INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A conformer-based classifier for variable-length utterance processing in anti-spoofing

Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, Antonio M. Peinado

University of Granada, Spain

erosrosello@ugr.es, agomezalanis@ugr.es, amgg@ugr.es, amp@ugr.es

Abstract

The success achieved by conformers in Automatic Speech Recognition (ASR) leads us to their application in other domains, such as spoofing detection for automatic speaker verification (ASV), where the conformer self-attention mechanism might effectively model and detect the artifacts introduced in spoofed speech signals. Also, conformers can naturally handle the variable duration of speech utterances. However, as with transformers, the conformer performance may degrade when trained with limited data. To address this issue, we propose utilizing conformers in conjunction with self-supervised learning, specifically leveraging a pre-trained model called wav2vec 2.0, which is pre-trained using a substantial amount of bonafide data. Our experimental results demonstrate that our proposed method achieves one of the best results in the recent ASVspoof 2021 logical access (LA) and deep fake (DF) databases.

Index Terms: Spoofing detection, deep learning, conformers, Deep fake detection, wav2vec 2.0.

1. Introduction

Voice biometrics systems authenticate the identity of a speaker through their voice using automatic speaker verification (ASV) technology. Recent advancements in deep neural networks have significantly improved the performance of ASV systems [1]. However, these systems are vulnerable to malicious attacks, including voice synthesis or text-to-speech (TTS), voice conversion (VC), replay, and impersonation attacks, which can compromise their security [2]. This paper focuses on voice conversion and synthesis attacks, usually associated with logical access (LA) to the biometric system and audio deep fakes. Both use high-quality synthesized speech generated by modern VC and TTS systems.

Thus, during the last few years, the scientific community has paid attention to the development of anti-spoofing techniques to detect spoofing attacks on ASV systems [2]. Several evaluation campaigns have been organized on this topic, including ASVspoof 2015 [3], 2017 [4], 2019 [5], and 2021 [6], focusing on LA attacks (TTS and VC), physical access attacks (replay attacks), and speech deep fake detection. These challenges highlight the importance of developing technologies that are robust to different types of attacks and environmental conditions, with deep neural networks (DNNs) being the most effective approach [7, 8, 9, 10, 11, 12].

The attention-based encoder-decoder architecture is a powerful tool for modeling speech dependencies. The transformer architecture leverages self-attention to establish global dependencies between the input and output sequences [13]. Recent studies have shown that their combination with convolutions allows to model both local and global dependencies [14], with conformers demonstrating exceptional performance in ASR [15]. However, these models require large amounts of training data [16, 17, 18]. To address this, we propose a combination of conformers and self-supervised models, whose potential has already been proven in anti-spoofing tasks [19], in order to mitigate the requirement of large training datasets.

Self-supervised learning (SSL) has gained attention due to its ability to provide pre-trained models that generalize well across different tasks with limited labeled data [17]. Several SSL speech models, such as auto-regressive predictive coding [20], wav2vec [21] and HuBERT [22] have shown promising results for speech processing tasks. HuBERT and wav2vec 2.0 are popular SSL approaches applied in ASR [23, 24], mispronunciation detection [25], speaker recognition [26], and emotion recognition [27]. This type of model is currently being explored for anti-spoofing. The authors in [19] explored the complementary benefits of data augmentation (DA) and achieved state-of-the-art results (in combination with AASIST [28]). Thus, we have applied the wav2vec 2.0 XLS-R (0.3B) model [23] as a front-end technique, which has been pre-trained on a diverse corpus of speech data over 120 different languages from various regions worldwide.

The main contributions of this work are: (i) an improved conformer-based architecture for anti-spoofing; (ii) a novel classifier that combines a self-supervised model (wav2vec 2.0) with a conformer encoder; (iii) the proposed architecture allows processing variable-length utterances for spoofing detection. This last capability has the advantage of not disregarding any information from the input speech signal, in contrast with current state-of-the-art techniques, which apply cropping and concatenation strategies for training and evaluation with fixed-length speech sequences [19, 29].

The rest of this paper is organized as follows. Section 2 presents our proposed model, including the adaptation that enables the use of conformers for classification. In Section 3, we describe our experimental setup and the DA techniques employed in the experiments. Then, Section 4 describes the experimental results obtained with multiple variants of the proposed architecture. We compare our model with other state-of-the-art anti-spoofing systems from the literature in Section 5 and summarize our research findings in Section 6.

2. Proposed Method

In this section, we provide an overview of our model architecture. First, we briefly detail the Wav2Vec 2.0 model employed as a feature extractor as well as its fine-tuning process. Then, we describe the adaptation of the conformer for classification tasks such as spoofing detection.

10.21437/Interspeech.2023-1820

2.1. Wav2vec-based feature extraction

The pre-training of the wav2vec 2.0 XLS-R (W2V) model has been carried out solely with bonafide data. As stated in [30], to improve the detection of spoofing attacks, fine-tuning with in-domain bonafide and spoofed training data is necessary. To achieve this, we optimize the pre-trained model jointly with the conformer encoder (detailed in the next subsection) through back-propagation during training. In Section 4, we examine the impact of utilizing this fine-tuned W2V model as opposed to not using it, or using it without fine-tuning. This analysis provides insight into the effectiveness of pre-trained fine-tuned models and their impact on the model's overall performance.

The raw input waveform signal sequence, $s(n)$ $(n = 0, \ldots, N-1)$, is first processed by the W2V model to extract an output sequence of feature vectors, $\boldsymbol{O}' = (\boldsymbol{o}'_i | i = 0, \ldots, M-1)$ with $\boldsymbol{o}'_i \in \mathbb{R}^{D'}$. The convolutional neural network (CNN) inside the W2V model converts the input into a hidden feature sequence, which is then converted by the transformer network into the output sequence $\boldsymbol{O}'$. The ratio between N and M is dictated by the CNN stride (20 ms in our case). The vector sequence $\boldsymbol{O}'$ is finally transformed, through a fully connected layer (FC) plus a batch normalization (BN) and a SeLU activation function, into a final feature sequence $\boldsymbol{O} = (\boldsymbol{o}_i | i = 0, \ldots, M-1)$, with $\boldsymbol{o}_i \in \mathbb{R}^D$ for $i = 0, \ldots, M-1$, where D is the dimension of the conformer encoder, that is,

$$\boldsymbol{O} = \mathrm{BN}(\mathrm{SeLU}(\mathrm{FC}(\boldsymbol{O}'))). \tag{1}$$

2.2. Conformer encoder adaptation to classification

The conformer encoder is typically used for solving sequence-to-sequence problems (e.g., in ASR). In order to adapt the conformer to a classification task, we use a learnable classification token that is initialized randomly, as in [16, 18]. This token allows the conformer to be customized for a specific classification task.

As shown in Figure 1, we prepend the token, denoted as $\boldsymbol{x}_{class}$, to the reduced output sequence of the W2V model. In particular, if $\boldsymbol{X}^0 = (\boldsymbol{x}^0_j | j = 0, \ldots, M)$ is the input sequence of the first conformer block, then we will note

$$\boldsymbol{x}^0_0 = \boldsymbol{x}_{class}, \quad \boldsymbol{x}^0_j = \boldsymbol{o}_{j-1} \ \text{ for } \ 1 \leq j \leq M, \tag{2}$$

where $\boldsymbol{x}^0_j \in \mathbb{R}^D$, for $j = 0, \ldots, M$. Now, this combined sequence is then processed by the conformer encoder blocks. Denoting the output of the l-th conformer block as the sequence $\boldsymbol{X}^l$ and its input as the sequence $\boldsymbol{X}^{l-1}$, then

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}^{l-1} + \frac{1}{2}\,\mathrm{FFN}(\boldsymbol{X}^{l-1}), \tag{3}$$

$$\boldsymbol{X}' = \tilde{\boldsymbol{X}} + \mathrm{MHSA}(\tilde{\boldsymbol{X}}) + \mathrm{Conv}(\tilde{\boldsymbol{X}} + \mathrm{MHSA}(\tilde{\boldsymbol{X}})), \tag{4}$$

$$\boldsymbol{X}^l = \mathrm{Layernorm}\left(\boldsymbol{X}' + \frac{1}{2}\,\mathrm{FFN}(\boldsymbol{X}')\right), \tag{5}$$

where FFN refers to a Feed Forward module, MHSA refers to a Multi-Head Self-Attention module, and Conv refers to a Convolution module. Finally, we take the state of the classification token at the output of the last conformer block (i.e., we only use the first element of the output sequence of the last conformer block $\boldsymbol{x}^L_0$) as the final audio representation. This representation is fed to a linear layer to classify the input speech signal as either bonafide or spoof. Thus, the classification token is trained to capture the relevant characteristics of the audio signal to determine whether the input signal is genuine or not.

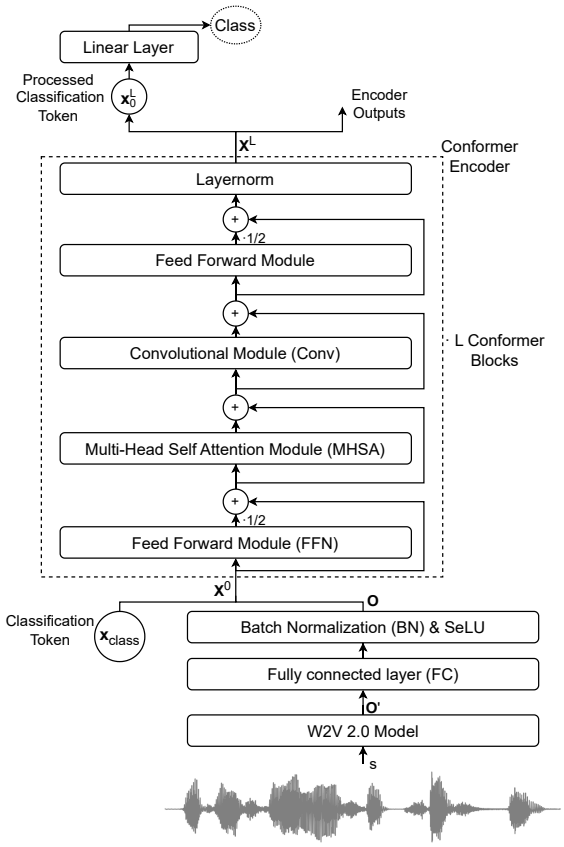

Figure 1: *Block diagram of the proposed architecture.*

By using this approach, we can effectively use the conformer encoder, a sequence-to-sequence model, for a classification task such as anti-spoofing. Moreover, as we will see in more detail in Section 4, by utilizing the pre-trained W2V model in conjunction with the conformer encoder we can surpass the need for large training datasets that conformers usually require.

3. Experimental Setup

In this section, we describe the datasets and evaluation metrics employed for our experiments. Also, we describe the training details and the data augmentation (DA) techniques applied.

3.1. Dataset and evaluation metrics

To evaluate our proposed method, we conducted experiments on two subsets of the ASVspoof 2021 database: logical access and deep fake [6]. The ASVspoof 2019 logical access training and development partitions were used for training and validation [5]. The databases consist of bonafide and spoofed speech generated using TTS and VC systems. The LA subset contains codec and transmission variability, while the DF subset introduces compression variability. Only six known attacks (2 VC-based and 4 TTS-based) are present in the training and development sets, while unseen attacks are present in the evaluation datasets [6].

We used the pooled equal error rate (EER) [31] as a primary metric and, for the LA evaluation dataset, we also report the minimum normalized tandem detection cost function (t-DCF)

[32] results.

3.2. Data Augmentation

To add variability to the existing training data, we made use of the RawBoost DA tool [33], which adds variation in the form of linear and non-linear convolutive noise, impulsive signal-dependent additive noise, and stationary signal-independent additive noise. Full details can be found in [33].

The DA applied in this work uses the same configuration and parameters as reported in [33]. For the LA database, a combination of linear and non-linear convolutive noise and impulsive signal-dependent additive noise strategies are used, while for the DF database, stationary signal-independent additive noise with random coloration is added.

3.3. Implementation details

Unless otherwise specified, we used a fixed-size input audio signal, which is cropped or extended, repeating its content, into segments of approximately 4 seconds duration. We employed the standard Adam optimizer with a learning rate of 10^{-6}, a weight decay of 10^{-4}, and a batch size of 20, in order to minimize a weighted cross-entropy loss function. The W2V model was implemented by using the Fairseq project toolkit [34].

The fully connected layer following the wav2vec 2.0 model has 144 output dimensions, which matches the dimension used for the conformer blocks. We utilized 4 conformer blocks with 4 heads and a kernel size of 31, making a total of 2.4M parameters for the conformer encoder. Also, we applied early stopping in order to end the training process when the weighted cross-entropy in the validation set did not improve across 7 iterations. The final system results from uniformly averaging the model weights of the top-5 epochs, i.e. the 5 epochs that demonstrated the best performance on the validation set. For more details, please refer to [35].

4. Results

This section begins with some preliminary results to discuss the influence of the fine-tuned Wav2vec 2.0 model and the chosen length of fixed-size inputs. Following that, we evaluate the performance of our proposed model in both the LA and DF ASVspoof 2021 evaluation datasets.

4.1. Preliminary results

Table 1 summarizes the impact of using the fine-tuned Wav2Vec 2.0 model. No data augmentation is considered here. We can see that replacing the Wav2Vec 2.0 model with a short time Fourier transform (STFT) using a Blackman analysis window of 30 ms length with 15 ms of frameshift and 256 frequency bins performs poorly, achieving only a 12.95% and 28.64% EER in the LA and DF evaluation sets, respectively. However, the use of the pre-trained Wav2Vec 2.0 model improves the EER to 7.65% and 8.67% in LA and DF, respectively. Fine-tuning (FT) the Wav2Vec 2.0 model further improves performance, achieving a 2.30% in LA and a 3.28% in DF, respectively. This last result suggests that the Wav2vec 2.0 model alleviates the need for a large dataset to train the conformer, making it able to achieve state-of-the-art results without using extra spoofing training data.

All these experiments were performed with a fixed speech sequence duration of approximately 4 seconds. However, as shown in Table 2, the selected input length, at which inputs are cropped or extended, can significantly impact the model's performance. Our results (with W2V + conformer) suggest that excessive cropping or length extension can lead to a lower performance. In this regard, it is worth noticing that the average utterance lengths (measured in seconds) for the training, LA evaluation, and DF evaluation sets are 3.43, 2.72, and 3.06 seconds, respectively, with standard deviations of 1.42, 1.30, and 1.26, respectively. Specifically, cropping results in the loss of available data for training or evaluation, leading to significant performance degradation. Length extension is also harmful, although not so much as cropping. These results reveal the importance of variable-length utterance processing to mitigate duration mismatch in both training/evaluation and, at the very least, the evaluation/use stage.

Table 1: *Comparison of different forms of processing the raw audio inputs. FT refers to fine-tuned.*

	LA		DF
Model	EER	min t-DCF	EER
STFT + conformer	12.95	0.4899	28.64
W2V (no FT) + conformer	7.65	0.3807	8.67
W2V + conformer	2.30	0.2491	3.28

Table 2: *Comparison of training and evaluating with different fixed size inputs. The first column indicates the approximate duration to which the audios were cropped or extended.*

	LA		DF
Length	EER	min t-DCF	EER
2 seconds	7.80	0.3987	5.26
4 seconds	2.30	0.2491	3.28
6 seconds	6.18	0.3569	4.31

In the following subsections, we report results from training with fixed-size data and evaluate using both fixed and variable-length inputs. We also report results from training with variable-length data to assess whether this (more costly) approach may yield better results.

4.2. Results on the ASVspoof 2021 Logical Access Corpus

The performance of our proposed system on the LA evaluation dataset is reported in Table 3. The results show that using variable size inputs with data augmentation techniques improves the system performance. Specifically, our system achieves an EER of 1.38% when using fixed-size inputs, improving up to

Table 3: *EER and min t-DCF results for the ASVspoof 2021 LA evaluation set, for variable or fixed length in training and evaluation. DA refers to the use data augmentation technique.*

Variable Length Train	Variable Length Eval	DA	Pooled EER (%)	Pooled min t-DCF
×	×	×	2.30	0.2491
×	×	✓	1.38	0.2216
×	✓	×	2.82	0.2632
×	✓	✓	**0.97**	**0.2116**
✓	✓	×	5.14	0.3297
✓	✓	✓	**0.87**	**0.2092**

Table 4: *EER results for the ASVspoof 2021 DF evaluation set for variable and fixed length in training and evaluation.*

Variable Length Train	Variable Length Eval	DA	Pooled EER (%)
×	×	×	3.28
×	×	✓	**2.27**
×	✓	×	3.47
×	✓	✓	**2.58**
✓	✓	×	9.91
✓	✓	✓	7.36

0.97% when evaluated with variable-length utterances. Moreover, training the model with variable-length utterances yields even better results, reducing the EER to 0.87% in the LA scenario. Interestingly, it is worth noticing that the opposite occurs when no data augmentation techniques are used. In this case, the equal error rate increases from 2.30% to 5.14% when variable length inputs are employed. We hypothesize that this may be because the model is overfitting, so DA helps to avoid it. Thus, when we add variability to the training data using DA techniques, the full variable-length speech sequence can be exploited without overfitting.

4.3. Results on the ASVspoof 2021 Deep Fake Corpus

Table 4 shows the experimental results evaluated in the ASVspoof 2021 DF evaluation dataset. In this case, the use of variable-size utterances cannot obtain further improvements, either combined with data augmentation or not. However, when we use variable length only for evaluation, there is little difference with respect to the results with fixed-length inputs. When combined with data augmentation, the training and evaluation with fixed-length inputs achieve an EER of 2.27%, which is, to the best of our knowledge, the best result in the literature. Evaluation with variable-length sequences achieves a similar EER of 2.58%, which also outperforms other state-of-the-art systems. However, training and evaluating with variable-length utterances increases the EER to 7.36%. This result suggests a possible mismatch between the training dataset and the DF evaluation set, but further investigation is needed to confirm this. As in the LA case, the addition of DA techniques always yields better results than when no DA techniques are applied.

5. Discussion

Table 5 presents a comparison of the performance of our proposed anti-spoofing system with other relevant systems from the literature on the ASVspoof 2021 evaluation sets for both the logical access (LA) and deep fake (DF) scenarios. The first system is a baseline from the ASVspoof 2021 challenge that uses a RawNet2 architecture [29]. Additionally, we compare our model with two of the best-performing models in the LA and DF scenarios from the ASVspoof 2021 challenge [11, 36]. The fourth model uses a self-supervised approach, similar to our proposal, in combination with the AASIST model [19, 28].

Our proposed model (W2V+Conformer with variable length during evaluation) achieves state-of-the-art results in the LA scenario, slightly outperforming the Wav2vec 2.0 model with the AASIST architecture [19]. In the DF scenario, our model also outperforms all other models from the literature, achieving an EER of 2.58%, which significantly outperforms the W2V+AASIST model (3.69% EER).

Table 5: *Comparison of our system with other systems in the literature on the ASVspoof 2021 LA and DF evaluation sets in terms of EER on both sets and min t-DCF on LA. To provide a fair comparison, we used the same evaluation protocol and datasets as specified in the ASVspoof 2021 challenge.*

	LA		DF
Model	EER	min t-DCF	EER
Rawnet2 [29]	9.50	0.4257	22.38
STC Antispoofing [11]	1.32	0.2177	15.64
Pindrop Labs [36]	3.21	0.2608	16.05
W2V + AASIST*[19]	1.00	0.2120	3.69
W2V + conformer	**0.97**	**0.2116**	**2.58**

* Averaged result from 3 different training sessions.

Overall, the results demonstrate that our proposal is effective in detecting spoofed audio in both the LA and DF scenarios. Furthermore, the model's ability to handle variable-length inputs during evaluation without altering the audio makes it more practical for real-world applications, where audio inputs may vary in length.

6. Conclusions

In this paper, we present a novel anti-spoofing system that can handle variable length inputs which has achieved outstanding performance in both the LA and DF scenarios. Our adaptation of the conformer encoder for spoofing detection has achieved state-of-the-art results in the LA scenario and outperformed all other models from the literature in the DF scenario.

One of the novel features of our model is its ability to handle variable length inputs without altering the audio, which might be important for practical applications since audio duration alteration can lead to a loss of information, thus affecting negatively the performance. We have also demonstrated that the selected length at which audios are cropped or extended has a significant impact on the model performance, highlighting the need of using models that can handle variable sequence-length utterances.

Furthermore, we have shown that self-supervised models can enhance the performance of conformers in anti-spoofing without requiring additional training data. Our experiments showed a relative improvement in the performance of over 80% in both the LA and DF scenarios without any data augmentation techniques.

While our proposed model achieved a remarkable performance in the DF scenario, we observed a small decline in performance when we used variable-sized input data during training. This result warrants further investigation into the characteristics of the datasets used for training and evaluation, the effects of different DA techniques, and the impact of the model architecture on its performance. Understanding these factors will help to improve the effectiveness of our model in order to achieve better results in the future.

7. Acknowledgements

This work was supported by the FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades Proyecto PY20_00902 and by the project PID2019-104206GB-I00 funded by MCIN/AEI/10.13039/501100011033.

8. References

[1] N. J. M. S. Mary, S. Umesh, and S. V. Katta, "S-vectors and TESA: Speaker embeddings and a speaker authenticator based on transformer encoder," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 404–413, 2020.

[2] Z. Wu, P. L. D. Leon, C. Demiroğlu, A. Khodabakhsh, S. King, Z. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "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, pp. 768–783, 2016.

[3] Z. Wu, T. H. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015.

[4] T. H. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. W. D. Evans, J. Yamagishi, and K.-A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017.

[5] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. H. Kinnunen, and K.-A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *Proc. Interspeech*, 2019.

[6] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *Proc. ASVspoof 2021 Workshop*, Sep 2021.

[7] A. G. Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. Interspeech*, 2019.

[8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1985–1999, 2019.

[9] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1–1, 05 2018.

[10] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "A kernel density estimation based loss function and its application to ASV-spoofing detection," *IEEE Access*, vol. 8, pp. 108 530–108 543, 2020.

[11] A. Tomilov, A. F. Svishchev, M. Volkova, A. Chirkovskiy, A. S. Kondratev, and G. Lavrentyeva, "STC antispoofing systems for the ASVspoof2021 challenge," *Proc. ASVspoof 2021 Workshop*, 2021.

[12] A. Gomez-Alanis, J. Gonzalez Lopez, and A. Peinado, "Ganba: Generative adversarial network for biometric anti-spoofing," *Applied Sciences*, vol. 12, p. 1454, 01 2022.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[14] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[15] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech*, Oct 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Jun. 2019, pp. 4171–4186.

[18] E. Rosello, A. Gomez-Alanis, M. Chica, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "On the application of conformers to logical access voice spoofing attack detection ," in *Proc. Iberspeech*, 2022, pp. 181–185.

[19] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop*, 2022.

[20] Y.-A. Chung, Y. Belinkov, and J. Glass, "Similarity analysis of self-supervised speech representations," *Proc. ICASSP*, Jun 2021.

[21] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech*, Sep 2019.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, p. 3451–3460, 2021.

[23] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *Proc. Interspeech*, Sep 2022.

[24] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS*, vol. 33, 2020.

[25] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection," *Proc. Interspeech*, Aug 2021.

[26] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *Proc. Interspeech*, Aug 2021.

[27] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech*, 2021.

[28] J. weon Jung, H.-S. Heo, H. Tak, H. jin Shim, J. S. Chung, B.-J. Lee, H. jin Yu, and N. W. D. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *Proc. ICASSP*, 2022.

[29] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. W. D. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," *ICASSP*, 2021.

[30] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *The Speaker and Language Recognition Workshop (Odyssey 2022)*, Jun 2022.

[31] N. Brümmer and E. De Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," *Documentation of BOSARIS toolkit*, vol. 24, 2011.

[32] T. H. Kinnunen, K.-A. Lee, H. Delgado, N. W. D. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *The Speaker and Language Recognition Workshop*, 2018.

[33] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," *ICASSP*, May 2022.

[34] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[35] Y. Gao, C. Herold, Z. Yang, and H. Ney, "Revisiting checkpoint averaging for neural machine translation," in *AACL-IJCNLP 2022*. Association for Computational Linguistics, Nov. 2022, pp. 188–196.

[36] T. Chen, E. el Khoury, K. Phatak, and G. Sivaraman, "Pindrop labs' submission to the ASVspoof 2021 challenge," *Proc. ASVspoof 2021 Workshop*, 2021.