The Vicomtech partial deepfake detection and location system for the 2023 ADD Challenge

Juan Manuel Martín-Doñas[1,*], Aitor Álvarez[1]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)

Abstract

This paper describes our submitted system to the 2023 Audio Deepfake Detection Challenge Track 2. This track focuses on locating the manipulated regions in partially fake audio. Our approach integrates a pre-trained Wav2Vec2 based feature extractor and two different downstream models for deepfake detection and audio clustering. While the detection module is composed of a simple but efficient downstream neural classification model, the clustering-based neural network was trained to first segment the audio and then discriminate between the original regions and the manipulated segments. The final segmentation was obtained by combining the clustering process with the decision score through the application of some post-processing strategies. We evaluate our system on the test set of the challenge track, showing good performance for partially fake detection and location in challenging environments. Our novel, simple and efficient approach ranked fourth in the mentioned challenge among sixteen participants.

Keywords

Partial deepfake, anti-spoofing, audio tampering, wav2vec2, deep clustering

1. Introduction

Through the last decades, the improvements in deep learning technologies have favored the development of speech synthesis and voice conversion algorithms that can achieve high-quality audio signals [1]. Despite its various benefits and potential applications, the generation of human-like speech signals involves a risk: the creation of audio deepfakes that can deceive people and bring potential harm to society [2]. Therefore, advanced audio deepfake detection technologies [3, 4] are required to fight against the misuse of synthetic audio and discriminate them from genuine speech. The research efforts in this direction have fostered several competition campaigns which aim researchers to push the technological limits of deepfake detection and face continuously increasingly challenging scenarios. ASVspoof series [5, 6] represent an example of such challenges, which focused on the development of anti-spoofing countermeasures for automatic speaker verification systems. More recently, the 2022 Audio Deepfake Detection (ADD) challenge [7] was organized, further exploring the potential of audio generation and detection technologies.

The continuous and widespread advances of anti-spoofing and deepfake detection neural networks have yielded the research community to consider a new scenario that poses a potential thread: partial deepfakes [8, 9]. In this scenario, a genuine utterance is manipulated by inserting a fake clip into the original audio. Thus, only part of the signal is synthetic, which makes this class of deepfakes harder to detect by the previous countermeasure systems. Recent works have carried out the development of partial deepfake detection systems [10, 11, 12], bringing database resources for interested researchers. Furthermore, Track 2 of the ADD 2022 challenge intended to develop these technologies in an even more complex scenario in which the manipulation can also be done using another real audio segment, leading researchers to adapt detection technologies to this particular scenario [13, 14, 15]. The recently launched 2023 ADD challenge [16] Track 2 advanced the technology to the next step by locating the manipulated regions in partial fake audios. Hence, the proposed systems should detect not only if the audio was a partial deepfake or not, but also divide the audio into genuine and fake segments.

This paper presents our contributions to the 2023 ADD challenge Track 2 for manipulation region location. Our proposed system is based on a pre-trained wav2vec2 (W2V2) feature extractor to compute high-level representative information from the audio. These deep features are used for downstream detection and clustering neural networks, which score the input audio and cluster the segments of it into two categories, respectively. This joint information is used for partial deepfake detection and the location of the fake clips from the audio. We explored different alternatives for combining these outputs and using post-processing strategies to refine the segmentation. Our best system got the fourth position among sixteen participants of Track 2, achieving competitive results with a novel approach in challenging conditions.

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

*Corresponding author.

jmmartin@vicomtech.org (J. M. Martín-Doñas); aalvarez@vicomtech.org (A. Álvarez)

0000-0003-4874-0166 (J. M. Martín-Doñas); 0000-0002-7938-4486 (A. Álvarez)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3597/paper7.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

The remainder of this paper is organized as follows. Section 2 briefly reviews related works for partial deepfake detection and location. Our proposed approach is described in Section 3, including the different neural network modules and post-processing techniques. Then, Section 4 discusses the experimental framework and challenge results. Finally, the conclusions are summarized in Section 5.

2. Related work

The increasing interest in partial deepfakes has contributed to creating related databases to train and evaluate countermeasure systems. This is the case of the PartialSpoof [8] or the Half-Truth databases [9]. For the former, the authors experimented with detection systems that combine self-supervised feature extractors with both utterance- and segment-level scoring [10, 12]. Several techniques were proposed for the partial deepfake detection track of ADD 2022 challenge [13, 14, 17, 18] considering different architectures and data augmentation techniques. Although the goal was the detection of partial fakes, the work in [15] included a fake span discovery mechanism via question answering to predict the boundaries of the fake clip, but the location performance was not evaluated. On the other hand, recent works have also explored the location of fake clips or manipulated audio by detecting the artifacts in the boundaries of the inserted segment [19, 20, 21]. Thus, these works commonly addressed the problem as a classification task to detect the frames containing the boundaries.

Our proposed approach is based on our previous system for partial deepfake detection presented to the 2022 ADD challenge [13]. We further improved our method by considering a clustering-based neural network to segment the original and manipulated parts of the audio. This approach is inspired by previous works in source separation [22] and speaker diarization [23], where this deep clustering technique proved to perform well. We explored the clustering approach for partial deepfake location for two reasons. Firstly, as the manipulation can be done with an original clip, directly detecting spoofed segments in the audio is not feasible. Secondly, the direct detection of boundaries can be challenging to address due to the class-imbalance behavior of the problem (few boundary segments in the whole audio). Our approach is based on first segmenting the audio by exploiting acoustic differences in consecutive regions generated by different speakers, genuine and fake audio segments, the background o boundary artifacts, among others. Then, this information is used along our detection-based module to classify the segments into original or fake clips. To the best of our knowledge, this is the first work exploring clustering-based methods for partial deepfake location.

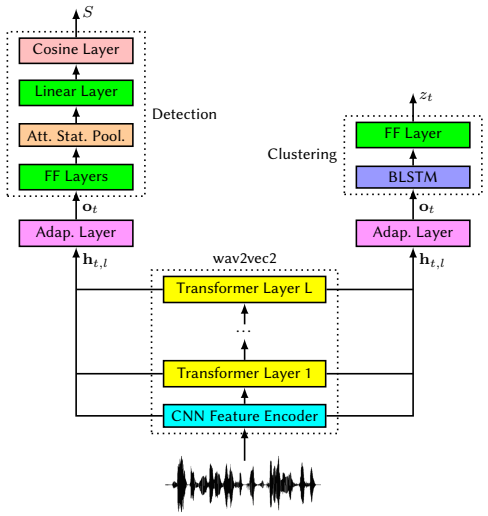

Figure 1: Overview of the proposed partial deepfake detection and location system based on wav2vec2.

3. Proposed approach

A diagram of our proposed system for the ADD Challenge 2023 is shown in Figure 1. Our system is composed of a W2V2 feature extractor and the corresponding detection and clustering neural networks. A detailed explanation of each part is presented in the following subsections, including the processing of the final system output.

3.1. Wav2Vec2 Feature Extractor

First, we extract deep features from the speech signal using pre-trained self-supervised neural networks. In this work we integrated the XLS-128 W2V2 model [24, 25], which includes 300M parameters and was trained with unlabelled speech data from 128 different languages. A 7-layer convolutional network encoder is first used to process the speech signal, extracting 1024-dimension vector representations every 20 ms with a receptive field of 25 ms. These features are further processed with a 24-layer Transformer to compute contextualized representations from the speech signal. The model was trained using self-supervised techniques to learn high-level representations of the speech data.

Instead of fine-tuning the W2V2 model for each specific task, we considered a general pre-trained feature extractor for the different downstream models. We exploited the information from the different transformer layers in order to robust the speech representation for the downstream task. Thus, a trainable adaptation layer is plugged between the W2V2 feature extractor and the downstream model. The adaptation layer consists of a temporal normalization layer [26] followed by a weighted

sum of the different layers $\mathbf{o}_t = \sum_{l=0}^{L} \alpha_l \mathbf{h}_{t,l}$, where $\mathbf{h}_{t,l}$ is the hidden representation for time step t and transformer layer l, L is the number of transformer layers, and α_l are trainable weights normalized to sum one. The output representations $\mathbf{o}_t$ are then used as the input for the specialized downstream models [27].

3.2. Deepfake detection and clustering networks

The objective of the task is not only to detect partially fake audio but also to indicate the segment of the speech that has been manipulated. Therefore, our approach is based on combining two different modules which we have named as detection and clustering networks. The architectures of both modules are described in the following paragraphs.

The detection network is identical to the one used in our previous work for the ADD2022 Challenge [13]. The network scores the audio signal, where the scoring is related to the likeness of the audio to be genuine. The output representations from the corresponding adaptation layer are first processed by two feed-forward (FF) layers with ReLU activations and dropout. A single embedding vector from the audio is then obtained using attentive statistical pooling followed by a linear activation layer. Finally, the final score is obtained using a cosine layer, which computes the cosine similarity between the embedding and a trainable vector network. The model is trained to give higher scores for genuine speech using a One-class (OC) softmax loss function [28].

The manipulated segment location involves a more challenging task. In this challenge, the partial deepfake can be created using synthetic or real audios, which makes it difficult to use previous methods for partial spoofing segment scoring [12]. Faced with this complex scenario, we propose using a clustering-based approach, where the network is trained to classify the output representations into two different clusters whenever they represent original audio, manipulated real clips, or synthetic speech. The underlying idea is that the network can learn to segment the audio when detecting different acoustic conditions or insertion artifacts. The output representations are fed into a recurrent bidirectional neural network based on LSTM (BLSTM) to exploit temporal information on the vectors fully. The output is processed by a FF layer with sigmoid activation that outputs a single value per time. The network is trained using binary cross-entropy (BCE) loss. To avoid the label ambiguity problem in the clusters assignment, we used permutation-invariant training (PIT) [22].

To summarize the network architecture and make its structure easier to understand for the reader, in Table 1 we show the layers and their output dimensions for each corresponding network module.

Table 1
Architecture of the involved network modules. It includes each layer and its output dimension, where T is the number of time frames and N the size of the mini-batch.

Layer name	Output dimension
Wav2Vec2	
CNN encoder	$N \times \mathsf{T} \times 1024$
Transformer	$N \times \mathsf{T} \times 1024$
Output features	$N \times \mathsf{T} \times 1024 \times 25$
Adaptation layer	$N \times \mathsf{T} \times 1024$
Detection network	
FF Layers	$N \times \mathsf{T} \times 128$
Att. Stat. Pool.	$N \times 256$
Linear layer	$N \times 128$
Cosine Layer	N
Clustering network	
BLSTM	$N \times \mathsf{T} \times 256$
FF Layer	$N \times \mathsf{T}$

3.3. Post-processing strategies

Our proposed systems outputs an score $S \in [-1, 1]$ from the detection model and a sequence $z_t = [z_1, \cdots, z_T] \in [0, 1]$ from the clustering module. We analyzed the results in the development set to design a group of post-processing strategies for the final clustering of the testing audios. We first binarized the sequence z_t by thresholding to segment the speech signal into two clusters, where the threshold is set experimentally. Sequences with only one segment (unique cluster) are classified as genuine or fake clips using the detection score S with a pre-defined optimum threshold. Moreover, during the PIT training, the network seems to have learnt a discriminative clustering between original and fake clips. We observed this on the development set, where the clustering module tends to assign lower values to the part of the original clip and higher values for manipulated or fake clips. This may be due to the artifacts in the fake clip, or the boundaries, yielding the network to detect these structures. Then, for the first attempt, we directly map the segments with zero value to genuine/real audio parts, and the rest with one value to fake parts.

However, there were some audios for which the clustering module tended to detect several short segments of fake audio. We define short segments as thoses with a duration spanning the receptive field of the W2V2 model (i.e., one or two W2V2 feature frames). This effect was more common in genuine clips manipulated with real audio segments, specially in those in which the genuine and manipulated segment belong to the same speakers. Thus, the network seems to detect the artifacts on the boundaries in the inserted clips as fake segments. In those cases, we hypothesized that this information could

help to discover the fake segments from a manipulated region. Following this assumptions, we evaluated three post-processing strategies applied to the initially computed clusters:

1. If two non-consecutive short segments are detected as fake, and the rest are tagged as genuine, the segments in the middle of the fake segments are labelled as fake as well.
2. If only a single short segment was detected as fake, we considered it as a single boundary dividing two segments, where the longest one is labelled as genuine and the shortest as fake.
3. As an extension of the first rule, if more than 2 short segments were initially tagged as fake, the first and last segments were considered as borders and all the segments in the middle were labelled as fake as well, by keeping the segments outside the borders as genuine.

Although these strategies are mainly based on empirical analysis of development data, we expected they help to improve the location accuracy of partially fake clips. We will discuss the benefits of the post-processing settings in the following subsection.

4. Experimental results

In this section, we first describe the speech databases used during the challenge to train and test the different approaches. Then, we explain the training setup procedure as well as the data augmentation techniques applied to the speech data. Finally, we show and analyze the experimental results obtained during the challenge evaluation phase and our best approach.

4.1. ADD 2023 Track 2 database

The ADD 2023 Track 2 database [16] comprises genuine and partially fake Chinese speech. The partial deepfakes can be complete fake audios generated by synthetic or voice conversion techniques, or manipulated audios where another real clip or synthetic segment has replaced a part of a genuine utterance. The database includes training, development, and evaluation sets with approximately 53K, 18K, and 50K utterances, respectively. The audios from the training and development partitions are annotated, indicating whether they are genuine or partial deepfakes, and the temporal boundaries of the original and inserted clips.

4.2. Experimental framework

The downstream models were fine-tuned using the data from the training set in the ADD2023 Track 2 database. The detection and clustering modules were trained independently using the same pre-trained W2V2 model as feature extractor. The training of both models was performed with the Adam optimizer. [29] with a learning rate of $3 \cdot 10^{-4}$ without weight decay, and the dropout rate for the FF layers in the detection network was set to 0.2. The W2V2 parameters were frozen during training. In order to cope with limitations in GPU memory, we used a mini-batch of 8 utterances and accumulated gradients of eight mini-batches, yielding an effective mini-batch of 64 utterances. The length of the utterances was adjusted using zero-padding during batch creation. To validate the model and keep the best parameters, we evaluated the performance on the development set after each epoch and considered an early-stopping strategy. Thus, we kept the model with the lowest validation loss, and the training stopped after 10 epochs without improvements on the development set.

Additionally, with the aim of improving our models' robustness for challenging conditions, we also explored data augmentation techniques during the training phase. We considered the recently proposed Rawboost method [30], which has shown outstanding performance for deepfake detection and anti-spoofing systems. Rawboost operates directly upon raw waveform speech signals by considering different distortions created *on the fly* for a given audio signal. In particular, we evaluated using linear and non-linear convolutive noise as well as impulsive signal-dependent additive noise. We chose this method to evolve the finite impulse response (FIR) data augmentation we analyzed for partial deepfake detection in the ADD 2022 challenge [13], which showed significant improvements in the tasks evaluated.

4.3. Challenge results

We evaluated our proposed approach in the test set of the ADD2023 Track2 database. The participant systems in the challenge were evaluated using a metric score defined by the organizers. This score is computed as follows,

$$\text{Score} = 0.3 \cdot A_{\text{sen}} + 0.7 \cdot F_{\text{1seg}}, \tag{1}$$

where A_{sen} is the average sentence accuracy (related to the classification of audios as genuine or fake), and F_{1seg} corresponds to an average segment F_1 score that measures the ability of the model to identify manipulated areas in speech correctly. The F_1 score is obtained by taking fake segments as positive samples and using a frame length of 10 ms. The weighted factors were defined by the organizers, giving more importance to the segmentation than the fakeness detection.

Table 2 shows the results obtained for our different evaluated systems in the challenge. For each method, we provide the resulting final score, the sentence accuracy, and the segment F_1 score. In addition, we also

Table 2
Final results of our submitted approaches to the ADD 2023 Track 2 challenge. They include our clustering and detection approach with different post-processing strategies. The systems are compared in terms of detection accuracy and segment classification.

System	$\mathbf{A}_{\text{sen}}$	$\mathbf{P}_{\text{seg}}$	$\mathbf{R}_{\text{seg}}$	$\mathbf{F}_{1\text{seg}}$	Score
Cluster	77.89	**68.66**	11.99	20.42	37.66
Cluster + Det.	**78.16**	41.27	17.53	24.61	40.67
Cluster + Det. + PP (1)	**78.16**	51.26	29.13	37.15	49.45
Cluster + Det. + PP (1+2)	**78.16**	48.42	44.77	46.52	56.01
Cluster + Det. + PP (2+3)	**78.16**	50.15	**53.29**	**51.67**	**59.62**

report the precision and recall for the fake segments. We compared five different alternatives: standalone clustering approach (Cluster), classification of single-segment audios using the detection module with an optimized threshold (Det.), and the combination of post-processing stages to refine the clusters (PP). As it can be observed, the best results were obtained when the clustering and detection modules were combined with post-processing strategies, especifically when applying one and multiple fake short-segments detection (rules 2 and 3).

An analysis of the results obtained gives some insights into how the different steps help improve the final score. The base system shows good sentence detection accuracy (close to 78%), but with the lowest recall of the fake segments, which indicates that the system does not recover many manipulated parts. The detection threshold optimization yields some improvements in the accuracy and the recall at the expense of a reduction in precision. Nevertheless, the final results show better performance in terms of F_1 score. Then, applying the post-processing techniques on detected manipulated utterances helps to better locate the regions with fake audio. The three evaluated combinations show similar precision around 50%, but the best combination of rules yields the best recall of 53% and a F_1 score close to 52%. It demonstrates how we can exploit the detected artifact regions during the clustering step to refine the final output and better catch insertion clips.

5. Conclusions

In this work, we have presented our proposed system for the 2023 ADD challenge Track 2 based on a pre-trained wav2vec2 feature extractor and downstream neural networks for detection and clustering of partial deepfakes. Our approach exploits the deep features from W2V2 to develop two modules: (1) a detection network that scores the audio in terms of its likeness to be a partial deepfake, and (2) a clustering-based network that segments the audio to discriminate between original and fake clips. This information is combined to obtain the final segmentation, using post-processing strategies to refine the result. Our proposed method shows competitive results in the 2023 ADD challenge, ranking fourth in Track 2 among the participants. As future work, we will further research the combination of both modules to improve the accuracy in the detection of manipulated segments. In addition, we will analyze different data augmentation techniques to better adapt the system for challenging scenarios with different background conditions, as well as other self-supervised models to compute robust speech representations.

Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation programme in the context of project EITHOS, under Grant Agreement No. 101073928.

References

[1] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 132–157.

[2] T.-N. Le, H. H. Nguyen, J. Yamagishi, I. Echizen, Robust deepfake on unrestricted media: Generation and detection, in: Frontiers in Fake Media Generation and Detection, 2022, pp. 81–107.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and countermeasures for speaker verification: A survey, Speech Communication 66 (2015) 130–153.

[4] C. B. Tan, et al., A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction, Multimedia Tools and Applications 80 (2021) 32725–32762.

[5] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, K. A. Lee, ASVspoof 2019: spoofing

countermeasures for the detection of synthesized, converted and replayed speech, IEEE Transactions on Biometrics, Behavior, and Identity Science 3 (2021) 252–265.

[6] J. Yamagishi, et al., ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection, in: Proc. ASVspoof Workshop, 2021, pp. 47–54.

[7] J. Yi, et al., ADD 2022: The first audio deep synthesis detection challenge, in: Proc. ICASSP, 2022, pp. 9216–9220.

[8] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, N. Evans, An initial investigation for detecting partially spoofed audio, in: Proc. InterSpeech, 2021, pp. 4264–4268.

[9] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, R. Fu, Half-Truth: A partially fake audio detection dataset, in: Proc. InterSpeech, 2021, pp. 1654–1658.

[10] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, Multitask learning in utterance-level and segmental-level spoof detection, in: Proc. ASVspoof Workshop, 2021, pp. 9–15.

[11] M. H. Rahman, et al., Detecting synthetic speech manipulation in real audio recordings, in: Proc. IEEE WIFS, 2022.

[12] L. Zhang, X. Wang, E. Cooper, N. Evans, J. Yamagishi, The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 813–825.

[13] J. M. Martín-Doñas, A. Álvarez, The Vicomtech audio deepfake detection system based on Wav2Vec2 for the 2022 ADD Challenge, in: Proc. ICASSP, 2022, pp. 9241–9245.

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

[15] H. Wu, et al., Partially fake audio detection by self-attention-based fake span discovery, in: Proc. ICASSP, 2022, pp. 9236–9240.

[16] J. Yi, et al., ADD 2023: The second audio deepfake detection challenge, in: Proc. IJCAI Workshop on Deepfake Audio Detection and Analysis (DADA), 2023.

[17] M. Li, Y. Ahmadiadli, X.-P. Zhang, A comparative study on physical and perceptual features for deepfake audio detection, in: Proc. DDAM workshop, 2022, pp. 35–41.

[18] X. Liu, et al., Deep spectro-temporal artifacts for detecting synthesized speech, in: Proc. DDAM workshop, 2022, pp. 69–75.

[19] Z. Zeng, Z. Wu, Audio splicing localization: Can we accurately locate the splicing tampering?, in: Proc. ISCSLP, 2022, pp. 120–124.

[20] B. Zhang, T. Sim, Localizing fake segments in speech, in: Proc. ICPR, 2022, pp. 3224–3230.

[21] Z. Cai, W. Wang, M. Li, Waveform boundary detection for partially spoofed audio, in: Proc. ICASSP, 2023.

[22] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (2017) 1901–1913.

[23] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, P. Garcia, Encoder-decoder based attractors for end-to-end neural diarization, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022) 1493–1507.

[24] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Proc. NeurIPS, 2020, pp. 12449–12460.

[25] A. Babu, et al., XLS-R: Self-supervised cross-lingual speech representation learning at scale, arXiv preprint arXiv:2111.09296 (2021).

[26] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022 (2016).

[27] L. Pepino, P. Riera, L. Ferrer, Emotion Recognition from Speech Using wav2vec 2.0 Embeddings, in: Proc. InterSpeech, 2021, pp. 3400–3404.

[28] Y. Zhang, F. Jiang, Z. Duan, One-class learning towards synthetic voice spoofing detection, IEEE Signal Processing Letters 28 (2021) 937–941.

[29] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proc. ICLR, 2015.

[30] H. Tak, M. Kamble, J. Patino, M. Todisco, N. Evans, Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing, in: Proc. ICASSP, 2022, pp. 6382–6386.