THE VICOMTECH AUDIO DEEPFAKE DETECTION SYSTEM BASED ON WAV2VEC2 FOR THE 2022 ADD CHALLENGE

Juan M. Martín-Doñas and Aitor Álvarez

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)
{jmmartin, aalvarez}@vicomtech.org

ABSTRACT

This paper describes our submitted systems to the 2022 ADD challenge withing the tracks 1 and 2. Our approach is based on the combination of a pre-trained wav2vec2 feature extractor and a downstream classifier to detect spoofed audio. This method exploits the contextualized speech representations at the different transformer layers to fully capture discriminative information. Furthermore, the classification model is adapted to the application scenario using different data augmentation techniques. We evaluate our system for audio synthesis detection in both the ASVspoof 2021 and the 2022 ADD challenges, showing its robustness and good performance in realistic challenging environments such as telephonic and audio codec systems, noisy audio, and partial deepfakes.

Index Terms— antispoofing, wav2vec2, audio deepfakes, self-supervised, data augmentation

1. INTRODUCTION

Speech synthesis and voice conversion technologies [1] have rapidly grown in the last years, mainly thanks to the development of the deep learning paradigm. Although this broadens the application of these technologies, a threat is also present: the generation of deepfake speech that can even foolish modern automatic speaker verification systems [2]. The need for reliable countermeasures has favored the research on audio deepfake detection systems able to detect spoofed speech [3]. An example of this effort is the ASVspoof series [4, 5], consisting of biannual challenges focused on the development of antispoofing countermeasures for verification systems.

The continuous improvements in audio deepfake detection have widespread interest in developing robust solutions in realistic challenging scenarios. For example, the systems robustness on noisy and reverberant scenarios was studied in [6, 7]. Likewise, the ASVspoof2021 challenge [8] considered synthesized speech transmitted over different telephone networks, as well as speech deepfakes modified through commercial compression algorithms. Moreover, recent works [9] have explored the detection of partially spoofed audio. In order to continue improving these systems in complex environments, the 2022 Audio Deep synthesis Detection (ADD) challenge has been recently launched [10]. Its main goal is the detection of deep synthesis and manipulated audios, and it includes three different tracks: 1) Low-quality fake audio detection (diverse background noises and disturbances are contained in the audios); 2) Partially fake audio detection (several small fake speech segments are hidden in real audio); and 3) Audio fake game.

This paper presents our contributions to the ADD 2022 challenge, based on the use of the wav2vec2 (W2V2) [11] approach to extract discriminative information that helps detect spoofed audio. These models are trained with self-supervised learning methods with a large amount of unlabelled speech data, allowing to learn high-level representations of the speech signal. W2V2 has been explored in different speech processing tasks, such as speech recognition, speaker verification [12] and emotion classification [13]. Regarding audio deepfake detection, only a few works have explored this approach or similar [14, 15]. Differently from previous works, we propose to use a pre-trained W2V2 model as a feature extractor, but using the encoded representations of the different transformer layers. This information can be then exploited by a simple but effective downstream model to detect spoofing attacks. Moreover, we analyzed different data augmentation techniques to adapt the classifier to the final application scenario. Our approach was evaluated in both tracks 1 and 2 of the ADD challenge, where it achieved first and fourth position among the participants, respectively. In addition, we include results on the logical access and speech deepfake tracks of the ASVspoof 2021.

The remainder of this paper is organized as follows. In Section 2, we present our proposal based on the W2V2 approach. Section 3 describes the speech databases used for the evaluations and the data augmentation training strategies. Then, in Section 4, the results are presented and analyzed. Finally, conclusions are summarized in Section 5.

This work was supported in part by the Spanish Centre for the Development of Industrial Technology (CDTI) through the Project ÉGIDA—RED DE EXCELENCIA EN TECNOLOGIAS DE SEGURIDAD Y PRIVACIDAD under Grant CER20191012.

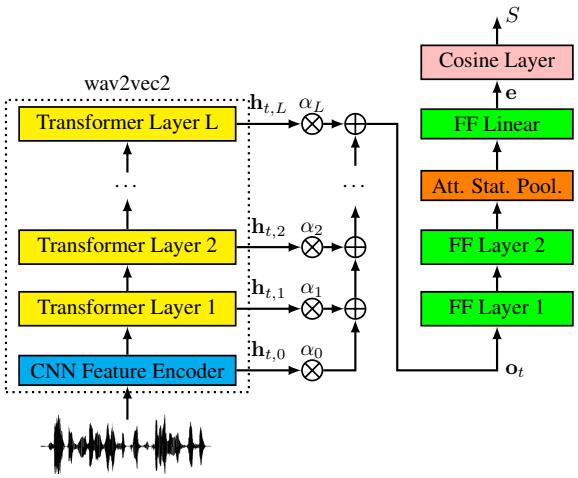

Fig. 1. Overview of the proposed audio deepfake detection approach based on wav2vec2.

2. WAV2VEC2-BASED AUDIO DEEPFAKE DETECTION APPROACH

A diagram of our proposed system for the ADD challenge is depicted in Figure 1. It is composed of a W2V2 feature extractor, which obtains the encoded speech representations, and a classification model that scores the input audio as genuine or spoof. A detailed explanation of each part is presented in the next subsections.

2.1. Wav2vec2 as Feature Extractor

As the base W2V2 architecture, we evaluated the Large models of 300M parameters, presented in [16], trained with 53 and 128 languages (XLS-53 and XLS-128, respectively). The raw speech signal is first processed by a feature encoder composed of several convolutional layers (CNN), which extracts vector representations of size 1024 every 20 ms, using a receptive field of 25 ms. Then, these encoder features are fed into a transformer network with 24 layers which is used to obtain contextualized representations of the speech signal. The model is trained in a self-supervised setting using a contrastive loss. The objective is to predict the quantized representations of certain masked encoded feature representations from a set of distractors using the contextualized representations. Thus, this model can learn high-level representations of the speech signal from a large amount of unlabelled data. The features extracted from the pre-trained W2V2 model can be used to train a downstream classifier in a specific speech processing task with a relatively few amount of labeled data. Moreover, the W2V2 and downstream models can be jointly trained in the related task. In this work, we explore the use of W2V2 as a pre-trained model, thus allowing to have a general feature extractor that can be used with specialized classification models adapted to each spoofing detection task.

Table 1. Architecture of the downstream classification model. 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 size
W2V2 features	$N \times \mathrm{T} \times 1024 \times 25$
Temp. Norm. + Layer weight.	$N \times \mathrm{T} \times 1024$
FF Layer (1 and 2)	$N \times \mathrm{T} \times 128$
Att. Stat. Pool.	$N \times 256$
FF Linear	$N \times 128$
Cosine Layer	N

2.2. Classification Model

The contextualized representations from the last transformer layer of a pre-trained model can be useful for certain speech tasks, like speech recognition. Nevertheless, previous works have shown that for other tasks, for example, speaker verification or emotion recognition, more discriminative information can be obtained from the first or intermediate layers [12, 13]. Therefore, we followed the methodology of these previous works and used the hidden representations of the different transformer layers as input for our downstream model.

The classification model processes the W2V2 hidden representations as follows. First, a temporal normalization [17] is applied at the input features from each transformer layer. Then, for each temporal step t, an output representation is computed from the normalized hidden representations $\mathbf{h}_{t,l}$ as $\mathbf{o}_t = \sum_{l=0}^{L} \alpha_l \mathbf{h}_{t,l}$, where l represents the hidden layer index, L the number of transformer layers, and α_l are network trainable weights, which are normalized to sum one [13]. The computed vectors at each time step are fed into two feed-forward (FF) layers with ReLU activation and dropout, and then they are passed to an attentive statistical pooling layer [18]. This attention layer computes and concatenates the temporal mean and standard deviation of the vectors at the different time steps, thus obtaining a single representation for the whole utterance. A linear layer (FF without non-linearity) is then applied to compute the embedding vector $\mathbf{e}$. Finally, the final score is obtained as a cosine similarity, $S = \cos(\mathbf{w}, \mathbf{e}) \in [-1, 1]$, where $\mathbf{w}$ is a vector network parameter representing the direction of genuine speech in the embedding space. The classification model is shown in more detail in Table 1. The model is trained to compute higher scores for genuine speech using a One-class softmax loss function [19].

3. EXPERIMENTAL FRAMEWORK

This section describes the speech databases used to train and evaluate our systems, as well as the data augmentation techniques and training setup procedure. For each of the challenges, we trained our systems using only the corresponding train data of that challenge (closed conditions).

3.1. ADD 2022 challenge database

The ADD 2022 database [10] comprises spoofed audio generated by speech synthesis and voice conversion systems. The train and development (dev) sets contain clean speech based on the multi-speaker Mandarin speech corpus AISHELL-3 [20]. Each set has about 28K utterances, with different speakers each. Tracks 1 and 2 include both an adaptation set with about 1K utterances, and a test set with about 100K utterances without labels. The adaptation set presents similar conditions that the audios in the test set, and it is provided to adapt the systems trained with the general train set. For track 1, the corresponding sets incorporate both genuine and spoof utterances with various real-world noises and background music. For track 2, the corresponding sets include genuine utterances and partially fake utterances generated by manipulating the original genuine speech with real or synthesized audio.

3.2. ASVspoof 2021 challenge database

We focused on the logical access (LA) and speech deepfake (DF) partitions of the ASVspoof 2021 challenge [8]. Both partitions contain spoofed audio generated by speech synthesis or voice conversion methods, with clean speech derived from the VCTK corpus [21]. The 2021 challenge only considered evaluation data, so the ASVspoof 2019 [5] LA train and dev sets were used for training. Each set contains about 25K utterances from 20 and 10 speakers, respectively. The 2021 LA and DF sets are similar to the 2019 LA data, but they consider more challenging scenarios. The LA set contains about 180K utterances of speech transmitted through real telephonic systems at different bandwidths and with different codecs. The DF set includes about 600K utterances of speech processed with various commercial audio codecs. Moreover, this set considers clean speech from other databases, making it more challenging.

3.3. Data augmentation techniques

In order to improve the performance of our approach and to adapt it to the different scenarios, we considered the use of data augmentation techniques to be applied *on the fly* during the training step. The main augmentation technique evaluated corresponded to the use of low-pass finite impulse response (FIR) filters on the speech signals. This procedure has shown its success for ASVspoof 2021 LA and DF tracks [22] as it emulates the effects of transmission artifacts and codecs. Moreover, they act by masking frequencies in the speech signal, improving the generalization capabilities of LA detection systems. Thus, they can be also useful for the conditions presented in the ADD database. In addition, they are directly applied to the raw waveform, making them compatible with the W2V2 network. We evaluated both narrowband (NB) and wideband (WB) FIR filters, following a similar procedure to that presented in [22].

For the ADD challenge, we joined the train and corresponding adaptation set of each track to improve the model generalization on the new challenging conditions. Although they are scarce, the adaptation data can be helpful for the W2V2-based system to be adapted to the corresponding scenario. Furthermore, for track 2 we also considered the generation of new partially fake audios. To this end, 20% of the genuine utterances in the train/dev set were randomly selected at each epoch. For each of these utterances, we selected a segment of variable duration, shorter than the original audio, from a different utterance (genuine or spoof) within the corresponding set. The segment was then overlapped over the original utterance in a random position.

3.4. Training setup

The models were trained using the Adam optimizer [23] with the default learning rate and a dropout of 0.2. During training, the W2V2 parameters were frozen, and only the classifier parameters were updated. A mini-batch of 8 utterances was used, and the parameters were updated every 8 iterations. At each epoch, the model was evaluated using the corresponding dev set, keeping the model with the lowest loss. The training stopped after 10 epochs without improvements on the dev set.

4. RESULTS

In this section, we describe the results obtained for our proposal and its different configurations in both ASVspoof 2021 and 2022 ADD challenges. The different systems were evaluated and compared in terms of equal error rate (EER), which is the most common metric used in biometric applications.

4.1. Results on ADD challenge

Table 2 shows the results obtained for tracks 1 and 2 of the ADD challenge. Our approach is evaluated in terms of the W2V2 model and data augmentation techniques, as well as the ADD data sets used for training. The XLS-128 model shows better results in general, but the main improvements come from the use of adaptation data along with the train set. The use of a small portion of data with similar conditions to the test set helps the model to adapt better to the scenario evaluated, reducing EER drastically. Moreover, in track 2 we generated more adaptation data through the partial fake augmentation strategy described in Subsection 3.3, increasing even further the model discrimination capabilities. Finally, we explored the use of NB FIR filters to increase the robustness under low-quality and partial fake conditions. This strategy reduced the EER by about 1% in both tracks and combined with the previous ones, it achieved the best results. Thus, this demonstrates that the FIR-based strategy is not only useful to emulate telephonic or codec conditions [22], but also helps to train better antispoofing systems in challenging conditions.

Table 2. Final results in terms of EER (%) of our submitted approaches to the ADD 2022 challenge. It also includes different variants for the W2V2 model and the data augmentation (DA) and adaptation strategies.

W2V2	Sets	DA	Track1	Track2
XLS-53	Train	-	32.96	38.09
	Tr.+Adap.	-	23.70	33.73
XLS-128	Train	-	32.20	45.88
	Tr.+Adap.	-	22.62	30.35
	Tr.+Adap.	FIR	**21.71**	-
	Tr.+Adap.	partial	-	17.58
	Tr.+Adap.	FIR+part.	-	**16.59**

Table 3. EER (%) results of our proposed W2V2 approach on ASVspoof 2021, considering different pre-trained models and FIR-based data augmentation strategies.

W2V2 model	Data augmentation	LA	DF
XLS-53	-	8.87	7.71
	FIR-NB	4.34	11.27
	FIR-WB	4.98	6.99
XLS-128	-	7.20	5.68
	FIR-NB	**3.54**	6.18
	FIR-WB	7.08	**4.98**

4.2. Results on ASVspoof 2021

Table 3 shows the results obtained for our approach in the LA and DF tracks of the ASVspoof21 database. We experimented with different combinations of W2V2 pre-trained models and low-pass FIR-based data augmentations, both NB and WB filters. Again, the XLS-128 model performs the best in both partitions. Furthermore, the FIR-based augmentations help to improve the discrimination capabilities of our proposal. Particularly, the NB filters are useful for LA scenarios as they emulate traditional telephonic systems, while the DF is more favored by the WB filters. WB filters emulate general audio codecs as the ones in the DF set.

In addition, in Table 4 we compare our approach with other systems in this challenge. As it is shown, our approach outperforms other methods in the DF set and achieves competitive results in LA track, despite being a single system (not an ensemble of classifiers). The W2V2 features show robustness to the varied speech content in the DF set, allowing to highly reduce EER compared to other systems. On the other hand, compared with [15], our approach exploits the representations of the different transformer layers, allowing comparative results while using a simpler downstream model. Moreover, the classifier can be effectively adapted using data augmentations techniques, while the W2V2 pre-trained model remains as a general feature extractor. It can be interesting in practical applications to save computational resources.

Table 4. Comparative EER (%) results of our proposed method with participant systems in the ASVspoof 2021 challenge and other self-supervised approaches.

System	LA	DF
LCNN+ResNet+RawNet [22]	**1.32**	15.64
GMM+LCNN (Ensemble) [24]	3.62	18.30
ECAPA-TDNN (Ensemble) [25]	5.46	20.33
ResNet (Ensemble) [26]	3.21	16.05
W2V2 (fixed)+LCNN+BLSTM [15]	10.97	7.14
W2V2 (finetuned)+LCNN+BLSTM [15]	7.18	5.44
Proposed system	3.54	**4.98**

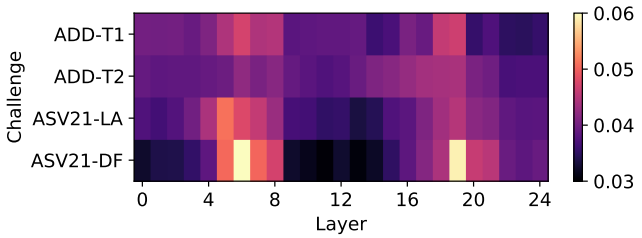

Fig. 2. Visualization of the weight values α_l for our best-proposed approach at each challenge and track.

Finally, Figure 2 draws the value of the weights α_l for our best systems in the tracks of ADD 2022 and ASVspoof21 challenges. As it can be observed, the classifier uses the information from the different transformer layers to detect spoofed audio, using different layer weights depending on the scenario considered.

5. CONCLUSION

In this work, we have presented our proposed system for the 2022 ADD challenge, which is based on the pre-trained W2V2 self-supervised architecture. Our approach exploits the contextualized representations at the different transformer layers in a finetuned classification model to detect spoofed speech. While the W2V2 model remains general, we adapted the downstream network to each scenario through data augmentation techniques and adaptation data. Our proposed method shows competitive results in both the ASVspoof 2021 and 2022 ADD challenges, which represent challenging realistic scenarios with synthesized spoofed audio. Thus, our system ranked first and fourth position in tracks 1 and 2 of 2022 ADD challenge, respectively. As future work, we will test other self-supervised models as well as additional data augmentation techniques.

6. REFERENCES

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

[2] 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*, vol. 80, no. 21, pp. 32725–32762, 2021.

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

[4] Z. Wu et al., "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech 2015*, 2015, pp. 2037–2041.

[5] M. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[6] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "An investigation of spoofing speech detection under additive noise and reverberant conditions.," in *Proc. Interspeech 2016*, 2016, pp. 1715–1719.

[7] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Trans. on Audio, Speech, and Lang. Processing*, vol. 27, no. 12, pp. 1985–1999, 2019.

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

[9] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, "An initial investigation for detecting partially spoofed audio," in *Proc. Interspeech 2021*, 2021, pp. 4264–4268.

[10] J. Yi et al., "ADD 2022: The first audio deep synthesis detection challenge," in *Proc. ICASSP*, 2022.

[11] A. Baevski, A. Zhou, H.and Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[12] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," *arXiv preprint arXiv:2110.05777*, 2021.

[13] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[14] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vec feature for spoofing speech detection," in *Proc. Interspeech*, 2021, pp. 4269–4273.

[15] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

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

[17] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[18] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[19] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[20] H. Shi, Y.and Bu, X. Xu, S. Zhang, and M. Li, "AISHELLl-3: A multi-speaker mandarin TTS corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[21] J. Yamagishi, "English multi-speaker corpus for CSTR voice cloning toolkit," 2012.

[22] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 ASVspoof Workshop*, 2021, pp. 61–67.

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

[24] R. K. Das, "Known-unknown Data Augmentation Strategies for Detection of Logical Access, Physical Access and Speech Deepfake Attacks: ASVspoof 2021," in *Proc. 2021 ASVspoof Workshop*, 2021, pp. 29–36.

[25] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021," in *Proc. 2021 ASVspoof Workshop*, 2021, pp. 75–82.

[26] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in *Proc. 2021 ASVspoof Workshop*, 2021, pp. 89–93.