Neural Semi-fragile Watermarking for Proactive Deepfake Speech Detection

Dohyun Yoon and Tomoki Toda
Nagoya University, Japan
E-mail: yoon.dohyun@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—Proactive methods for detecting deepfake speech, which utilize audio watermarks, have been studied as an alternative to passive detection methods, which are generally considered weak at generalizing to unseen spoofing attacks. However, existing proactive methods focus mainly on the imperceptibility of watermarks and their detection accuracy. Consequently, attempts to link the deepfake generation process with the training of the watermarking system are limited. In this paper, we introduce a novel watermark distortion process for generalizing to diverse speech generation systems. This approach allows the method to generate semi-fragile watermarks that are vulnerable to the watermark distortion but robust to various acoustic distortions. Experiments using open-source speech generation systems demonstrate the effectiveness of the method for deepfake speech detection.**

I. Introduction

Deepfake speech refers to audio content created by text-to-speech (TTS) or voice conversion (VC) systems based on deep generative methods [1]. In recent years, due to the rapid development of deep learning algorithms and model architectures, it has become increasingly difficult to distinguish between fake and genuine media [2]. Deepfake speech can undermine the authenticity of content, leading to the spread of disinformation, identity abuse, and privacy violations [3], [4]. As the technologies used to create deepfake speech become more sophisticated, reliable ways to detect them are essential to maintain trust in digital communications.

Current approaches for the deepfake speech detection are based on a two-class classification [5], [6], which has been shown not to generalize well [7]. In order to successfully learn the features that distinguish real from fake speech, a large dataset of natural and synthetic speech from a variety of sources is required. While previous approaches are reasonably effective under controlled conditions, it is difficult to deal with unseen data, such as synthesized speech from newer and more advanced deep generative networks [8]. The fundamental limitation of supervised learning-based detection methods—being forced to struggle with generalization—reveals a critical weakness and highlights the need for specialized solutions that can adapt to the evolving nature of deepfake speech generation technologies.

Audio watermarking based on deep neural networks is a useful way for rights holders to preserve the integrity of their original content by embedding an imperceptible signal into the content [9], [10]. Proactive methods in the context of watermarking for deepfake speech detection refer to preventive techniques that embed identifying information into the original speech data prior to potential tampering. While there are several possible ways to design the watermark for the deepfake detection, we focus on the possibility of using the semi-fragile watermarking in this paper. The properties of the semi-fragile watermark required by our proactive method are that the watermark must be robust to "weak" distortions that can be caused by user actions, such as background noise, but easily destroyed by "strong" distortions that can be caused by third-party actions, such as TTS or VC. In other words, the watermark is used to guarantee the originality. In the following, we will use the term "acoustic distortion" for the former and "watermark distortion" for the latter. These properties can be used to distinguish between watermarked original speech and (watermark-distorted) deepfake speech. However, current proactive methods suffer from limitations in their lack of consideration of the process of generating deepfake speech. More details on the existing methods will be provided in Section II.

In this paper, we propose a novel proactive deepfake speech detection method that utilizes semi-fragile watermarking via simultaneous incorporation of the watermark distortion module and the acoustic distortion module into the training phase of our watermark encoder and detector. Specifically, our method induces the watermark to be undetectable in situations where the watermark is distorted by the watermark distortion process. These are realized by primitive operations of speech modification, such as fragmenting and concatenating multiple watermarked speech samples, or by overlapping them, to address wide varieties of the deepfake speech generation technologies. At the same time, the objective function that takes into account auditory features and the acoustic distortion module is included in the training phase to ensure the stealthiness and robustness of the watermark. The experimental results show that our encoder is able to produce watermarks that are not easily perceivable by humans, while being resistant to several types of acoustic distortion. Furthermore, the experimental results on the use of open-source TTS and VC models (which are easily accessible to those who want to create deepfake speech) show that our detector can sufficiently distinguish between natural and synthetic speech.

II. Related works

Collaborative watermarking [11], which simultaneously trains a speech synthesis system and a passive deepfake detector, enables the system to generate "detector-visible"

979-8-3315-7206-8/25/$31.00

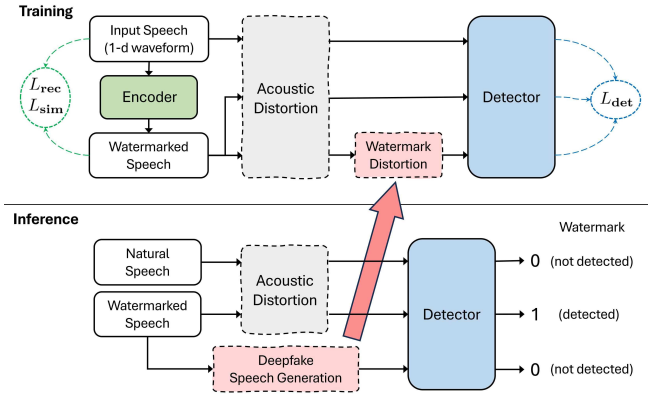

Fig. 1. Training and inference pipeline of our proposed networks. A simplified version of the watermark distortion process that can occur in deepfake speech generation is used in the training phase.

watermarks. This allows us to overcome the limitations of data-dependent supervised learning approach and improve detection accuracy compared to conventional passive methods. However, it requires individual training for each speech synthesis system, and the quality of the watermarked output is strongly affected by the performance of the system. WavMark [12] proposes a proactive method that can maintain the quality of the original speech through an independent watermark encoder and decoder structure utilizing invertible neural networks. AudioSeal [13] employs an encoder-decoder based on a neural audio codec structure that enables sample-level watermark embedding and detection, and generates watermarks that are more perceptually similar and robust to various acoustic distortions. Although these methods have demonstrated satisfactory results in terms of audio watermarking, they have not yet fully addressed the underlying mechanism by which high-level processing, such as deepfake manipulation, can affect the embedded watermark.

III. Proposed Method

A. Audio watermarking for deepfake speech detection

The main goal of audio watermarking based on an encoder-detector structure is as follows: The encoder must generate a watermark that is robust to acoustic distortions, yet subtle enough not to degrade the auditory characteristics of the original input speech, and the detector must be able to recognize the existence of the watermark. Figure 1 shows our proposed method. In the task of our proposed proactive watermarking for deepfake speech detection, we take advantage of the property that the semi-fragile watermark embedded in the speech is modified by deepfake processing, so that the detector is unable to detect the watermark. The overall framework consists of the encoder and the detector to be trained, and modules to perform acoustic distortions and watermark distortions.

B. Encoder and detector

The encoder utilizes a Wave-U-Net [14] structure, which is a specialized variant of U-Net for one-dimensional (1-d) data, since both input and output are 1-d time-domain audio signals. The detector requires a powerful network that can tell the difference between two similar speech waveforms (input and watermarked), so we use a combination of discriminators that are used to train neural vocoders, especially UnivNet's [15] Multi-Resolution Spectrogram Discriminators (MRSDs) and Multi-Period Waveform Discriminators (MPWDs).

C. Distortions

For each training step, one random acoustic distortion $\Phi(\cdot)$ and one random watermark distortion $\Theta(\cdot)$ are selected and applied as follows: $z_{\text{a}} = \Phi(z), z_{\text{w}} = \Theta(z), z_{\text{aw}} = \Theta(\Phi(z))$, where z is either original speech or watermarked speech.

1) Acoustic distortions: Acoustic distortions are applied in training and inference to verify that the watermark is preserved even after the process. Note that these distortions are also applied to the input speech that does not contain a watermark, since the detector may be able to distinguish between the raw speech and the distorted speech. Each of the acoustic distortions, inspired by WavMark, is implemented using differentiable functions provided by PyTorch [16] library as follows:

- ND: No distortion
- BN: Add a background noise with a random amount between 30 and 60 dB SNR
- GN: Add a Gaussian noise with a random amount between 30 and 60 dB SNR
- LP: Low-pass filtering with a cutoff frequency of 5 kHz
- HP: High-pass filtering with a cutoff frequency of 1 kHz
- BP: Band-pass filtering with a central frequency of 3 kHz and Q factor of 0.707
- MS: Median smoothing with a random window size of between 2 and 10
- RL: Resample to 16 kHz and then back to 24 kHz
- RH: Resample to 48 kHz and then back to 24 kHz
- AL: Reduce audio amplitude by a factor of 0.9
- AH: Increase audio amplitude by a factor of 1.1
- TS: Randomly adjust audio length between 0.9 and 1.1
- EA: Add an echo by a factor of 0.3 with a gap of 0.1 second
- SS: Randomly change 0.1% of samples to 0
- CO: Encode to Opus format and then back to WAV format
- CM: Encode to MP3 format and then back to WAV format
- CR: Crop first 1,000 and last 1,000 samples.

2) Watermark distortions: In order to generate watermarks that are vulnerable to deepfake manipulation, it is advisable to include actual generative systems while training the networks. However, it is not feasible to cover a large number of existing systems. Since speech synthesis or conversion is essentially a process of generating new data from existing ones, the most basic information, such as whether the watermark is mixed (distorted) or not, can be a useful clue if multiple watermarked speech samples are used to generate deepfakes. The processes of distorting watermarks are inspired by, and simplified versions of, two classical speech synthesis methods:

concatenative synthesis [17] based on the concatenation operation and statistical synthesis [18] based on the overlapping (i.e., averaging) operations. They are implemented as follows:

- Concatenation: assigns a random masking sequence to each speech data and performs the following operation:

$$y_{\mathrm{w}} \leftarrow y \otimes (1 - \mathrm{mask}(y)) + y' \otimes \mathrm{mask}(y), \quad (1)$$

  where $\mathrm{mask}(\cdot)$ generates mask bits, each corresponding to 4096 samples of the input utterance, then extended by the length of y, where y and y' are the same speech with two different watermarks. The element-wise multiplication $\otimes$ takes samples of y' where the mask bit of y is 1, and leaves it unchanged if the mask bit is 0.
- Overlapping: performs overlapping for each speech data, based on a random factor α between 0.3 and 0.5 as follows:

$$y_{\mathrm{w}} \leftarrow \alpha \cdot y + (1 - \alpha) \cdot y'. \quad (2)$$

D. Training objectives

The encoder and detector are simultaneously trained for the goals of the proactive watermarking for deepfake speech detection. First, the watermarked speech is made close enough to the input signal that the difference is unnoticeable to the human ear by the reconstruction loss L_{rec} and the similarity loss L_{sim}:

$$L_{\mathrm{rec}} = \frac{1}{2}\left[\mathrm{L1}(x, y) + \mathrm{L1}(\Gamma_{f,t}(x), \Gamma_{f,t}(y))\right], \quad (3)$$

$$L_{\mathrm{sim}} = 1 - \cos(x, y), \quad (4)$$

where x is the raw speech, and y is the watermarked speech, that is, $y = E(x)$, if the encoder is represented by the function $E(\cdot)$. In Equation 3, $\mathrm{L1}(\cdot, \cdot)$ indicates the L1 loss function, and $\Gamma(\cdot)$ is a function that calculates the Gammatonegram [19] computed by passing the audio signal through a gammatone filterbank that takes into account human auditory characteristics on a frame-by-frame basis, where f and t are the frequency bins and frame units of the Gammatonegram, respectively. In Equation 4, $\cos(\cdot, \cdot)$ denotes the cosine similarity. The encoder can output a signal that is acoustically similar to the input speech by minimizing these two objective functions. Then, we adjusted the object of LSGAN [20] as the detection loss L_{det} so that the detector outputs a low score for a missing or mixed watermark, and a high score for a present watermark:

$$L_{\mathrm{det}} = \frac{1}{3N}\sum_{i}^{N}\left[D_i(x_{\mathrm{a}})^2 + (D_i(y_{\mathrm{a}}) - 1)^2 + D_i(y_{\mathrm{aw}})^2\right], \quad (5)$$

where N is the total number of MRSDs and MPWDs, and $D_i(\cdot)$ is the i-th discriminator among them. The final objective function $\mathrm{L}_{\mathrm{total}}$ optimized during the training phase is represented as the sum of the three losses:

$$\mathrm{L}_{\mathrm{total}} = L_{\mathrm{rec}} + \lambda_0 L_{\mathrm{sim}} + \lambda_1 L_{\mathrm{det}}, \quad (6)$$

where λ_0 and λ_1 are hyperparameters.

IV. Experimental evaluations

A. Experimental conditions

We used three datasets for our experiments: English multi-speaker CSTR VCTK corpus [21], AudioSet [22], and MUSAN [23]. From the VCTK corpus, we used 40,327 sentences from 99 speakers for training, 2,309 sentences from 5 speakers for validation and 250 sentences from 5 speakers for evaluation. We selected 231 audio recordings from "speech" subset of the AudioSet for evaluation only. This was done by two preprocessing steps: utilizing Silero VAD [24], an open-source voice activity detector, to extract speech segments with a length of longer than 2 seconds, and then manually filtering out those segments that were actually recognized as human speech. The purpose of using this dataset is to verify that our proposed method is applicable to untrained languages and recording environments. 823 real-world environmental sounds from the "noise" subset of the MUSAN were used to add background noise to the speech (BN). All data was downmixed to a mono channel and resampled from 48 kHz or 44.1 kHz to 24 kHz sampling frequency. For the training phase, a sequence of 49,152 samples (corresponding to about 2 seconds) was randomly selected from each data for batch computation, and if the data was shorter than that, it was selected after repeating the data.

The encoder is composed of 12 pairs of residual blocks, each comprising convolution layers that perform upsampling and downsampling, with 24 filters. The detector was set up as MPWDs with 3 different reshape factors and MRSDs with 5 different spectrogram resolutions, following the preferences of the original paper. Given these settings, our encoder and decoder were trained with the Adam optimizer for a total of 400,000 steps, a batch size of 16, and hyperparameters of the objective function $\lambda_0 = 1.0$ and $\lambda_1 = 5.0$.

B. Evaluation metrics and situation design

For the objective evaluation of the watermarked speech without acoustic distortion, the sound quality was measured by means of signal-to-noise ratio (SNR), scale-invariant SNR (SI-SNR), spectrogram distortion on Gammatonegram (GD), perceptual evaluation of speech quality [25] (PESQ), and short-time objective intelligibility [26] (STOI) in comparison to the original speech. Except for GD, higher values indicate better quality.

The performance of the detector was evaluated in two scenarios: (1) how well it can distinguish the watermarked speech from the original speech under the same acoustic distortion settings as in the training phase, and (2) how well it can distinguish the deepfake speech from the watermarked speech, assuming that the watermark is distorted when the deepfake is created using the watermarked speech. We chose the area under the curve (AUC) and the equal error rate (EER), as well as the accuracy (ACC) and the F1 score (F1), calculated at a threshold of approximately 0.325, the average value of the EER thresholds among the acoustic distortions, as the evaluation metrics.

TABLE I

THE OBJECTIVE EVALUATION RESULT OF THE QUALITY OF THE WATERMARKED SPEECH BY COMPARING TO THE ORIGINAL SPEECH. THE RESULT OF AUDIOSEAL [13] IS ALSO SHOWN JUST AS A REFERENCE ALTHOUGH IT IS DIFFICULT TO FAIRLY COMPARE IT WITH OUR PROPOSED METHOD BECAUSE THE AMOUNT OF TRAINING DATA FOR AUDIOSEAL IS MUCH LARGER THAN OURS.

method	**SNR** [dB]	**SI-SNR** [dB]	**GD** [dB]	**PESQ**	**STOI**
ours	15.03	17.58	2.89	4.02	0.988
(AudioSeal)	26.02	26.07	0.51	4.30	0.990

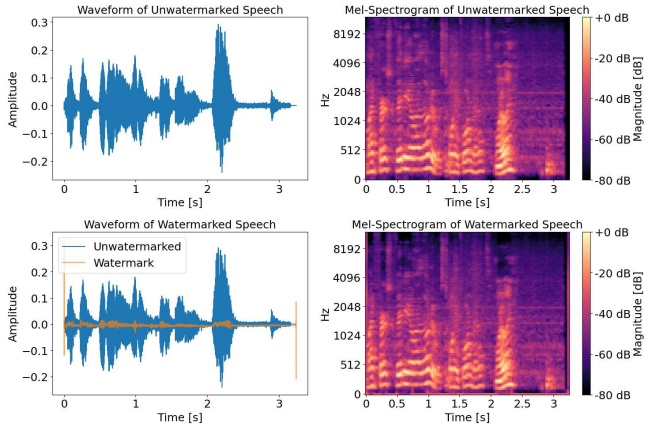

Fig. 2. An example of the watermarked speech. The top row shows a waveform and a Mel-spectrogram of the original speech, and the bottom row shows those of the watermarked speech. The orange color represents the watermark, i.e., the difference between the two speech.

In scenario (2), the experiment basically assumed voice cloning using watermarked speech. The pre-trained XTTSv2, YourTTS [27] for text-to-speech synthesis, and FreeVC [28] for voice conversion provided by the Coqui TTS [29] toolkit, were used to generate deepfake speech. For the XTTSv2 and YourTTS, the source text providing the models with non-speaker-related content was randomly selected from the VCTK corpus, while the target speech providing the models with speaker-related content was selected from the watermarked speech. In the case of FreeVC, a target speech was selected from the watermarked speech, while a source speech was randomly selected not only from the original, but also from the watermarked speech, in order to analyze the case of more significant watermark distortion. This experiment was run 10 times and subsequently the averages were calculated.

C. Speech quality evaluation results

Table I shows the quality of the watermarked speech by the proposed method. The result of AudioSeal is also shown as a reference. Each score was calculated with the evaluation dataset downsampled to a 16 kHz sampling rate in order to match the open-source environment provided by AudioSeal. It is important to note that AudioSeal is not optimized for semi-fragile watermarks, i.e., it is relatively straightforward in comparison to our method. Besides, the amount of training data used in AudioSeal is much larger than our proposed system.

TABLE II

DETECTOR PERFORMANCE UNDER EACH ACOUSTIC DISTORTION. (BOLD INDICATES POOR RESULTS.)

acoustic distortion	**AUC**	**EER [%]**	**ACC [%]**	**F1-score**
ND	1.0	0.208	99.646	0.995
BN	1.0	0.187	99.625	0.995
GN	1.0	0.769	99.360	0.990
LP	1.0	0.416	99.645	0.996
HP	**0.746**	**33.056**	**74.890**	**0.688**
BP	**0.749**	**32.017**	**69.784**	**0.677**
MS	0.966	7.256	91.709	0.937
RL	1.0	0.208	99.125	0.998
RH	1.0	0.208	99.666	0.995
AL	1.0	0.832	99.630	0.995
AH	1.0	0.208	99.823	0.998
TS	1.0	0.083	99.822	0.997
EA	1.0	0.0	99.969	1.0
SS	1.0	0.104	99.687	0.995
CO	1.0	0.208	99.635	0.995
CM	1.0	0.416	99.696	0.996
CR	1.0	0.624	99.205	0.987
Mean	0.968	4.518	95.932	0.955

TABLE III

DETECTOR PERFORMANCE UNDER EACH GENERATIVE MODEL. THE SECOND AND THIRD COLUMNS INDICATE THE WATERMARKING STATUS OF THE SOURCE AND TARGET SPEECH, WHERE x DENOTES ORIGINAL SPEECH AND y DENOTES WATERMARKED SPEECH. (BOLD INDICATES BETTER RESULTS.)

deepfake model	**source**	**target**	**AUC**	**EER [%]**	**ACC [%]**	**F1-score**
XTTSv2	-	y	1.0	0.0	99.491	0.995
YourTTS	-	y	1.0	0.0	99.491	0.995
FreeVC	x	y	0.914	11.206	87.620	0.889
	y	y	**0.958**	**6.778**	**92.069**	**0.926**

Consequently, while there is no necessity to surpass the result of AudioSeal, our method can still generate watermarks that are sufficiently imperceptible to the human ear (PESQ > 4.0 and STOI > 0.98). The comparison of waveforms and Mel-spectrograms of the original and watermarked speech is shown in Figure 2. We have often found the presence of a pulse at the beginning and ending of the watermarked speech. As will be demonstrated in the subsequent section, the acoustic distortion CR does not affect the detection performance while removing these pulses. Therefore, we can implement post-processing to remove them by simply discarding 300 samples (equivalent to 0.025 seconds at a 24 kHz sampling rate) from the starting and ending points of the watermarked speech as the default output of our encoder.

D. Watermark detection evaluation results

The evaluation results for scenarios (1) are shown in Table II. From the results, we found that the watermarks generated by our method are stable in the majority of acoustic distortion conditions. However, the watermarks were not effectively recognized when the watermarked speech was passed through a filter above frequency of 1 kHz (HP, BP). In general,

human speech has a high power in the low frequency band, which makes it a reasonable place to embed watermarks. Nevertheless, future research on generating watermarks that can be embedded in higher frequency bands is necessary to enhance the stability of our method.

Table III shows that when a watermarked speech is used to inform the deepfake generative model about the speaker identity as the target speech, the distortion of the watermark can be detected. The results show that a significant amount of watermark is destructed in the cases of speech synthesis. On the other hand, in the cases of voice conversion, the performance degradation was relatively pronounced, but the overall metrics improved when both source and target speech are watermarked and mixed. These results suggest that our proposed method effectively manages the distorted watermarks, but further improvements are necessary.

V. Conclusion and limitations

In this paper, we proposed a method for discriminating deepfake speech by embedding and detecting the semi-fragile watermarks based on the encoder-detector network. Experimental results showed that the watermarks generated by our approach do not significantly degrade the auditory quality of the original content and are robust to various acoustic distortions that may occur in practice. Moreover, we introduced the concept of watermark distortion and applied it to the training of our model with the objective of improving the explainability of the proactive deepfake speech detection task. It is our hope that this method will contribute to future research in the field of audio watermarking aimed at generalization of deepfake generation.

However, since the deepfake detection task depends on the reliability of the detector, further study is necessary to enable flawless detection under more severe and unseen conditions. In addition, the neural watermarking method fundamentally can be vulnerable to malicious attacks such as model extraction [30] and adversarial attacks [31], so countermeasures should be prepared for real-world applications.

Acknowledgment

This work is partly supported by projects, JPNP25006, commissioned by NEDO and JST AIP Acceleration Research JPMJCR25U5, Japan.

References

[1] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," *Frontiers in Big Data*, vol. 5, p. 1 001 063, 2023.

[2] A. Triantafyllopoulos, B. W. Schuller, G. İymen, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, 2023.

[3] R. Katarya and A. Lal, "A study on combating emerging threat of deepfake weaponization," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, IEEE, 2020, pp. 485–490.

[4] N. Amezaga and J. Hajek, "Availability of voice deepfake technology and its impact for good and evil," in *Proceedings of the 23rd Annual Conference on Information Technology Education*, 2022, pp. 23–28.

[5] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, 2022.

[6] X. Liu, X. Wang, M. Sahidullah, *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] H. Shim, R. G. Hautamäki, M. Sahidullah, and T. Kinnunen, "How to construct perfect and worse-than-coin-flip spoofing countermeasures: A word of warning on shortcut learning," *arXiv preprint arXiv:2306.00044*, 2023.

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

[9] K. Pavlović, S. Kovačević, and I. Durović, "Speech watermarking using deep neural networks," in *2020 28th Telecommunications Forum*, IEEE, 2020, pp. 1–4.

[10] A. Patil and R. Shelke, "Digital audio watermarking: Techniques, applications, and challenges," *Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 2*, pp. 679–689, 2022.

[11] L. Juvela and X. Wang, "Collaborative watermarking for adversarial speech synthesis," in *Proc. ICASSP*, arXiv preprint arXiv:2309.15224, 2024.

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

[13] R. San Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon, and T. Tran, "Proactive detection of voice cloning with localized watermarking," in *International Conference on Machine Learning*, vol. 235, 2024.

[14] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[15] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint arXiv:2106.07889*, 2021.

[16] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[17] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, IEEE, vol. 1, 1996, pp. 373–376.

[18] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *2007 IEEE International*

Conference on Acoustics, Speech and Signal Processing-ICASSP'07, IEEE, vol. 4, 2007, pp. IV–1229.

[19] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Cascade classifiers trained on gammatonegrams for reliably detecting audio events," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2014, pp. 50–55.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[21] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https : / / api . semanticscholar . org / CorpusID : 213060286.

[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 776–780.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, https://github.com/snakers4/silero-vad, 2021.

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings*, IEEE, vol. 2, 2001, pp. 749–752.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] 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.

[28] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[29] T. C. T. Team, *Coqui tts: A deep learning toolkit for text-to-speech, battle-tested in research and production*, https://github.com/coqui-ai/TTS, 2021.

[30] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[31] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.