On the Application of Conformers to Logical Access Voice Spoofing Attack Detection

Eros Rosello[1]*, Alejandro Gomez-Alanis*[1]*, Manuel Chica*[1]*, Angel M. Gomez*[1]*, José A. Gonzalez-Lopez*[1]*, Antonio M. Peinado*[1]

[1]University of Granada, Spain

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

Abstract

Biometric systems are exposed to spoofing attacks which may compromise their security, and automatic speaker verification (ASV) is no exception. To increase the robustness against such attacks, anti-spoofing systems have been proposed for the detection of spoofed audio attacks. However, most of these systems can not capture long-term feature dependencies and can only extract local features. While transformers are an excellent solution for the exploitation of these long-distance correlations, they may degrade local details. On the contrary, convolutional neural networks (CNNs) are a powerful tool for extracting local features but not so much for capturing global representations. The conformer is a model that combines the best of both techniques, CNNs and transformers, to model both local and global dependencies and has been used for speech recognition achieving state-of-the-art performance. While conformers have been mainly applied to sequence-to-sequence problems, in this work we make a preliminary study of their adaptation to a binary classification task such as anti-spoofing, with focus on synthesis and voice-conversion-based attacks. To evaluate our proposals, experiments were carried out on the ASVspoof 2019 logical access database. The experimental results show that the proposed system can obtain encouraging results, although more research will be required in order to outperform other state-of-the-art systems.

Index Terms: Spoofing detection, deep learning, conformers.

1. Introduction

Automatic speaker verification systems (ASV) are designed to determine whether a given speech signal proceeds from a given individual [1]. As the interest in this technology grows, due to its commercial application [2], the same happens to the concerns about its security. ASV systems can be fooled by presenting speech samples resembling the voice of a genuine user. We can distinguish four types of spoofing attacks [3]: (i) replay (i.e. replay of a pre-recorded voice of a genuine user), (ii) impersonation (i.e. mimicking a genuine voice), and (iii) text-to-speech systems (TTS) or (iv) voice conversion (VC) systems to generate artificial speech resembling the voice of a legitimate user. In this paper, we will focus on the detection of logical access (LA) attacks, i.e., TTS- and VC-based spoofing attacks.

Spoofing detection for ASV has become an important topic for researchers in recent years as evidenced by the organization of several evaluation campaigns (challenges) on this specific topic: (i) ASVspoof 2015 [4], which focused on LA attacks (TTS and VC); (ii) ASVspoof 2017 [5], which focused on physical access (PA) attacks (replay attacks) under noisy environments; (iii) ASVspoof 2019 [6], which addressed both the detection of LA attacks generated with the latest TTS and VC technologies and simulated replay attacks under different reverberant acoustic conditions; and the most recent one (iv) ASVspoof 2021 [7], which extended both evaluation datasets, LA and PA, the first with a focus on robustness to channel variation and the second with recordings made in real physical environments and adding a speech deepfake detection subchallenge. One of the main conclusions extracted from these challenges is that the use of deep neural networks (DNNs) outperforms other conventional approaches [8]–[20].

Recent works have shown that combining transformers and convolutions may yield better performance than using them individually [21]. Thus, CNNs can efficiently extract local features and the attention mechanisms can learn content-based global interactions. In particular, conformers have shown excellent performance in automatic speech recognition (ASR) tasks [22]. In this preliminary work, we attempt to study how to adapt the conformer to a classification and detection problem such as anti-spoofing for ASV. We hypothesize that global and local correlations are relevant for this classification task, thus bringing out conformers as a powerful modeling tool to be reckoned with.

The rest of this paper is organized as follows. Section 2 summarizes related work and describes the proposed adaptation that allowed us to use the conformer for classification. Section 3 describes the experimental setup used in our experiments. Section 4 describes the results achieved with the different adaptation approaches. Finally, we summarize the conclusions derived from this research in Section 5.

2. Conformer-based spoofing detection

This section starts by describing the conformer encoder employed in sequence-to-sequence tasks such as ASR [22]. Then, we describe the two different approaches that we propose to adapt the conformer to classification tasks.

2.1. Review of Conformers

The attention-based encoder-decoder architecture allows the modeling of dependencies without regard to their distance. Transformers employ a self-attention mechanism to draw global dependencies between input and output [23]. Conformers add convolutional layers into the transformer architecture, thus improving their robustness and accuracy [22].

The conformer encoder architecture is composed of four modules stacked together: a feed forward module, a multi-headed self-attention block that makes use of a sinusoidal positional encoding (a technique borrowed from Transformer-XL [24]), then the convolutional module, and finally another feed forward module and followed by a layer normalization. This

structure is illustrated in Figure 1.

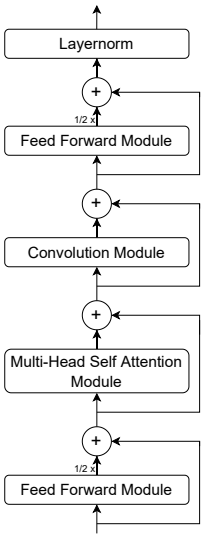

Figure 1: *Diagram depicting the main processing blocks of the Conformer encoder architecture.*

2.2. Conformer adaptation to classification tasks

We propose two alternative approaches to adapt the conformer architecture to classification tasks. The first approach, depicted in Figure 2 and described in subsection 2.2.1, is inspired by the visual transformer (ViT) [25], which uses an extra learnable class embedding that is prepended to the transformer encoder input. The output of that embedding is the one used for classification.

The second approach, depicted in Figure 3 and described in subsection 2.2.2, makes use of the transformer decoder [23] with some adjustments to make it work for classification. This adaptation allows us to use all the outputs of the classification-adapted conformer encoder. To the best of our knowledge, this is the first time that the transformer decoder structure is adapted to a classification task.

2.2.1. Classifier with conformer encoder and classification token

A block diagram of our first approach for classification using the conformer is illustrated in Figure 2. The first layer is a linear layer that receives the log magnitude spectrogram features with shape 400×256 and reduces them to $400 \times d$, where d is the dimension of the conformer and dimension 400 corresponds to the number of frames selected (details in Section 3.2). This layer allows us to reduce the dimension of the input sequence. Then we feed the conformer encoder with the output of this previous layer. Thus, the conformer encoder receives a sequence of vectors $x \in \mathbb{R}^d$.

Similar to ViT [25], we prepend a learnable embedding to the conformer input sequence. We refer to this embedding as classification token. The conformer encoder processes it along with the output of the previous linear layer. Finally, the transformed classification token is forwarded to the final linear layer in order to classify the input speech signal as either bonafide or spoofing.

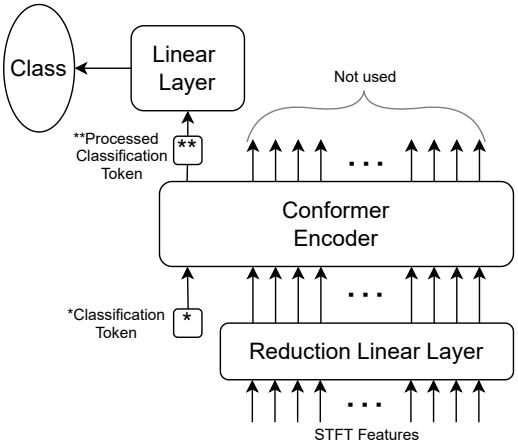

Figure 2: *Overview of the first proposed classifier. We feed the log magnitude spectrogram features to a linear layer and feed the resulting sequence of vectors to a conformer encoder. To perform classification we add an extra learnable classification token to the sequences and use its output to classify.*

2.2.2. Classifier with conformer encoder and decoder

In our second approach (Fig 3), we have placed an adapted transformer decoder at the output of the conformer encoder so that we can use an additional attention mechanism to process all the outputs of the conformer encoder. This decoder has two inputs: the memory input, which is used as the key (K) and value (V) for the attention blocks, and the target input, which is fed to the attention blocks as the query (Q). As previously, the first layer is a linear layer that compresses the log magnitude spectrogram features and feeds the conformer encoder. In turn, its outputs feed the first block of the decoder, a multi-head attention block, as memory input. Regarding the target input we have considered two options:

- In the first option, we take the classification token as target input, identical to the one employed in the previous approach. In this option, the target input will be the learnable classification token that in Figure 3 is labeled as the *Classification Token*.
- In the second option, the attention block (input Q) is fed with the classification token after going through the encoder block. That is, we feed the adapted transformer decoder with the vector that in the first approach is used for classification. In this option, the target input will be the state of the classification token at the output of the conformer encoder. This is depicted in Figure 3 as the *Processed Classification Token*.

The idea behind these two approaches, establishing the classification token as the network target, is to train the decoder so that it is able to predict a new refined classification token which is finally processed by the classification linear layer, the same as in the approach of the previous subsection.

Then, the outputs of the multi-head attention block feed a layer normalization and a feed-forward layer. This structure is very similar to the transformer decoder [23], although we do not use the first self-attention block since our target input is not a sequence. The output of the decoder can be processed by another decoder block or by a linear layer, as depicted in Figure 3, before performing the final classification.

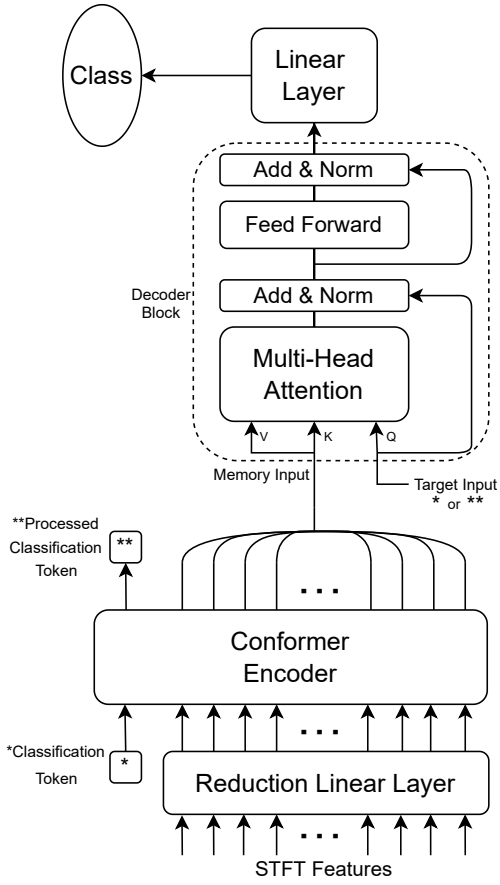

Figure 3: *Overview of our second proposed classifier with one decoder block.*

3. Experimental Setup

In this section, we describe the datasets and evaluation metrics employed for the experiments as well as the spectral analysis performed and the training details.

3.1. Dataset and evaluation metrics

We conducted experiments on the ASVspoof 2019 logical access subset [6]. A summary of their composition in terms of speakers and number of utterances is presented in Table 1.

The database contains bonafide speech and spoofed speech data generated using 17 text-to-speech and voice conversion systems. Six of them are used as known attacks, and the other 11 as unknown attacks. As in the challenge, the training and development sets only contain known attacks, whereas the evaluation set contains 2 known and all 11 unknown spoofing attacks. Among the 6 known attacks, there are 2 voice conversion systems and 4 TTS systems. Voice conversion systems use neural-network-based and spectral-filtering-based approaches [26]. TTS systems use either waveform concatenation or neural-network-based speech synthesis using a conventional source-filter vocoder [27] or a WaveNet-based vocoder [28]. The 11 unknown systems comprise 2 voice conversion systems, 6 TTS, and 3 hybrid systems and were implemented with waveform generation methods including classical vocoding, GriffinLim [29], generative adversarial networks [30], neural waveform models [31], waveform concatenation, waveform filtering [32], spectral filtering, and their combination.

Table 1: *Structure of the ASVspoof 2019 logical access data corpus divided by training, development, and evaluation sets [6].*

	#speakers		#utterances	
Subset	Male	Female	Bona fide	Spoof
Training	8	12	2580	22800
Development	8	12	2548	22296
Evaluation	21	27	7355	63882

We used the pooled equal error rate (EER) [33] as the primary metric and also report results in terms of the minimum normalized tandem detection cost function (t-DCF) [34].

3.2. Spectral Analysis

Speech signals were analyzed using a Blackman analysis window of 25 ms length with 10 ms of frame shift. Log magnitude spectrogram features (STFT) with 256 frequency bins (512-points FFT) were obtained to feed the neural network.

We truncated the spectrum along the time axis with a fixed size of $T = 400$ frames in order to feed the first layer of the neural network. During this procedure, short utterances were extended by repeating their contents if necessary to match the required length.

3.3. Training Details

The network is trained using the ASVspoof 2019 LA training partition to minimize a cross-entropy loss function. We used the Adam optimizer with a fixed learning rate of $6 \cdot 10^{-5}$. The batch size is set to 128.

According to some preliminary experiments, training computational burden and overfitting can be reduced by selecting 2100 utterances of the development set for validation, for which 0% EER is achieved sooner than for the whole set. Thus, the results in the next section will be given directly on the evaluation set. To further prevent overfitting we use dropout in the decoder and each conformer module, with a 0.35 dropout for the convolutional and attention modules, and a 0.3 dropout for the feed-forward module and the decoder. Also, early stopping was applied when no improvement of the EER of the validation sub-set was obtained after nine epochs.

We also trained the best-performing models with another loss function for comparison. We used the one-class softmax (OC-Softmax) [35] loss function with hyper-parameters: $\alpha = 20$ and $m = 0.9$. For this loss function, we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to update the weights in the model. We used stochastic gradient descent (SGD) optimizer for parameters in the loss functions. The learning rate is initially set to $3 \cdot 10^{-4}$ with a 50% decay every 10 epochs.

4. Results

This section starts with the results obtained with our first adaptation of the conformer to classification tasks (W/O-DECODER), which uses a class embedding similar to the one used in vision transformers [25]. After that, Section 4.2 describes the results of the second proposed approach (W-DECODER) with different options for the target input. Finally, Section 4.3 describes the results achieved with the OC-softmax loss function [35]. Also, it must be pointed out that the goal of this section is not so much

to achive competitive results as to study what performance can be achieved.

4.1. Results with encoder and classification token

Table 2 shows the results of our first proposed approach. We can see how the dimension of the model and the number of attention heads affect the number of parameters of the model. We decided to analyze first how the dimension of the model impacts its performance. First, we can observe that the dimension of the model has the greatest impact on the number of parameters and that a model with a high dimension does perform poorly. This seems to indicate that the model is not able to generalize well for the new spoofing attacks in the evaluation dataset. On the other hand, a model with a very low dimension works better than the oversized one, but is not as good as the models with medium size dimension. After adjusting the dimension of the model we tried different numbers of attention heads and anti-spoofing classes, i.e., two classes (bonafide and spoof) or seven classes (the six types of attack in the training data plus bonafide).

Table 2: *Results of the W/O-DECODER approach and model size.*

Dim	Attention Heads	Classes	Param	EER (%)	t-DCF
256	8	7	2.6M	10.09	0.2563
128	4	7	0.8M	7.80	0.1653
100	8	7	0.79M	9.14	0.1698
100	8	2	0.79M	8.25	0.1855
100	4	7	0.59M	**7.56**	0.1668
100	4	2	0.59M	8.93	0.1884
64	4	7	0.37M	8.58	0.1686

The best results are obtained with the medium-size models with 4 attention heads and dimensions 128 and 100 yielding an EER of 7.80% and 7.56%, respectively. This result shows that conformers are an effective tool for discriminating between spoofed and bonafide voices. Thus, in the next experiments, we will employ a conformer encoder with dimension size of 100 and 4 multi-attention heads, as well as 7 training classes.

4.2. Results with encoder and decoder

The results of the second approach when feeding the decoder with the token directly (W-DECODER1) are shown in Table 3. This approach has a slightly worst performance in terms of EER than the previous one. However, when we take as target input the state of the classification token after going through the conformer (W-DECODER2) we achieve a slight improvement in terms of both EER and t-DCF with respect to the approach which only applies the encoder as shown in Table 4.

These results show that our new approach to adapting an encoder/decoder structure to classification can achieve even better results than the well-known method of using the encoder structure with an extra learnable class embedding [25].

4.3. Results with OC-Softmax loss function

All the previous results are the ones obtained using the classical softmax loss function. We also trained our best two models using the OC-Sofmax loss function [35]. The results, which are shown in Table 5, are better than the ones obtained with the models which have the same number of parameters but trained with only 2 classes. However, they do not outperform the ones trained with 7 classes using the classical softmax loss.

Table 3: *Results of the W-DECODER1 approach.*

Decoder Dimension	Attention Decoder Heads	Decoder Blocks	EER (%)	t-DCF
100	10	2	8.851	0.1670
100	10	1	10.005	0.2051

Table 4: *Results when the target input is the state of the classification token after going through the conformer and the dimension size of the decoder is 100 (W-DECODER2).*

Attention Decoder Heads	Decoder Blocks	#Classes	EER (%)	t-DCF
10	2	7	**7.517**	0.1531
10	2	2	8.197	0.1787
10	1	7	8.835	0.1616

Table 5: *Results of the two approaches trained using OC-Softmax. The parameters selected are the ones that yielded the best performance with the previous loss function.*

Model	EER (%)	t-DCF
W/O-DECODER	7.695	0.1697
W-DECODER2	8.867	0.1801

5. Conclusions and future work

In this paper, we have explored the use of conformers for the classification task involved by ASV anti-spoofing. We have shown that the conformer encoder can obtain encouraging results but not better than those of other state-of-the-art anti-spoofing systems, like the ones obtained by the gated recurrent convolutional neural network (GRCNN) in [10] which achieved a 3.85% in EER and 0.0952 in t-DCF (notice however that GRCNN includes phase features and time modeling). We also introduced a new approach for adapting the seq2seq structure of encoder and decoder to classification, showing that this new approach can provide a slight improvement over the previous that only uses the encoder structure with an extra learnable embedding. We theorize this is due to the use of all the outputs of the encoder for classification and not only the output of the classification token. In future work, it would be worthwhile to investigate a sliding-window encoder-decoder approach in order to fully exploit the sequential nature of speech. Under this new approach, the decoder could lead to further reductions of EER since the decoder will be able to process its previous outputs similarly to a sequence-to-sequence model.

6. 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.

7. References

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

[2] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Automatic speaker recognition for mobile forensic applications," *Mobile Inf. Syst.*, vol. 2017, 2017, Art. no. 6986391.

[3] Z. Wu *et al* "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 768–783, Apr. 2016.

[4] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah and Aleksandr Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 2037–2041.

[5] T. Kinnunen, Md Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi and K. Aik Lee "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017.

[6] M. Todisco, X. Wang, V. Vestman, Md Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen and K. Aik Lee "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *Proc. INTERSPEECH*, September 2019.

[7] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans and Hector Delgado "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, September 2021.

[8] Alejandro Gomez-Alanis, Antonio Peinado, Jose A. Gonzalez Lopez and Angel Gomez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection," *Interspeech*, Sep. 2019, pp. 1068–1072.

[9] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[10] Alejandro Gomez-Alanis and Antonio M. Peinado and Jose A. Gonzalez and Angel Manuel Gomez, "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1985-1999, 2019.

[11] A. Gomez-Alanis and J. A. Gonzalez-Lopez and S. P. Dubagunta and A. M. Peinado and M. Magimai.-Doss, "On Joint Optimization of Automatic Speaker Verification and Anti-spoofing in the Embedding Space," in *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579-1593, 2021.

[12] A. Gomez-Alanis and J. A. Gonzalez-Lopez and A. M. Peinado, "A Kernel Density Estimation Based Loss Function and its Application to ASV-Spoofing Detection," in *IEEE Access*, vol. 8, pp. 108530-108543, 2020.

[13] A. Gomez-Alanis and J. A. Gonzalez-Lopez and A. M. Peinado, "GANBA: Generative Adversarial Network for Biometric Anti-Spoofing," in *Applied Sciences*, vol. 12, no. 3, 1454, 2022.

[14] Alejandro Gomez-Alanis and Antonio M. Peinado and Jose A. Gonzalez, "Adversarial Transformation of Spoofing Attacks for Voice Biometrics," in *Proc. Iberspeech*, pp. 255-259, 2021.

[15] Tomilov, A., Svishchev, A., Volkova, M., Chirkovskiy, A., Kondratev, A., Lavrentyeva, G., "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021 61-67, doi: 10.21437/ASVSPOOF.2021-10.

[18] Alejandro Gomez-Alanis, Antonio Peinado, Jose A. Gonzalez Lopez and Angel Gomez, "A deep identity representation for noise robust spoofing detection," in *Proc. Interspeech*, Sep. 2018, pp. 676–680.

[19] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. Interspeech*, Sep. 2019, pp. 1078–1082.

[20] Alejandro Gomez-Alanis, Antonio Peinado, Jose A. Gonzalez Lopez and Angel Gomez, "Performance evaluation of front- and back-end techniques for ASV spoofing detection systems based on deep features," in *Proc. IberSPEECH*, Nov. 2018, pp. 45–49.

[21] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3286–3295.

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS,* 2017.

[24] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR),* 2021.

[26] M. Morise, F. Yokomori, and K. Ozawa, "Effect of speech transformation on impostor acceptance," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process. Proc.*, May 2006, pp. 933–936.

[27] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "WORLD: A vocoder-based highquality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[28] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499.*

[29] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP),* 1983, pp. 804–807.

[30] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-Natural speech waveform conversion using cycle-consistent adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT),* Dec. 2018, pp. 632–639.

[31] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* May 2019, pp. 5916–5920.

[32] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. Interspeech,* 2014, pp. 2514–2518.

[33] N. Brummer and E. de Villiers, "The BOSARIS toolkit ¨ user guide: Theory, algorithms and code for binary classifier score processing", 2011.

[34] T. Kinnunen, K. Lee, H. Delgado, N. Evans, M. Todisco, J. Sahidullah, M.and Yamagishi, and D. A Reynolds, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification" in *Proc. Speaker Odyssey Workshop*, 2018, pp. 312–319.

[35] Zhang, You and Jiang, Fei and Duan, Zhiyao, "One-class Learning Towards Synthetic Voice Spoofing Detection" in *IEEE Signal Processing Letters*, 2021, doi: 10.1109/LSP.2021.3076358