Juvela, Lauri; Wang, Xin

Collaborative Watermarking for Adversarial Speech Synthesis

Published in:
ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI:
10.1109/ICASSP48485.2024.10448134

Published: 18/03/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Published under the following license:
Unspecified

Please cite the original version:
Juvela, L., & Wang, X. (2024). Collaborative Watermarking for Adversarial Speech Synthesis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11231-11235). (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10448134

IEEE Copyright Notice

This work has been accepted to the IEEE International Conference on Acoustics, Speech and Signal Processing.

Please cite the published version of the paper.

```
@INPROCEEDINGS{10448134,
  author={Juvela, Lauri and Wang, Xin},
  booktitle={Proc. ICASSP},
  title={Collaborative Watermarking for Adversarial Speech Synthesis},
  year={2024},
  pages={11231-11235},
  doi={10.1109/ICASSP48485.2024.10448134}}
```

COLLABORATIVE WATERMARKING FOR ADVERSARIAL SPEECH SYNTHESIS

Lauri Juvela *

Aalto University, Espoo, Finland
lauri.juvela@aalto.fi

Xin Wang †

National Institute of Informatics, Tokyo, Japan
wangxin@nii.ac.jp

ABSTRACT

Advances in neural speech synthesis have brought us technology that is not only close to human naturalness, but is also capable of instant voice cloning with little data, and is highly accessible with pre-trained models available. Naturally, the potential flood of generated content raises the need for synthetic speech detection and watermarking. Recently, considerable research effort in synthetic speech detection has been related to the Automatic Speaker Verification and Spoofing Countermeasure Challenge (ASVspoof), which focuses on passive countermeasures. This paper takes a complementary view to generated speech detection: a synthesis system should make an active effort to watermark the generated speech in a way that aids detection by another machine, but remains transparent to a human listener. We propose a collaborative training scheme for synthetic speech watermarking and show that a HiFi-GAN neural vocoder collaborating with the ASVspoof 2021 baseline countermeasure models consistently improves detection performance over conventional classifier training. Furthermore, we demonstrate how collaborative training can be paired with augmentation strategies for added robustness against noise and time-stretching. Finally, listening tests demonstrate that collaborative training has little adverse effect on perceptual quality of vocoded speech.

Index Terms— Voice cloning, Generated speech detection, Watermarking, HiFi-GAN, ASVspoof

1. INTRODUCTION

Modern speech synthesis systems have achieved nearly human level naturalness and are capable of zero-shot voice cloning from a few seconds of adaptation data [1]. Open-source implementations, sharing pre-trained models, and good software packaging has made voice cloning with TTS easily accessible also outside the research community [2]. Furthermore, the use of voice cloning technology as a service (for a small monthly subscription fee) has recently emerged as a popular past-time on the internet. With this growing user base, the amount of generated speech in the wild is increasing, which poses a risk of casually created misinformation and malicious deepfakes.

Research on audio deepfake detection focuses mostly on passive protection using machine learning methods, which is also referred to as speech anti-spoofing [3]. This scenario assumes that defenders have no prior knowledge of what attackers will use to generate the audio deepfake. The attackers can use any speech generation models to create the audio deepfake, while the defenders only have a limited number of audio deepfake types in the training set, which does not necessarily cover those from the attackers. Hence, the key question for the defender is how to develop a detection model on the basis of the limited training data and make it generalize to unseen deepfakes.

Research outcomes from, for example, the ASVspoof [4] and Audio Deep synthesis Detection (ADD) challenges [5], have demonstrated some deep learning-based detectors can detect certain unseen deepfakes in the benchmark datasets with error rates smaller than 5%. However, many detectors were found to be vulnerable to spurious features in the training set [4, 6, 7] and generalize poorly to data from different domains [8, 9]. Even though the generalization capability may be improved by including more diverse training data, the long-term equilibrium of this adversarial game of passive audio deepfake detection is yet to be seen. Attackers can always create audio deepfake from newer and stronger speech generation models.

Generative adversarial networks (GANs) [10] take a mirrored perspective on playing the adversarial game of detection and generation. In this setting, a Discriminator network attempts to classify between samples from the real and generated data distributions, while another network, the Generator synthesizes new samples and tries to spoof the discriminator into classifying the generated samples as real. The theoretical equilibrium for the GAN game is that the generator learns to match the real data distribution exactly, and discriminating between real and generated samples becomes impossible [10]. Despite practical challenges with finite model capacity, numerics, and unstable training dynamics, GAN-based synthesis methods [11, 12] can be used for realistic speech waveform synthesis. In particular, HiFi-GAN [13] is widely used as a high-quality neural vocoder in text-to-speech synthesis and voice-conversion.

The adversarial perspective remains highly relevant in defence against black-hat attack scenarios. However, the intended use of a generative model is often not malicious, and deceptive realism is merely a side-product of a high-quality system. In these use cases, the generative model provided is likely happy to comply to any regulation and make an active effort watermark their generated model outputs as such. The relevance of watermarking is amplified by the common use of pre-trained models and hosted services: if a model has built-in watermarking that is not trivial to remove and does not affect the perceptual quality of the output, most users will not make an effort to remove the watermark.

This paper proposes a collaborative training scheme that tasks a generative model to watermark its output to be more easily detectable by a specific classifier without degrading the perceptual quality. The experiments use HiFi-GAN [13] as the generative model and ASVspoof 2021 challenge [4] baseline countermeasure models as watermark detector models. Additionally, the detection performance under additive noise and time-stretching is shown to improve by differentiable augmentation during training. Results show that collaborative training consistently improves the detection performance compared to the corresponding passively trained countermeasure model.

*We acknowledge the computational resources provided by the Aalto Science-IT project.

†This work was supported by JST CREST Grant JPMJCR18A6 and PRESTO Grant JPMJPR23P9, and MEXT KAKENHI Grant 21H04906.

2. RELATED WORK

Watermarks can be evaluated using multiple characteristics, including robustness, perceptibility, and applications [14]. Some applications, such as proof-of-ownership or object identification, require a watermark payload with high enough bit count to encode sufficient information, but this paper focuses on a simple zero-bit watermark: the watermark is present in synthetic speech and not present in natural speech. Sometimes researchers use the terms *watermarking* and *fingerprinting* interchangeably. We use fingerprint to denote the summary of the perceptible information in the carrier speech signal, such as linguistic content, expression, speaker identity, or acoustic channel information. In contrast, a watermark aims to convey hidden information and not be perceptible to the human listener.

Audio watermarking usually consists of an embedder and a detector. The embedder embeds a watermark or message, e.g., pseudo-random numbers or a hash, into the input audio, and the detector verifies the message's existence in a watermarked audio. Typical algorithms using DSP include adding pseudo-noise in the temporal or spectral domain (a.k.a. spread spectrum watermarking [15]), phase coding [16], and echo hiding [17]. Some recent methods also implement the embedded and detector using DNNs [18, 19].

A notable approach that combines DSP and statistics is patchwork watermarking [20]. When implemented in the spectral domain, this algorithm embeds a single bit m in the spectrum of an audio frame by selecting two sets of frequency bins and changing their amplitude values in opposite ways. Let s_k be the spectral amplitude at the k-th frequency bin and $\mathcal{A}$ and $\mathcal{B}$ be the two sets of frequency bins. The algorithm can be written as

$$\begin{cases} s_i \leftarrow \exp(s_i, 1+d), \quad s_j \leftarrow \exp(s_j, 1-d), & \text{if } m = 1 \\ s_i \leftarrow \exp(s_i, 1-d), \quad s_j \leftarrow \exp(s_j, 1+d), & \text{if } m = 0 \end{cases}, \tag{1}$$

where $i \in \mathcal{A}$, $j \in \mathcal{B}$, and d is a strength parameter. The general idea to detect the single bit of message is to compare the mean values of the two sets

$$\begin{cases} \widehat{m} = 1, \text{if } \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} s_i - \frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}} s_j > 0 \\ \widehat{m} = 0, \text{else} \end{cases}. \tag{2}$$

A larger strength d value increases the robustness against distortions but degrades the quality of the watermarked audio.

For generative models in the image domain, watermarking the training data with standard DSP techniques has been shown to transfer to GAN outputs [21]. Yu *et al.* [22] propose a jointly trained watermark encoder-decoder scheme for GANs. This is closely related to our work, both in motivation ("from *passive classifiers* to *proactive fingerprinting*"), and implementation (their Generator acts as a watermark embedding model). Main differences are different domains and our work adds augmentation for robustness.

3. PROPOSED METHOD

Figure 1 depicts an overview of the system. A Generator model takes a mel-spectrogram as input and outputs a corresponding synthetic speech waveform. Detector models all try to classify between real and generated speech, and the training dynamics change based on which role the classifier takes:

1. **Discriminator** is adversarial to the Generator. The Generator attempts to fool the discriminator into classifying generated samples as real.

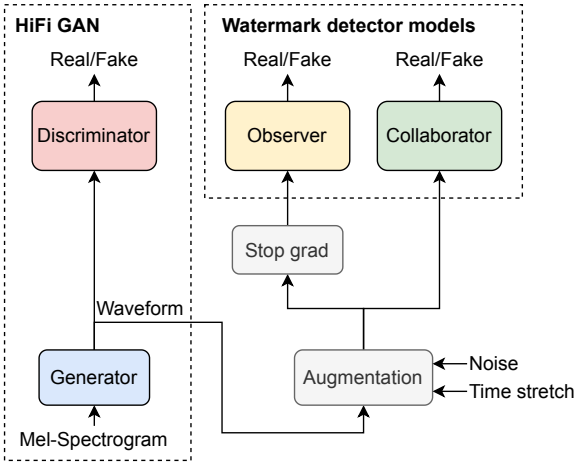

Fig. 1: A Generator can view detector models in three distinct roles: fool the Discriminator to produce more realistic samples, ignore the Observer, or help the Collaborator to extract a watermark from generated speech.

2. **Observer** acts as a passive detector. Gradient flow from the Observer to Generator is detached. This corresponds to traditional ASVspoof countermeasure training.
3. **Collaborator** shares its objective with the Generator. The pair attempt to discover a *watermark* that is embedded into the generated signal to aid in the binary classification task, while not hindering Generator's other objectives.

Notably, the general approach is not specific to GANs or ASVspoof detector models, and can be adapted to a wide range of detectors and generative models. For simplicity, we chose to limit the experiments to neural vocoding instead of end-to-end TTS. Previous research has demonstrated that training ASV Spoofing countermeasures on vocoded speech transfers well to TTS systems using the same vocoders [23].

3.1. Loss functions

Generator and Discriminator loss functions and architectures follow the HiFi-GAN recipe [13]. We use the large Generator model configuration (V1) and base the experiments on the official implementation[1].

Denote generated speech signal as $x_{\text{gen}} = G(m)$, where G is the Generator model, and m is the mel-spectrum of the target waveform. Discriminator, D, is trained with the least-squares GAN loss function, with the target score at one for real samples and at zero for generated

$$\mathcal{L}_D = \mathbb{E}\left[(D(x_{\text{real}}) - 1)^2 + D(x_{\text{gen}})^2\right], \tag{3}$$

where the expectation, $\mathbb{E}$, is approximated by minibatch averages over batch elements and timesteps. Each of the sub-discriminators in the HiFi-GAN discriminator ensemble share the same objective. Similarly, the Generator loss functions are from HiFi-GAN, comprising the adversarial loss

$$\mathcal{L}_{\text{G-adv}} = \mathbb{E}\left[(D(x_{\text{gen}}) - 1)^2\right], \tag{4}$$

feature matching loss at each hidden activation $D_i(\cdot)$

$$\mathcal{L}_{\text{G-FM}} = \sum_i \mathbb{E}\left[(D_i(x_{\text{real}}) - D_i(x_{\text{gen}}))^2\right], \tag{5}$$

[1] https://github.com/jik876/hifi-gan

and L1 regression loss on log-mel-spectrograms

$$\mathcal{L}_{\text{G-mel}} = \mathbb{E}\left[|\log(\boldsymbol{M}\cdot|\text{STFT}(x_{\text{real}})|) - \log(\boldsymbol{M}\cdot|\text{STFT}(x_{\text{gen}})|)\right], \quad (6)$$

where $\boldsymbol{M}$ is a tensorized mel-filterbank matrix.

The watermark detector model, WM, has exactly the same objective as D: assign a high score to x_{real} and a low score to x_{gen}

$$\mathcal{L}_{WM} = \mathbb{E}\left[(WM(x_{\text{real}})-1)^2 + WM(x_{\text{gen}})^2\right]. \quad (7)$$

G can either share the WM objective or ignore it. These two scenarios are called Collaborator and Observer, respectively.

In practice, the system is trained by alternating minibatches that switch between the D objectives and the joint objective of G and WM. In Observer mode, the gradient flow from WM to G is detached and $\mathcal{L}_{WM}$ has no effect on G.

3.2. Watermark detector models

Both ASVspoof 2021 baseline models are designed to operate at at 16 kHz sample rate, whereas the HiFi-GAN defaults to 22.05 kHz sample rate. To maintain compatibility with pre-trained models, we use a differentiable resampling layer between the Generator and watermark detector models. The resampler uses a Hann-windowed sinc interpolation filter with six filter taps.

Pre-trained models use batch normalization in evaluation mode with stored statistics. For joint training, we removed batch normalization layers after identifying they led to poor generalization with collaborative training. Recent research has found that removing batch normalization boosts adversarial training [24], and we speculate the underlying cause may be related due to similar mathematics of adversarial and collaborative training.

LFCC-LCNN is the first detector used in this study, as the baseline of the ASVspoof 2021 challenge[2]. Its front end extracts linear frequency cepstrum coefficients (LFCC) [25], which is similar to the Mel frequency cepstrum coefficients but uses filters placed in equal sizes on a linear scale. The frame length and hop size are 20 and 10 ms, respectively. The FFT size is 1,024, and the number of dimensions is 60, including delta and delta-delta components. The maximum frequency covered by the filter is half of the Nyquist frequency. The back end is a light convolution neural network (LCNN) [26], followed by two recurrent layers using long-short term memory units, global average pooling, and a linear output layer [27].

RawNet2 is another ASVspoof 2021 baseline used in this study is [28]. RawNet2 uses convolution (in DNNs) to implement the differentiable band-pass filters. The convolution kernel, or the filter coefficients, are pre-calculated in the same manner as the windowed-sinc filters, and the cut-off frequencies are co-located with the filter bank in MFCC. The input waveform is processed by the convolution layer and transformed by multiple blocks with 2D convolution, max pooling, and filter-wise feature map scaling. The hidden features are then processed by a recurrent layer with gated recurrent units and a linear output layer. Its configuration follows the official implementation, except that the batch normalization layer is removed. The input waveform is padded or truncated to a fixed length of 65,536.

3.3. Differentiable augmentation

We apply two differentiable augmentation techniques to improve the robustness of the proposed collaborative watermarking method. First, time-stretching is implemented using linear interpolation. The time-scale factor is randomized uniformly between time-scale factors 0.9 and 1.1 and kept constant for each mini-batch. Second, additive noise samples are randomly drawn from the MUSAN database [29] noise subset. For simplicity, we keep the noise level constant at 10 dB SNR. More fine-grained analysis on the effect of varying noise levels is left as future work.

4. EXPERIMENTS

4.1. Datasets

The Voice Cloning Toolkit (VCTK) [30] corpus is used for all the speech data in the experiments. VCTK has a data split protocol for ASVspoof aimed at evaluation purposes [4], but we deemed the training set too small for training the synthesis system. Instead, we opted for a custom 80-10-10% split to training, validation, and test sets. Split details are provided with the source code. Each subset has distinct speakers, but has overlaps in text content. Noise data for augmentation consists of the noise subset from the MUSAN database [29]. We apply a 80-10-10% split and use unseen noise samples during testing.

4.2. Training details

Following HiFi-GAN, we use the AdamW optimizer with learning rate 2e-4, $\beta_1 = 0.8$, $\beta_2 = 0.99$ and learning rate decay with exponential decay factor 0.999. Training segements are padded or randomly cropped to 8,192 samples with LFCC-LCNN and 65,536 samples with RawNet. In Collaborator models, the gradients are allowed to flow back from the WM to G, while in Observer models the gradient is detached. Each model configuration is trained for 50 epochs, amounting to approximately 110k iterations. Note that the HiFi-GAN paper [13] reports training for 2.5M iterations, which does yield higher synthesis quality. This work focuses on demonstrating the relative benefit of collaborative training over baseline observer training in many scenarios. We believe the current setup offers a reasonable trade-off between the computational cost of training many systems, and the perceptual quality of generated speech. The various model configurations were trained on available hardware from a pool of GPUs consisting of NVIDIA A100, V100 and P100 models.

4.3. Patchwork watermark baseline

As a DSP baseline, patchwork watermarking is applied to vocoded speech from a HiFi-GAN trained without detector collaboration. The patchwork watermarking algorithm described in Section 2 is included for reference. The open-sourced toolkit Audiowmark [3] is used due to its high-quality implementation with error-correction in watermarking detection. A fixed-length 128 bit text string is watermarked. The hyper-parameter d was selected for each test utterance through grid search. Given the range between 0.01 and 0.2, the value of d was decided so that the detector outputs the correct watermark without other hypothesis. During test conditions, the watermarked audio is added with noise or stretched before sent to the watermark detector.

4.4. Listening test

We conducted a mean opinion score (MOS) test to evaluate the perceptual effect of the proposed method. Listeners were asked to rate the naturalness of the presented speech samples on a five-point scale ranging from 1 (bad) to 5 (excellent). The listening test stimuli consist of 50 utterances selected randomly from test set. After pooling

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

[3] https://github.com/swesterfeld/audiowmark

Table 1: EER (%) of experiment systems in clean, stretch, noise, and stretch plus noise (S+N) testing conditions. A darker cell color indicates a higher EER value. Each EER was averaged over 20 independent rounds of generation and evaluation. Error rates (%) of spectral-domain patchwork watermarking are shown at the top for reference.

			Clean	Stretch	Noise	S+N				
Spectral patchwork watermarking			0.21	32.65	85.29	98.25				
WM configuration			LFCC-LCNN				RawNet			
Training	Augmentation	Role	Clean	Stretch	Noise	S + N	Clean	Stretch	Noise	S + N
Joint	None	collaborator	1.61	2.95	34.66	42.90	0.12	43.25	44.33	48.62
		observer	3.46	5.13	37.87	46.20	0.34	17.86	35.35	49.65
	Stretch + noise	collaborator	1.05	1.38	15.25	34.94	1.36	2.33	4.03	54.10
		observer	3.73	4.24	27.72	41.35	3.72	4.87	13.31	46.48
Pre-trained	None	collaborator	17.73	20.87	42.13	45.19	10.64	21.36	28.32	51.48
		observer	49.47	49.50	49.05	49.56	47.76	48.35	48.55	50.15
	Stretch + noise	collaborator	32.83	32.79	40.79	44.37	45.91	46.75	47.18	48.40

Table 2: Subjective MOS results with confidence interval (95%). Proposed models use WM as collaborator.

Ref.	Natural recording		4.13 ± 0.11	
	Spectral patchwork		3.49 ± 0.13	
	Baseline HiFi-GAN		3.54 ± 0.13	
Proposed	Training	Augmentation	LFCC-LCNN	RawNet
	Joint	None	3.38 ± 0.17	3.59 ± 0.13
		Stretch + noise	3.46 ± 0.13	3.60 ± 0.13
	Pre-trained	None	3.58 ± 0.13	3.51 ± 0.13
		Stretch + noise	3.46 ± 0.14	3.67 ± 0.12

together test utterances from each system, the stimuli were randomly batched to 100-sample listening sessions, and each batch was rated by at least five listeners on the Prolific crowd-sourcing platform. After screening, 3884 total ratings by 35 listeners were used in the analysis. Table 2 shows MOS values with t-statistic based confidence intervals. Differences between vocoded systems are not statistically significant.

5. RESULTS AND DISCUSSION

Table 1 displays test set equal error rates (EER) for a range of systems. When training watermark detector models jointly with the generator, collaborative training consistently outperforms conventional observer training. This remains consistent over the detector model type, optional augmentation during training, and different test conditions, from clean to time-stretching and/or additive noise. All LFCC-LCNN configurations struggle with addivive noise, whereas collaborative training with RawNet and data augmentation still performs well under noise (EER 4.03). However, combining noise with time-stretching remains a challenging for both LFCC-LCNN and RawNet detectors, even when the models were trained with matching data augmentation.

Pre-trained models struggle as watermark detectors, as listed in the bottom part of Table 1. In collaborative training, the synthesis model manages to embed some detectable information in clean conditions, but detection performance deteriorates when noise is added. In a zero-shot scenario (i.e., pre-trainer observer) the performance is near chance level for both LFCC-LCNN and RawNet. It appears that the Generator's attempt to fool the Discriminator transfers to the ASVspoof pre-trained baselines, which have not been trained on HiFi-GAN vocoded speech.

The DSP-based watermarking algorithm achieves a low error rate in the clean condition since the strength d was decided to allow perfect watermark detection in this condition. However, there are a few utterances in which the embedded watermark failed to be detected given the largest d. Its error rate increases when the watermarked audio is stretched, even when the provided compensation for playback was used. The performance degrades dramatically when additive noise is applied. Note that the patchwork baseline is not one-to-one comparable to the detector models, since baseline watermark uses a 128-bit payload and does not output detection scores required for EER calculation.

A central limitation of the current experimental setup is the focus on neural vocoding. Real-world voice cloning applications use a full text-to-speech or voice conversion system. Nevertheless, we are optimistic on the transferability of training on vocoded speech based on recent results in the ASVspoof scenario [23]. However, security critical applications still need to follow an adversarial countermeasure protocol. Collaborative watermarking is only useful when the generative model user has an incentive to watermark their model outputs. Furthermore, collaborative training is not limited to GANs. Any deep generative model with differentiable sampling (including diffusion models [31, 32]) can collaborate with a detector model to improve the odds of detection.

6. CONCLUSIONS

This paper proposes synthetic speech watermarking scheme based on collaborative training between a generative synthesis model and a watermark detector. The results show that collaborative training consistently improves detection performance compared to baseline passive countermeasure training. Source code and demonstration samples are available for readers[4]. Future work includes extending the study to full text-to-speech voice cloning systems, and developing specialized architectures for the speech watermarking task. Further, more informative watermarks are appealing: it would be useful to know who generated the samples, or what data was used to train the model.

[4] https://ljuvela.github.io/CollaborativeWatermarkingDemo/

7. REFERENCES

[1] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[2] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

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

[4] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. 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, 2023.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "ADD 2022: The first Audio Deep Synthesis Detection Challenge," in *Proc. ICASSP*, May 2022, pp. 9216–9220.

[6] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn?," in *Proc. ASVspoof Challenge workshop*, 2021, pp. 55–60.

[7] Y. Zhang, W. Wang, and P. Zhang, "The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System," in *Proc. Interspeech*, 2021, pp. 4279–4283.

[8] D. Paul, M. Sahidullah, and G. Saha, "Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora," in *Proc. ICASSP*, 2017, pp. 2047–2051.

[9] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?," *Proc. Interspeech*, pp. 2783–2787, 2022.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[11] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," in *Proc. Interspeech*, 2019, pp. 694–698.

[12] R. Yamamoto, E. Song, and J.-M. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," in *Proc. Interspeech*, 2019, pp. 699–703.

[13] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[14] F. Petitcolas, "Watermarking schemes evaluation," *IEEE Signal Processing Magazine*, vol. 17, no. 5, pp. 58–64, 2000.

[15] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," in *Proc. ICIP*, 1996, vol. 3, pp. 243–246.

[16] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM systems journal*, vol. 35, no. 3.4, pp. 313–336, 1996.

[17] D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Information Hiding: First International Workshop Cambridge, UK, May 30–June 1, 1996 Proceedings 1*. Springer, 1996, pp. 295–315.

[18] K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski, "Robust speech watermarking by a jointly trained embedder and detector using a DNN," *Digital Signal Processing*, vol. 122, pp. 103381, 2022.

[19] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "Wavmark: Watermarking for audio generation," *arXiv preprint arXiv:2308.12770*, 2023.

[20] M. Steinebach, *Digitale Wasserzeichen fuer Audiodaten*, Shaker, 2004.

[21] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *Proc. ICCV*, 2021, pp. 14448–14457.

[22] N. Yu, V. Skripniuk, D. Chen, L. S. Davis, and M. Fritz, "Responsible disclosure of generative models using scalable fingerprinting," in *Proc. ICLR*, 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, pp. 1–5.

[24] H. Wang, A. Zhang, S. Zheng, X. Shi, M. Li, and Z. Wang, "Removing batch normalization boosts adversarial training," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23433–23445.

[25] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015, pp. 2087–2091.

[26] 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.

[27] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Interspeech*, 2021, pp. 4259–4263.

[28] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. ICASSP*, 2020, pp. 6369–6373.

[29] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015.

[30] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[31] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. ICLR*, 2021.

[32] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, 2021.