol. 2015, 2015.

[32] J. Huh, H. S. Heo, J. Kang et al., "Augmentation adversarial training for unsupervised speaker recognition," in *Workshop on SAS, NeurIPS*, 2020.

[33] U. Tiwari, M. Soni et al., "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *Proc. ICASSP*, 2020.

[34] R. Yang, "Additive noise detection and its application to audio forensics," in *Proc. APSIPA*, 2014.

[35] T. Kinnunen, H. Delgado et al., "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM TASLP*, vol. 28, 2020.

[36] A. Tomilov, A. Svishchev et al., "STC antispoofing systems for the ASVspoof2021 challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[37] T. Chen, E. Khoury et al., "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[38] W. H. Kang, J. Alam et al., "CRIM's system description for the ASVspoof 2021 Challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[39] H. Zhang, M. Cisse et al., "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.