Interspeech 2024
1-5 September 2024, Kos, Greece

Source Tracing of Audio Deepfake Systems

Nicholas Klein, Tianxiang Chen, Hemlata Tak, Ricardo Casal, Elie Khoury

Pindrop, Atlanta, GA, USA

{nklein,tchen,Hemlata.Tak,rcasal,ekhoury}@pindrop.com

Abstract

Recent progress in generative AI technology has made audio deepfakes remarkably more realistic. While current research on anti-spoofing systems primarily focuses on assessing whether a given audio sample is fake or genuine, there has been limited attention on discerning the specific techniques to create the audio deepfakes. Algorithms commonly used in audio deepfake generation, like text-to-speech (TTS) and voice conversion (VC), undergo distinct stages including input processing, acoustic modeling, and waveform generation. In this work, we introduce a system designed to classify various spoofing attributes, capturing the distinctive features of individual modules throughout the entire generation pipeline. We evaluate our system on two datasets: the ASVspoof 2019 Logical Access and the Multi-Language Audio Anti-Spoofing Dataset (MLAAD). Results from both experiments demonstrate the robustness of the system to identify the different spoofing attributes of deepfake generation systems.

Index Terms: Anti-spoofing, audio deepfake detection, explainability, ASVspoof

1. Introduction

In recent years, deepfake generation and detection have attracted significant attention. On January 21, 2024, an advanced text-to-speech (TTS) system was used to generate fake calls to manipulate the voice of US President, Joe Biden, encouraging voters to skip the 2024 primary election in the state of New Hampshire [1]. This incident underscores the critical need for deepfake detection that is reliable and trusted. Thus, explainability in deepfake detection systems is crucial. Within this research area, the task of deepfake audio source attribution has recently been gaining interest [2–10]. The goal of this task is to predict the source system that generated a given utterance. For example, the study in [2] aims to predict the specific attack systems used to produce utterances in ASVspoof 2019 [11]. This approach of directly identifying the name of the system misses the opportunity to categorize the spoofing systems based on their attributes. Such attribute-based categorization allows for better generalization to spoofing algorithms that are unseen in training but are composed of building blocks, such as acoustic models or vocoders, that are seen. Along these lines, authors in [3] propose a more generalizable approach by classifying the vocoder used in the spoofing system. Authors in [4] explore classifying both the acoustic model and vocoder, finding that the acoustic model is more challenging to predict. The work in [5] takes this further by proposing to classify several attributes of spoofing systems in ASVspoof 2019 LA: conversion model[1], speaker representation, and vocoder. However, their findings demonstrate accuracy challenges in discerning speaker representation. Another drawback of their evaluation protocol is that the ASVspoof 2019 dataset is relatively outdated as there have been many advancements in voice cloning techniques in the last five years. Finally, their choice of categories for acoustic model and vocoder are very broad (e.g."RNN related" for acoustic model and "neural network" for vocoder) and may not be that useful in narrowing down the identity of the spoofing system.

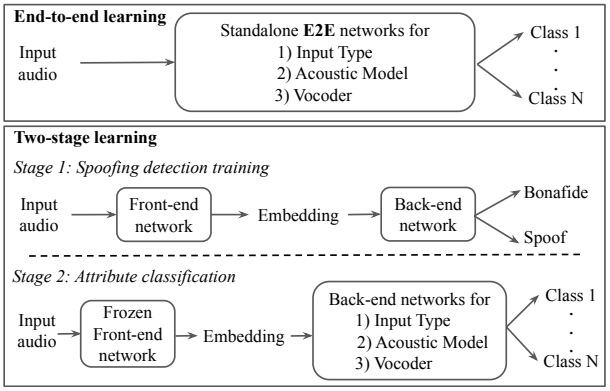

Figure 1: *Illustration of proposed frameworks for spoofing attribute-classification. Top: End-to-end learning from audio. Bottom: Two-stage learning that includes a traditional countermeasure (CM) and an auxiliary classifier trained on embeddings.*

In this work, we investigate two attribute classification strategies as illustrated in Fig. 1: an end-to-end learning method which trains standalone systems for each attribute and a two-stage learning method which leverages the learned representations of existing countermeasure systems. To this end, we leverage three state-of-the-art systems, namely ResNet [12], self-supervised learning (SSL) [13], and Whisper [14]. In addition to identifying the acoustic model and vocoder, we propose classifying the input type (i.e. speech, text, or bonafide) rather than speaker representation. This allows for distinguishing between TTS and VC systems. As an anchor to previous work, we evaluate our methods on the ASVspoof 2019 protocol designed by [5]. To address the limitations of the outdated ASVspoof-based protocol, we design a new protocol based on the recent MLAAD dataset which consists of multilingual utter-

[1]In [5], the term "conversion model" is used instead of "acoustic model" to refer more generally to the encoder part of the system for both TTS and VC systems.

10.21437/Interspeech.2024-1283

Table 1: *ASVspoof 2019 LA protocol for attribute-classification tasks, adapted from [5].*

	# Bonafide	# Spoofed		
Sets	**-**	**Input type**	**Acoustic model**	**Vocoder**
Train	7,796	71,824	71,824	71,824
Eval	1,638	4,194	4,194	4,194

ances produced by 52 systems comprising a variety of state-of-the-art TTS systems. Compared to the ASVspoof-based protocol, this protocol uses more modern attack systems and replaces vague categories with specific acoustic models and vocoders. We make this novel MLAAD source tracing protocol publicly available[2]. To the best of our knowledge, this is the first study of source tracing on a multi-lingual TTS dataset.

2. Attribute classification of spoof systems

In this section, we describe our approaches for classifying the input type, acoustic model, and vocoder of the spoofing system used to generate a given audio.

2.1. Proposed strategies

We present two strategies for leveraging existing state-of-the-art (SOTA) spoofing countermeasure (CM) systems for the task of component classification:

- Our *End-to-End* (E2E) approach takes an existing CM architecture and trains the whole model for each of the multi-class component classification tasks separately, as depicted in the top part of Fig. 1.
- The *Two-Stage* approach, shown in the bottom of Fig. 1, splits training into two steps: first an existing CM is trained for the standard binary spoof detection task; next, the CM backbone is frozen and a lightweight classification head is trained on the CM's embeddings for each separate component classification task. For the classification head, we use the simple feed forward architecture from the back-end model of the ResNet spoof detection system described in [12].

While the second approach is limited to the information that the binary-trained CM learns, it is very attractive in practice: in addition to the reduction in computational costs, existing binary systems can be trained on significantly more data than we have component labels for and enhancing them with an auxiliary head rather than replacing them with a modified E2E system is much safer for models that run in production.

2.2. Countermeasures

We used three different CMs to validate our hypothesis. These systems are well known and have reported excellent detection performance on several datasets.

ResNet. This system consists of a front-end spoof embedding extractor and a back-end classifier. The front-end model is known as the ResNet18-L-FM model, as detailed in [12, 15]. To enhance the model's generalization capability, large margin cosine loss [16] (LMCL) and random frequency masking augmentation are applied during training. The back-end model is trained using the spoof embedding vectors for the classification tasks described in Section 2. The back-end classifier is a feed forward neural network with one FC layer described in [12].

[2]MLAAD protocol: `doi.org/10.5281/zenodo.11593133`

Self-supervised learning. SSL-based front-ends have attracted significant attention in the speech community, including spoofing and deepfake detection [13, 17–23]. The SSL-based CM architecture[3] is a combination of SSL-based front-end feature extraction and an advanced graph neural network based back-end, named AASIST [24]. The 160-dimensional CM embeddings are extracted prior to the final fully-connected output layer. The SSL feature extractor is a pre-trained wav2vec 2.0 model [25, 26], the weights of which are fine-tuned during CM training.

Whisper. The Whisper model is based on an off-the-shelf encoder-decoder Transformer architecture for automatic speech recognition (ASR) [27]. The Whisper-based CM architecture [14][4] is a combination of Whisper-based front-end feature extraction and light convolution neural network (LCNN) [28] as a back-end. For the front-end feature extraction, the Whisper embedding is concatenated with 128-dimensional linear frequency cepstral coefficients (LFCCs) [29] along with their delta and double-delta features. The 768-dimensional CM embeddings are extracted prior to the final fully-connected output layer. The reader is referred to [14] for further technical details.

3. Datasets and protocols

Two publicly available spoofing detection benchmarks are used in our study: the ASVspoof 2019 LA [11, 30] and the most recent MLAAD dataset [31].

3.1. ASVspoof 2019

The ASVspoof 2019 LA dataset has three independent partitions: train, development, and evaluation. Spoofed utterances are generated using a set of different TTS, VC, and hybrid TTS-VC algorithms [11]. To compare our methods against those presented in [5], we adopt their protocol partition as detailed in Table 1. Notably, it only includes a train and development set, so we do not do any hyper-parameter search on this protocol. While we use the same categories as [5] for the acoustic and vocoder tasks, we create a new "Input type" task which is helpful to separate between TTS and VC systems. Table 1 summarises the statistics for each partition used for the different attribute classification tasks on the ASVspoof 2019 dataset.

3.2. MLAAD

MLAAD consists of TTS attacks only, however it includes 52 different state-of-the-art spoofing algorithms [31]. We manually label the acoustic models and vocoders based on the available metadata.[5] Since MLAAD includes only TTS systems, we focus on acoustic model and vocoder classification without any input-type prediction. For end-to-end systems such as VITS and Bark, we use the name of the full system as the acoustic model and vocoder labels. Additionally, while the MLAAD dataset labels 19 different architectures, our protocol groups several systems that are identical aside from their training data. For example, the systems "Jenny", "VITS", "VITS-Neon", and "VITS-MMS" are all labeled with the same acoustic model and vocoder category "VITS". For the bonafide class, we include bonafide samples from the multilingual M-AILABS

[3]`github.com/TakHemlata/SSL_Anti-spoofing`

[4]`github.com/piotrkawa/deepfake-whisper-features`

[5]We use the "model_name" field provided in the dataset's accompanying "meta.csv" file. System descriptions for each model_name can be found in the Coqui-TTS [32] and HuggingFace repositories.

Table 2: *MLAAD protocol for acoustic model classification task. Tacotron2: T*

	# Bonafide	# Spoofed												
Sets	**-**	**bark**	**capacitron**	**fastpitch**	**glowtts**	**neural-hmm**	**overflow**	**T**	**T-dca**	**T-ddc**	**tortoise tts**	**vits**	**xtts-v1**	**xtts-v2**
Train	28,345	762	845	859	1,866	855	846	859	856	2,802	834	15,633	4,789	4,758
Dev	6,584	159	84	61	1,049	65	72	83	72	225	77	4,877	1,251	1,688
Eval	6,390	79	71	80	1,085	80	82	58	72	973	89	12,490	1,960	2,554

Table 3: *MLAAD protocol for vocoder classification task. Multiband-mel:mul; Wavegrad:w-grad*

	# Bonafide	# Spoofed					
Sets	-	**bark**	**hifi-gan**	**mul-gan**	**univnet**	**vits**	**w-grad**
Train	28,345	762	9,135	2,680	6,473	15,633	859
Dev	6,584	159	2,112	150	1,392	4,877	83
Eval	6,390	79	3,753	170	2,135	12,490	58

dataset [33]. We divide the data into train, development, and evaluation partitions while preventing speaker overlap. To enable this for the spoof samples, we assign voice labels using spherical k-means clustering on embeddings from the state-of-the-art speaker verification system, ECAPA-TDNN [34]. We use the elbow criteria on the inertia values to select K=75 clusters. We remove two vocoders, Griffin-Lim [35] and Fullband-MelGAN [36], since they each have a cluster containing most of their samples. The resulting acoustic model and vocoder labels along with their number of examples in each partition are presented in Table 2 and Table 3, respectively.

4. Experimental Results

4.1. Implementation details

ResNet and SSL models use 4 second (s) raw audio as input, whereas the Whisper model processes on $30s$ audio. For ResNet, LFCC features are extracted using $20ms$ window and $10ms$ frame-shift along with its delta and double delta features. Since fine-tuning large SSL models requires high GPU computation, experiments with SSL are performed with a smaller batch-size of 16 and a lower learning rate of 10^{-6}. We used the same set-up for SSL and Whisper based models as describe in [13] and [14], respectively. SSL and Whisper based models are fine-tuned on ASVspoof and MLAAD datasets in their respective experiments, whereas the ResNet model is trained from scratch. For the auxiliary classifier, a batch size of 256 and a learning rate of 10^{-3} is used with no hyper-parameter tuning. The best model is chosen based on Dev set accuracy and average F1-score for ASVspoof and MLAAD experiments, respectively.

4.2. Results on ASVspoof 2019

Our results are compared with the previous study [5] on ASVspoof 2019 in terms of unweighted accuracy in Table 4.

Input type classification: This study introduces a novel task, predicting input types, which the previous study did not explore. We train classification heads using fixed ResNet, SSL, and Whisper based binary spoof detection models named as, ResNet (two-stage), SSL (two-stage), and Whisper (two-stage). These experiments achieve 97.8%, 96.7% and 78.4% accuracy, respectively. Our SSL model fine-tuned end-to-end, SSL (E2E), further improves accuracy to 99.9%.

Acoustic model classification: Several of our models surpass the previous study's highest accuracy of 88.4%, achieved by the multi-task-trained RawNet2 model in [5]. Specifically, SSL (two-stage), ResNet (two-stage), and SSL (E2E) achieve accuracies of 91.4%, 92.6%, and 99.4% (a 12.4% relative improvement over the previous study), respectively. The substantial increase in accuracy may be due to the fact that our models are specifically trained for these tasks, unlike the previous study's multi-task approach that jointly trained on acoustic, vocoder, and speaker representation tasks.

Table 4: *Results in terms of Accuracy (%) on the ASVspoof 2019 LA dataset. Methods presented in [5] are included in the top two rows for comparison with our methods. We show our results when training a classification head on top of fixed embeddings from the binary CM backbone ("two-stage") as well as when training the CM backbone end-to-end for this task ("E2E").*

Method	**Input type**	**Acoustic model**	**Vocoder**
ResNet34 [5]	-	86.5	84.5
RawNet 2 [5]	-	88.4	77.5
ResNet (two-stage)	97.8	92.6	81.4
SSL (two-stage)	96.7	91.4	73.7
Whisper (two-stage)	78.4	64.4	63.8
ResNet (E2E)	90.5	84.3	83.8
SSL (E2E)	99.9	99.4	84.6
Whisper (E2E)	77.5	72.3	59.5

Table 5: *Results in terms of macro-averaged Accuracy / F1-score (%) on the MLAAD dataset. We show our results when training a classification head on top of fixed embeddings from the binary CM backbone ("two-stage") as well as when training the CM backbone end-to-end for this task ("E2E").*

Method	**Acoustic model**	**Vocoder**
ResNet (two-stage)	18.8 / 12.0	30.3 / 26.5
SSL (two-stage)	36.6 / 16.7	50.4 / 34.9
Whisper (two-stage)	49.6 / 31.5	48.1 / 40.2
ResNet (E2E)	85.4 / 82.3	97.4 / 93.3
SSL (E2E)	60.0 / 59.3	93.5 / 89.4
Whisper (E2E)	58.6 / 47.9	62.8 / 60.3

Vocoder classification: Our SSL (E2E) model slightly outperforms the previous study with an accuracy of 84.6% (a 0.1% relative improvement). Unlike the acoustic model, we do not see the same level of improvement. Analyzing errors from our top-performing model, SSL (E2E), we find that 882 out of 896 mis-predictions occur from predicting attack A07 as "Neural Network". Attack A07 uses a non-neural WORLD vocoder, however it also uses a GAN-based post filter that identifies areas of the waveform to mask out (See [11] for further details). This post-filter is not seen in training and must have consistently affected the final waveform in a way that mangled the resemblance to traditional vocoder audio. Aside from this one kind of error, our SSL (E2E) model's accuracy is 99.7%.

4.3. Results on MLAAD

We report results in terms of macro-averaged F1 and accuracy scores in Table 5. With the larger number of specific vocoder and acoustic model categories compared to the ASVspoof pro-

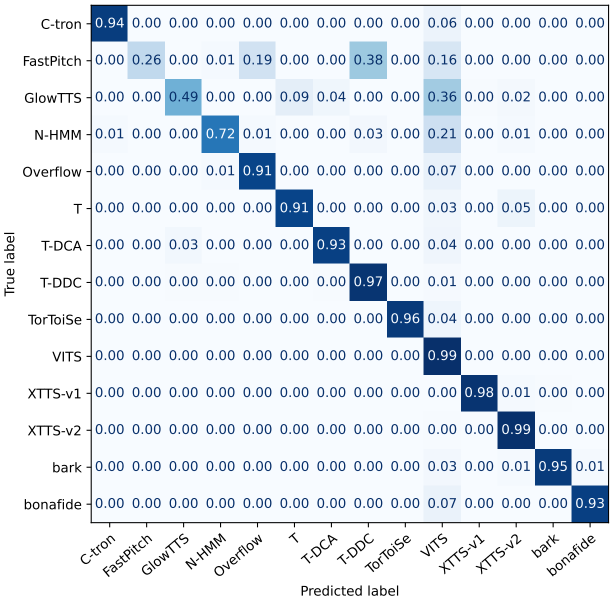

Figure 2: *Confusion matrix for ResNet (E2E) acoustic model predictions on the MLAAD evaluation set. Prediction counts are normalized by true label counts (by row). T: Tacotron2*

tocol, we find that the vocoder is easier to distinguish than the acoustic model, as observed in [4]. Our best performance on each of these tasks is achieved by our ResNet (E2E) model, with average F1-scores of 93.3% for the vocoder and 82.3% for the acoustic model task. Our two-stage strategy performed noticeably worse here, indicating that the binary spoof detection models omitted much architecture-specific information when fitting to the binary task. The auxiliary head models that performed the worst on the acoustic and vocoder classification tasks are the ones that leveraged the ResNet architecture. This is likely due to the ResNet model's use of the LMCL loss function [16] which minimizes intra-class variation and thus reduces the separability of deepfake examples produced by different architectures.

Error analysis: We analyze the mistakes most commonly made by our top-performing ResNet (E2E) model. In the acoustic model task, we get <90% accuracy on three categories, as can be seen in the confusion matrix illustrated in Fig. 2. Fastpitch is mistaken for Tacotron2-DDC 38% of the time, Overflow 19% of the time, and VITS 16% of the time; GlowTTS is mistaken for VITS 36% of the time; and Neural-HMM is mistaken for VITS 21% of the time. In each of these cases, the predicted and actual acoustic models have a high degree of overlapping voice clusters in the test set. This indicates that the acoustic model embeddings are capturing voice information, and systems that share a common voice in the test set are more challenging to distinguish. In the vocoder task, the ResNet (E2E) model's performance on the different categories is high. The most mistaken category is bonafide, in which case VITS is mistakenly predicted 7% of the time.

4.4. Embedding visualization

Our top performing models' embeddings for the acoustic classification task using ASVspoof and MLAAD protocols are visualized using UMAP in Fig. 3. Notably, the acoustic models in the MLAAD dataset exhibit more difficulty in separation. This

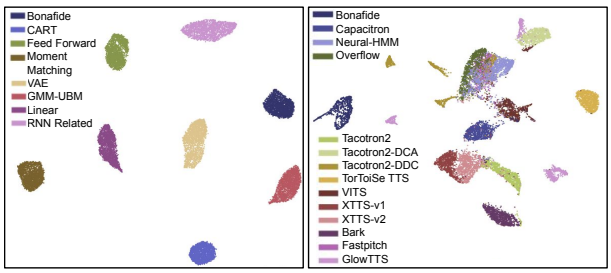

Figure 3: *Embeddings from our top performing models on the acoustic model classification task of each of our protocols, plotted using UMAP dimensionality reduction with* $n\_neighbors = 50$*. Left: ASVSpoof embeddings from SSL (E2E) model. Right: MLAAD embeddings from ResNet (E2E) model.*

challenge may stem from overlapping voices among different models in the test set, as discussed in the previous error analysis section. Additionally, we observe distinct clusters of acoustic models with similar architectures: XTTS-v1 and XTTS-v2; as well as Neural-HMM [37] and Overflow [38] (which combines Neural-HMM with normalizing flows).

5. Conclusions and Discussions

In this paper, we propose three multi-class classification tasks to give more explanatory predictions in the place of traditional binary spoof detection: input-type, acoustic model, and vocoder classification. We experiment with two methods of leveraging open source spoof detection systems to accomplish this task and evaluate them on a recently introduced ASVspoof 2019 protocol as well as a new protocol that we design using the more modern MLAAD dataset. Our SSL (E2E) method outperforms the previous study on ASVspoof that we compare to on the acoustic and vocoder tasks with relative improvements in accuracy of 12.4% and 0.1% respectively while achieving 99.9% accuracy on our newly introduced input-type classification task. On our MLAAD protocol which includes a greater number of vocoder and acoustic categories from more modern TTS systems, our ResNet (E2E) model yields an average f1 score of 82.3% for the acoustic model and 93.3% for the vocoder classification task. Our findings support existing literature that suggest that the vocoder is easier to distinguish than the acoustic model. Additionally, we observe that the acoustic models of systems that produce similar voices are more challenging to discriminate. Thus, a potential area of future study is to more explicitly ignore voice-specific information.

Our experiments with two-stage classification methods that leverage embeddings from binary spoof detection systems show promise, though they underperform on MLAAD compared to the full model fine-tuning methods. Future research in this area is crucial as models that augment rather than replace existing binary spoof detection systems are attractive, especially in industry where changes in the behavior of the binary detection system require thorough evaluation. Thus, one possible future experiment is to assess where in the binary model contains the most useful information for discriminating the different spoof system components. Additionally, assessing how the choice of loss function for the binary model affects the downstream multi-class performance could give insight into which existing models are best suited to being leveraged for two-stage learning.

6. References

[1] "Fake biden robocall tells voters to skip new hampshire primary election - bbc news," https://www.bbc.com/news/world-us-canada-68064247, Last Accessed: 05/03/2024.

[2] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, pp. 1–14, 2021.

[3] X. Yan, J. Yi, J. Tao, C. Wang, H. Ma, T. Wang, S. Wang, and R. Fu, "An initial investigation for detecting vocoder fingerprints of fake audio," in *Proc. of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022.

[4] C. Y. Zhang, J. Yi, J. Tao, C. Wang, and X. Yan, "Distinguishing neural speech synthesis models through fingerprints in speech waveforms," *ArXiv*, vol. abs/2309.06780, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261705832

[5] T. Zhu, X. Wang, X. Qin, and M. Li, "Source tracing: Detecting voice spoofing," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022.

[6] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "ADD 2023: the second audio deepfake detection challenge," in *Proc. IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023.

[7] X.-M. Zeng, J.-T. Zhang, K. Li, Z.-L. Liu, W.-L. Xie, and Y. Song, "Deepfake algorithm recognition system with augmented data for add 2023 challenge," in *Proc. IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023.

[8] Y. Tian, Y. Chen, Y. Tang, and B. Fu, "Deepfake algorithm recognition through multi-model fusion based on manifold measure," in *Proc. IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023.

[9] J. Lu, Y. Zhang, Z. Li, Z. Shang, W. Wang, and P. Zhang, "Detecting unknown speech spoofing algorithms with nearest neighbors," in *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023.

[10] J. Deng, Y. Ren, T. Zhang, H. Zhu, and Z. Sun, "Vfd-net: Vocoder fingerprints detection for fake audio," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 151–12 155.

[11] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[12] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020.

[13] 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 *Proc. The Speaker and Language Recognition (Speaker Odyssey) Workshop*, 2022.

[14] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *Proc. INTERSPEECH*, 2023.

[15] T. Chen and E. Khoury, "Spoofprint: a new paradigm for spoofing attacks detection," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 538–543.

[16] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] Z. Jiang, H. Zhu, L. Peng, W. Ding, and Y. Ren, "Self-supervised spoofing audio detection scheme." in *INTERSPEECH*, 2020.

[18] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vec feature for spoofing speech detection." in *Interspeech*, 2021.

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

[20] Y. Eom, Y. Lee, J. S. Um, and H. Kim, "Anti-spoofing using transfer learning with variational information bottleneck," in *Proc. INTERSPEECH*, 2022.

[21] X. Wang and J. Yamagishi, "Investigating active-learning-based training data selection for speech spoofing countermeasure," in *Proc. SLT*, 2023.

[22] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *Proc. ICASSP*, 2022.

[23] X. Wang and J. Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *Proc. ICASSP*, 2023.

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

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in neural information processing systems (NIPS)*, 2020.

[26] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. INTERSPEECH*, 2022.

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023.

[28] 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, no. 11, pp. 2884–2896, 2018.

[29] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. INTERSPEECH*, 2015.

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

[31] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," *arXiv preprint arXiv:2401.09512*, 2024.

[32] G. Eren and The Coqui TTS Team, "Coqui TTS," Jan. 2021. [Online]. Available: https://github.com/coqui-ai/TTS

[33] T. M. S. Dataset, "The m-ailabs speech dataset," https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/, 2023, Last accessed on 05/03/2024.

[34] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. INTERSPEECH*, 2020.

[35] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[36] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *IEEE Proc. Spoken Language Technology Workshop (SLT)*, 2021.

[37] S. Mehta, É. Székely, J. Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)," in *Proc. ICASSP*, 2022.

[38] S. Mehta, A. Kirkland, H. Lameris, J. Beskow, Éva Székely, and G. E. Henter, "OverFlow: Putting flows on top of neural transducers for better TTS," in *Proc. INTERSPEECH*, 2023.