Anti-spoofing Ensembling Model: Dynamic Weight Allocation in Ensemble Models for Improved Voice Biometrics Security

Eros Rosello[1], *Angel M. Gomez*[1], *Iván López-Espejo*[1], *Antonio M. Peinado*[1], *Juan M. Martín-Doñas*[2]

[1]Dept. Signal Theory, Telematics and Communications and CITIC, University of Granada, Spain
[2]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), San Sebastián, Spain

{erosrosello,amgg,iloes,amp}@ugr.es, jmmartin@vicomtech.org

Abstract

This paper proposes an ensembling model as spoofed speech countermeasure, with a particular focus on synthetic voice. Despite the recent advances in speaker verification based on deep neural networks, this technology is still susceptible to various malicious attacks, so that some kind of countermeasures are needed. While an increasing number of anti-spoofing techniques can be found in the literature, the combination of multiple models, or ensemble models, still proves to be one of the best approaches. However, current iterations often rely on fixed weight assignments, potentially neglecting the unique strengths of each individual model. In response, we propose a novel ensembling model, an adaptive neural network-based approach that dynamically adjusts weights based on input utterances. Our experimental findings show that this approach outperforms traditional weighted score averaging techniques, showcasing its ability to adapt to diverse audio characteristics effectively.

Index Terms: Anti-spoofing, deep learning, ensemble model, wav2vec 2.0, fake audio.

1. Introduction

Voice biometrics systems employ automatic speaker verification (ASV) technology to confirm a speaker's identity through their unique voice characteristics. Recent advances in deep neural networks (DNNs) have notably enhanced ASV system performance [1]. However, these systems are still vulnerable to various malicious attacks, such as voice synthesis (text-to-speech, TTS), voice conversion (VC), replay, and impersonation attacks, posing significant security risks [2].

In this context, the development of countermeasures against ASV spoofing attacks or deepfakes has notoriously attracted the attention of the scientific community. Numerous evaluation campaigns, such as ASVspoof 2015 [3], 2017 [4], 2019 [5], and 2021 [6], have been focused on logical access (LA) attacks (TTS and VC), physical access attacks (replay attacks), and speech deepfake detection. These campaigns underscored the need for robust technologies resilient to diverse attack vectors and environmental conditions, with DNNs emerging as the most effective approach [7–12]. In this paper, we focus our attention on TTS and VC attacks, which rely on high-quality synthesized speech. These are commonly associated with LA attacks on biometrics systems as well as audio deepfakes [2].

While an increasing number of anti-spoofing techniques can be found in the literature, the fusion of a set of different systems, usually referred to as ensemble model, has provided the best results in recent challenges [5, 13]. Weighted score averaging (WSA), employed in the recent 2019 and 2021 ASVspoof challenges, has showcased superior performance to other ensemble techniques or to single systems [11, 14]. Nevertheless, this conventional ensemble model adheres to a rigid practice of assigning weights based solely on overall model performance, potentially neglecting the nuanced strengths inherent to each individual model.

We hypothesize that each model may exhibit proficiency in dealing with certain attacks or spoofing clues, potentially identifying subtle nuances, beyond attack modes or noise profiles, that each model best deals with. This observation suggests the potential for an ensemble model whose weight allocation dynamically adapts to the input data. Such adaptability might significantly enhance the model ensemble efficacy and resilience in real-world scenarios, where different attack types and acoustic environments are expected.

Following this idea, in this paper we introduce a DNN-based ensemble model, referred to as *ensembling model*, which dynamically adjusts the weighting based on input utterances, thus leveraging the advantages of each anti-spoofing model used in the ensemble. As a proof of concept, we test the feasibility of this new ensemble technique with four simple anti-spoofing models: the baseline models of the ASVspoof 2021 challenge. We not only compare our ensemble model with the classical WSA, but also with an equivalent model trained for binary (bonafide or spoof) audio classification in order to demonstrate the potential of the idea.

The rest of this paper is organized as follows: Section 2 presents our proposed ensembling model technique and the model used in this work. Section 3 outlines our experimental setup. Section 4 presents the experimental results obtained. Finally, in Section 5, we summarize our research findings and draw some conclusions.

2. Proposed ensembling model

In this section, we provide an overview of our ensembling model. First, we briefly detail the mathematical framework used for dynamically-extracting weights to calculate the final score. Then, we describe the architecture used to extract these weights. Finally, we provide a brief overview of the anti-spoofing models used for the ensemble.

2.1. Score computation

As mentioned earlier, our ensembling model dynamically adjusts weights for each anti-spoofing model based on input utterances. Specifically, our goal is to determine a function $f(\cdot) : \mathbb{R}^l \longrightarrow \mathbb{R}^M$, that, given an input utterance, $\mathbf{x} \in \mathbb{R}^l$, returns the optimal set of weights $\mathbf{w} = (w_1, \ldots, w_M)^\top = f(\mathbf{x})$, where l is the length of the utterance and M is the number of anti-spoofing models to be used. In our case, $M = 4$ (see Subsection 2.3).

To train our ensembling model we employ a cross-entropy

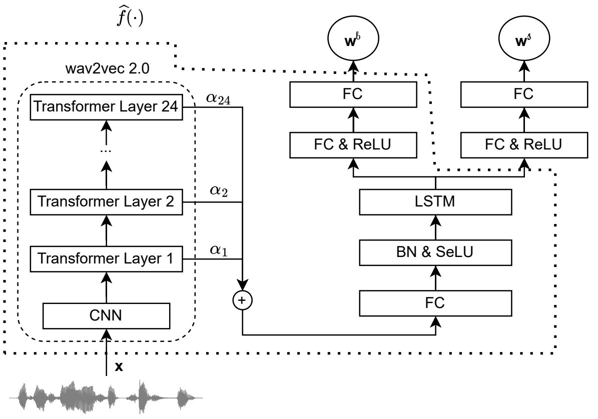

Figure 1: *Block diagram of the proposed ensembling model.*

loss which needs both bonafide and spoof scores. Consequently, during training, our ensembling model computes $2M$ weights from an input utterance ($\mathbf{x}$), two for each anti-spoofing model: one for calculating the final bonafide score, $\mathbf{w}_x^{\mathscr{b}} = (w_1^{\mathscr{b}}(\mathbf{x}), \ldots, w_M^{\mathscr{b}}(\mathbf{x}))^\top$, and one for computing the final spoof score, $\mathbf{w}_x^{\mathscr{s}} = (w_1^{\mathscr{s}}(\mathbf{x}), \ldots, w_M^{\mathscr{s}}(\mathbf{x}))^\top$. These weights are used to compute the final scores, $o_x^{(\mathscr{b},\mathscr{s})}$, according to the formula

$$o_x^{(\mathscr{b},\mathscr{s})} = \left(\mathbf{w}_x^{(\mathscr{b},\mathscr{s})}\right)^\top \cdot (\mathbf{s}_x - \boldsymbol{\beta}), \tag{1}$$

where $\mathbf{s}_x = (s_1(\mathbf{x}), \ldots, s_M(\mathbf{x}))^\top$ are the bonafide scores from the anti-spoofing models and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^\top$ are learnable parameters independent from the input utterance.

Note that only bonafide scores from anti-spoofing models are used to calculate our final scores due to the fact that many models only provide the former scores. In addition, it must be taken into account that the final spoof score and its corresponding weights are discarded during inference (i.e., they are only used for training).

2.2. Weight extraction

A diagram illustrating the proposed weight extraction system is presented in Figure 1. The system leverages the pre-trained wav2vec 2.0 (W2V2) XLS-128 model [15] as a feature extractor with no fine-tuning. In this extractor, the raw speech waveform undergoes processing by a feature encoder consisting of several convolutional layers. This encoder extracts vector representations of size $1,024$ every 20 ms, employing a receptive field of 25 ms. Subsequently, these encoder features are input to a transformer network comprising 24 layers to derive contextualized representations of the speech signal.

Studies have demonstrated that, for tasks like speaker verification and emotion recognition, when using self-supervised models as feature extractors, more discriminative information can be derived from the first or intermediate layers rather than the last layer [16, 17]. Therefore, following the approach of prior research [16–18], instead of only using the output from the last layer, we utilize the hidden representations of the various transformer layers as input to our model. Similar to [18], temporal normalization [19] is applied to the hidden representations of the different transformer layers. For each of the T temporal steps, an output representation is computed by applying a weighted sum of the normalized hidden representations. These weights, denoted as $\{\alpha_k;\, k = 1, \ldots, 24\}$ in Figure 1, are network trainable parameters normalized to sum one [16].

Table 1: *Architecture of the ensembling model, where T is the number of time frames and N is the batch size.*

Layer	Output size
Input (W2V2 features)	$N \times T \times 1,024 \times 24$
Temp. norm. and weighted sum	$N \times T \times 1,024$
FC layer, BN & SeLU	$N \times T \times 128$
LSTM	$N \times 80$
FF module	$N \times M$
Final scores	$N \times 2$

As a result, we obtain a vector sequence (of length T) that undergoes further processing through a fully-connected (FC) layer along with batch normalization (BN) and a scaled exponential linear unit (SeLU) activation function, ultimately reducing the vectors' dimension. The resulting vector sequence is processed by a long short-term memory (LSTM) network. The last output vector of the deepest LSTM layer is then fed to a feed-forward (FF) module which consists of a FC layer that halves the dimension followed by a rectified linear unit (ReLU) activation, and another FC layer that extracts M weights. During training, the FF module is replicated so one is utilized to extract the bonafide weights, while the other is employed to extract the spoof weights (see Figure 1). The proposed ensembling model is summarized in Table 1.

2.3. Anti-spoofing models

As explained previously, our ensembling model dynamically assigns weights to a set of anti-spoofing models. In this study, we utilize the four baseline models from ASVspoof 2021 [6] to this end. Briefly, these models are:

- **CQCC GMM**: A Gaussian mixture model (GMM)-based system that operates on constant-Q cepstral coefficients (CQCCs) [20];
- **LFCC GMM**: Another GMM-based system that operates on linear frequency cepstral coefficients (LFCCs) [21];
- **LFCC LCNN**: A model based on a lightweight convolutional neural network (LCNN) architecture [22], incorporating LSTM layers and average pooling, that utilizes LFCC features;
- **RawNet2** [23]: A DNN-based system employing the fully end-to-end RawNet2 architecture [24], which operates directly on raw audio waveforms.

Each of these systems was trained exclusively on the corresponding ASVspoof 2019 training data and optimized using the respective development data. No data augmentation techniques were applied during training.

2.3.1. Comparative performance of the anti-spoofing models

In order to assess the feasibility of our proposal, we examine the performance of the above baseline models on the ASVspoof 2021 LA database. Our hypothesis is that if all baseline models misclassify the same utterances, our proposal could not bring any improvements. However, if the anti-spoofing models exhibit complementary behavior, wherein audios misclassified by one model are correctly classified by another model, our proposal could exploit this by prior analyzing the utterance and assigning it to the most suitable anti-spoofing model. Note that, despite this may be seen as a detour to a direct classification, the

Table 2: *Number of incorrectly classified utterances. Diagonal: Number of utterances incorrectly classified by each standalone model. Off-diagonal: Number of utterances incorrectly classified by two models. The last row shows the performance of each anti-spoofing model in terms of equal error rate (EER), in percentages.*

Model	CQCC GMM	LFCC GMM	LFCC LCNN	RawNet2
CQCC GMM	$23,142$	$11,231$	$7,340$	$4,314$
LFCC GMM	$11,231$	$28,593$	$3,688$	$2,501$
LFCC LCNN	$7,340$	$3,688$	$13,721$	$4,678$
RawNet2	$4,314$	$2,501$	$4,678$	$11,071$
EER (%)	15.62	19.30	9.26	7.47

goal is that the ensembling model actually seizes the different strengths of the different anti-spoofing models.

In Table 2, we present the equal error rate (EER) performance of each baseline model on the ASVspoof 2021 LA eval database. Additionally, this table displays the number of utterances misclassified by each model at the EER threshold (diagonal entries) and the number of utterances misclassified by two models simultaneously (off-diagonal entries).

As can be seen from Table 2, the anti-spoofing models show varied misclassification patterns, which can be derived from the significantly smaller number of utterances jointly misclassified by any two models compared to that of every standalone model. Moreover, some model combinations demonstrate superior complementary behavior compared to others. For example, the LFCC GMM (row 2) model shows a better complementary behavior with RawNet2 (row 2, column 4) than with CQCC GMM (row 2, column 1). Furthermore, we also have examined the behavior of all baseline models simultaneously (not reported in Table 2), and we have found that only $1,107$ utterances —approximately 10% of those misclassified by the best model (see RawNet2 in Table 2)— are misclassified by all anti-spoofing models.

The above results suggest that it is possible to significantly enhance the joint performance of these models if we were able to assign each utterance to the appropriate model, which could be achieved by means of our proposed ensembling framework.

3. Experimental setup

In this section, we describe the datasets and evaluation metrics considered in our experiments as well as the training details.

3.1. Models, datasets and evaluation metrics

For the anti-spoofing models, we use publicly available model weights and scores[1], except for RawNet2, which was trained from scratch due to issues with the provided weights. Models were originally trained using the ASVspoof 2019 LA training and development partitions for training and validation, respectively [5]. Additionally, the scores returned by each anti-spoofing model are normalized prior to be fed to the ensemble models. We apply a min-max normalization followed by a sigmoid function, ensuring that all scores fall within the range $[0, 1]$, easing comparison across models with similar thresholds.

Our ensembling model is not trained with the exact same data as the anti-spoofing models (ASVspoof 2019 LA training), due to the exceptional performance of the anti-spoofing models on their training data (we could not assign weights to the most appropriate models if all models performed very well on the data we used). Instead, we use the ASVspoof 2019 development set for training and the ASVspoof 2021 LA progress partition as the validation set. No data from the Deep Fake (DF) partition are used for validation.

To evaluate our proposed method, we conduct experiments on the LA and DF evaluation subsets of the ASVspoof 2021 database [6]. These subsets contain both bonafide and spoofed speech, the latter generated using TTS and VC systems. The LA subset encompasses codec and transmission variability, while the DF subset introduces compression variability. In addition, while the 2019 training and development sets include only six known attacks (2 VC-based and 4 TTS-based), the 2021 evaluation datasets incorporate unseen attacks [6].

As primary metric, we employ the pooled EER [25]. Additionally, for the LA subset, we also report minimum normalized tandem detection cost function (t-DCF) [26] scores.

3.2. Implementation details

We create, by either cropping or padding the content as needed, 4-second long input audio signals to be processed. For training, we employ the standard Adam optimizer [27] with an initial learning rate of 10^{-4}, a weight decay of 10^{-4}, and a batch size of 32 training samples. Weighted cross-entropy is used as the loss function. After the W2V2 feature extraction, a FC layer produces 128 output dimensions. Consequently, we incorporate an LSTM network with an input size of 128, a hidden size of 80, and three layers, resulting in a total of 300k trainable parameters. We train 5 different realizations of the proposed model using 5 different random seeds, which allows us to study potential statistically significant performance gains. In instances where we compare our model with other results lacking multiple runs, we utilize the model that best performs in our validation dataset (ASVspoof 2021 LA progress). To prevent overfitting, we implement an early-stopping scheme [28], finishing the training process if the weighted cross-entropy on the validation set fails to improve over 8 consecutive epochs. We also halve the learning rate if the validation loss does not decrease for 3 epochs in a row. Finally, all experiments were done in a Nvidia 3090 GPU.

4. Results

In this section, we present our experimental results, comparing our proposed ensembling model with other WSA ensembles and a direct classifier with identical architecture to our model.

4.1. Comparison with classical weighted score averaging models

We compare our proposal with the classical WSA ensemble technique used in other works [11, 13]. To this end, we con-

Table 3: *Comparison of our proposed ensembling model with other WSA ensemble models on the ASVspoof 2021 LA and DF evaluation sets in terms of EER (%) on both sets and min t-DCF on LA.*

	LA		DF
Model	EER (%)	min t-DCF	EER (%)
GS Ensemble	5.784	0.2963	23.633
EB Ensemble	6.641	0.3051	20.150
Ensembling Model	**2.315**	**0.2339**	**5.596**

[1] `https://github.com/asvspoof-challenge/2021`

Table 4: *Architecture of the classification model, where, as before, T is the number of time frames and N is the batch size.*

Layer	**Output size**
Input (W2V2 features)	$N \times T \times 1,024 \times 24$
Temp. norm. and weighted sum	$N \times T \times 1,024$
FC layer, BN & SeLU	$N \times T \times 128$
LSTM	$N \times 80$
FF module	$N \times 2$

sider two different weight selection techniques:

- **Grid search (GS)**: We conduct a grid search over a predefined range of weights (with a minimum of 0.1 and all weights adding up 1) to identify the combination that minimizes the EER on the progress phase of ASVspoof 2021 LA and DF, for the LA and DF ensemble models, respectively.
- **Error-based (EB)**: We use the performance of the anti-spoofing models on the progress phase of ASVspoof 2021 for each track (LA and DF) to assign the ensemble weights, in particular, the inverse of the EER values. Higher EER values indicate worse performance, so taking their inverse ensures that models with better performance receive larger weights.

We report the results of this classical WSA ensemble with the 4 baseline models (Subsection 2.3) in Table 3. We can observe that our proposed ensembling model significantly outperforms these WSA techniques on both LA and DF subsets.

4.2. Comparison with an equivalent classification model

In order to prove that our ensembling model truly leverages the underlying anti-spoofing models, we have also developed a classification model with identical structure to our ensembling model. The architecture of this classification model, detailed in Table 4, closely resembles that of the proposed weight extraction system and, thereby, has a similar number of parameters. The main difference lies in the fact that the ensembling model employs two FF modules to estimate bonafide and spoof weights during training, whereas the current classification model only needs one FF module, since this model directly outputs bonafide and spoof scores.

Table 5 presents the experimental results of our ensembling model and the equivalent classification model when evaluated on the ASVspoof 2021 LA and DF evaluation subsets. We use 5 different random seeds, in order to present the results of the best model (selected through the development set) and the average across the 5 runs (in parentheses). To ensure fairness, we train the classification model using the same data as the ensembling model (see Subsection 3.1). According to Table 5, on the LA partition, we observe that our ensembling model remarkably outperforms the classification model. In particular, a Welch's t-test [29], which is performed using the results derived from the different random seeds, yields a p-value of 2.76%, indicating that the proposed ensembling model offers statistically significant enhancements (given a standard significance level of 5%) over the classification model on LA. On the other hand, on DF, there are no statistically significant differences between the two models compared.

Table 5: *Comparison of our proposed ensembling model with a classification model of similar characteristics on the ASVspoof 2021 LA and DF evaluation subsets in terms of EER (%). Reported results are the best (average) obtained from five runs with different random seeds.*

Model	LA	DF
Classifier Model	9.961 (12.801)	6.517 (**8.043**)
Ensembling Model	**2.315** (**4.484**)	**5.596** (8.743)

Table 6: *Comparison of our proposed ensembling model with other WSA ensemble models (5 anti-spoofing models) on the ASVspoof 2021 LA and DF evaluation sets in terms of EER (%) on both sets and min t-DCF on LA.*

	LA		DF
Model	EER (%)	min t-DCF	EER (%)
GS Ensemble	5.210	0.2943	11.393
EB Ensemble	5.757	0.3040	12.979
Ensembling Model	**2.315**	**0.2339**	**5.596**

Additionally, we also report the results of a classical WSA ensemble that uses 5 anti-spoofing models: the 4 baseline models plus the equivalent classification model, trained with the same data as the anti-spoofing models (using 5 different random seeds and selecting the best model for the ensemble), in comparison with our ensebling model. These results are detailed in Table 6, where we can observe that our proposed ensembling model (with 4 models) still outperforms the extended WSA techniques (with 5 models) on both LA and DF.

5. Conclusions

This paper introduces a novel ensembling model capable of dynamically weighing the scores provided by a set of underlying anti-spoofing models depending on the input utterance. The goal is that the ensemble seizes the most suitable model for detection according to the characteristics or clues contained in the utterance. Our ensembling model is compared to other WSA ensemble techniques under LA and DF scenarios.

Our results show the effectiveness of leveraging dynamic weight adjustments in ensemble models, since our approach obtains significantly better results than classical WSA ensembles. Results suggest that, by incorporating adaptive weighting mechanisms, our ensembling model can exploit the strengths of each anti-spoofing model for different types of input utterances, showing improved overall system resilience against spoofing attempts when compared to classical WSA techniques. To further support this hypothesis, our proposed approach has been compared with a direct classification model with similar characteristics. On the LA partition, our ensemble model shows a significant improvement over the classification model. This seems to indicate that our model is able to leverage different strengths of the anti-spoofing models. In addition, our technique obtains better results than a classical ensemble containing this classification model, suggesting that, in a real scenario, where the actual performance of the models to be used in the final ensemble is unknown, our approach might obtain better results than a classical WSA ensemble, even when the latter includes an additional classification model of identical characteristics to our ensembling model.

Moving forward, future work will focus on *1)* refining weight allocation strategies to further optimize the performance of ensemble models, *2)* exploring alternative architectures, and *3)* extending the scope of application to domains other than anti-spoofing. Finally, the potential of this ensemble technique with more powerful anti-spoofing models is also worth exploring.

6. Acknowledgements

This paper is part of the project PID2022-138711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, and the FPI grant PREP2022-000363. Also, it has been supported in part by European Union's Horizon Europe research and innovation programme in the context of project EITHOS under Grant Agreement No. 101073928.

7. 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. 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.

[8] A. Gomez-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.

[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] E. Rosello, A. Gomez-Alanis, A. M. Gómez, and A. M. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," *INTERSPEECH 2023*, 2023.

[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] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. H. Kinnunen, M. Todisco, J. Yamagishi, N. W. D. Evans, A. Nautsch, and K.-A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2022.

[14] A. Nautsch, X. Wang, N. W. D. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K.-A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, pp. 252–265, 2021.

[15] 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 2021.

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

[17] 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," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, 2021.

[18] J. M. Martín-Doñas and A. Álvarez, "The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9241–9245, 2022.

[19] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *ArXiv*, vol. abs/1607.08022, 2016.

[20] M. Todisco, H. Delgado, and N. W. D. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

[21] M. Sahidullah, T. H. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Interspeech*, 2015.

[22] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Interspeech*, 2019.

[23] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. W. D. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2," *ICASSP*, 2021.

[24] J. weon Jung, S. bin Kim, H. jin Shim, J. ho Kim, and H. jin Yu, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms," in *Interspeech*, 2020.

[25] 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.

[26] 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.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR 2015 – 3^{rd} International Conference on Learning Representations, May 7-9, San Diego, USA*, 2015.

[28] N. Gershenfeld, "An experimentalist's introduction to the observation of dynamical systems," in *Directions in Chaos — Volume 2*. World Scientific, 1988, pp. 310–353.

[29] B. L. Welch, "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved," *Biometrika*, vol. 34, pp. 28–35, 1947.