arXiv:2504.14832v2 [cs.CR] 21 Jun 2025

Protecting Your Voice: Temporal-aware Robust Watermarking

Yue Li, Weizhi Liu, Dongdong Lin, Hui Tian, *Senior Member, IEEE*, Hongxia Wang, *Member, IEEE*

***Abstract*—The rapid advancement of generative models has led to the synthesis of real-fake ambiguous voices. To erase the ambiguity, embedding watermarks into the frequency-domain features of synthesized voices has become a common routine. However, the robustness achieved by choosing the frequency domain often comes at the expense of fine-grained voice features, leading to a loss of fidelity. Maximizing the comprehensive learning of time-domain features to enhance fidelity while maintaining robustness, we pioneer a temporal-aware robust watermarking (*True*) method for protecting the speech and singing voice. For this purpose, the integrated content-driven encoder is designed for watermarked waveform reconstruction, which is structurally lightweight. Additionally, the temporal-aware gated convolutional network is meticulously designed to bit-wise recover the watermark. Comprehensive experiments and comparisons with existing state-of-the-art methods have demonstrated the superior fidelity and vigorous robustness of the proposed *True* achieving an average PESQ score of 4.63.**

***Index Terms*—Audio watermarking, Temporal-aware watermarking, Proactive forensics**

I. Introduction

GENERATIVE models have significantly advanced text-to-speech and text-to-music synthesis technologies [1]–[4], breaking the high-tech barrier of voice cloning technology. With their ability to closely mimic voices, these innovations raise increasing concerns, particularly in facilitating events such as fraud [5] and misinformation campaigns [6]. In response, governments and regulatory bodies are developing policies (CHN [7], EU [8] and USA [9]) aimed at regulating AI-generated content (AIGC) to mitigate these risks.

All of the aforementioned regulatory policies emphasize the responsibility of AIGC companies to implement specific marks for generated content. This underscores the significance of *watermarking* technology as an essential solution for proactive regulation of deepfake content [10]–[13]. As a consequence, research has increasingly shifted from traditional handcrafted watermarking methods to deep-learning-based approaches in the field of multimedia watermarking.

In the field of audio and voice watermarking, it has seemingly become an established principle that achieving good robustness necessitates a frequency domain transformation (FDT). As depicted in the upper branch of Fig. 1, this holds true both in the earlier era of handcrafted watermarking [14]–[19] and in the current epoch dominated by deep-learning-based approaches [20]–[27]. Concretely, Liu et al. [23] embedded the watermark into the approximate coefficients derived from the Discrete Wavelet Transform (DWT). Fully leveraging the short time window properties of the Short-Time Fourier Transform (STFT), Liu et al. [22] employed frequency features for robust watermarking. Additionally, the advantageous properties of STFT magnitudes have motivated subsequent works, including Chen et al. [25], Pavlovic et al. [26], and O'Reilly et al. [27], to embed watermarks into these magnitudes. For DL-based approaches mentioned above, another notable observation is that achieving high robustness requires more than just FDTs—the use of an attack simulator (AS) is essential. This observation prompts us to rethink: *Between FDT and AS, which serves as the cornerstone of robustness? Can robust performance be achieved with the AS alone, in the absence of FDT?* If FDT proves non-essential, it may be feasible to shift the watermark embedding process from the frequency domain to the temporal domain, potentially enabling fine-grained manipulation of timbre features for improved fidelity.

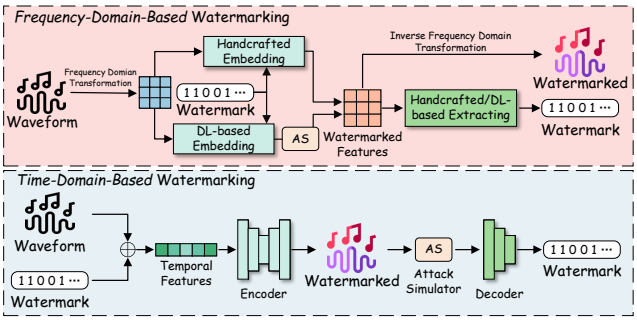

Fig. 1. Two branches of audio watermarking methods. Upper: Frequency-domain watermarking dominates existing SOTA methods. Lower: The proposed temporal-aware watermarking preserves robustness without compromising fine-grained temporal features.

To validate the feasibility of the aforementioned assumption, we propose a temporal-aware and robust watermarking (**True**) method tailored for speech and singing voices. The proposed method directly embeds the watermark into the voice waveform, as illustrated in the lower branch of Fig. 1, thereby avoiding the timbre detail loss typically associated with frequency-domain transformations in conventional frequency-domain-based watermarking approaches. Meanwhile, the AS has borne the responsibility of ensuring the watermark's robustness. Thus, the contributions can be boiled down to:

- *New Paradigm.* We pioneer a novel temporal-aware watermarking method, *True*, designed to proactively protect the copyright of singing and speech voices, with only AS

Yue Li, Weizhi Liu, Dongdong Lin and Hui Tian are with the College of Computer Science and Technology, National Huaqiao University, Xiamen 361021, China, and also with the Xiamen Key Laboratory of Data Security and Blockchain Technology, Xiamen 361021, China (e-mail: liyue_0119@hqu.edu.cn; lwzzz@stu.hqu.edu.cn; dongdonglin8@gmail.com; htian@hqu.edu.cn).

Hongxia Wang is with the School of Cyber Science and Engineering, Sichuan University, Chengdu 610207, China (e-mail:hxwang@scu.edu.cn)

Corresponding author: Hui Tian

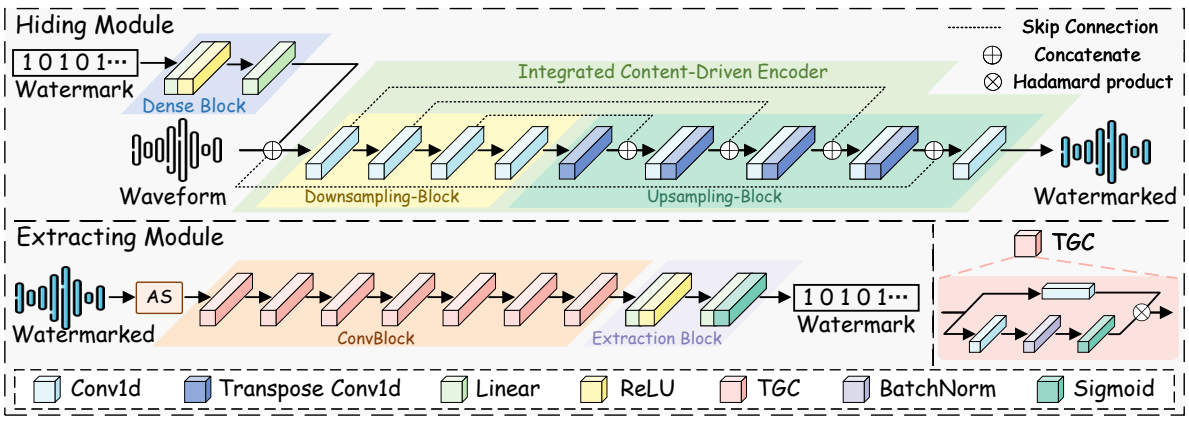

Fig. 2. **Detailed Architecture of True.** The hiding module consists of a dense block and an integrated content-driven encoder for watermarking. The extracting module includes the attack simulator (AS), the convolutional block (ConvBlock), and an extraction block. The ConvBlock is composed of several temporal-aware gated convolution (TGC) networks.

needed to guarantee the robustness of this paradigm.

- *Novel Architecture.* To preserve fine-grained voice features, an integrated content-driven encoder is proposed for watermarking. For bit-wise extraction, a temporal-aware gated convolutional network (TGC) is designed.
- *Sound Performance.* Comprehensive experiments and comparisons with state-of-the-art (SOTA) methods demonstrate the superior fidelity and enhanced robustness.

II. Methodology

The proposed method, *True*, aims to directly leverage the time-domain features from the voice waveform, achieving the trade-off between robustness and fidelity for the watermarked waveform. To this end, the method encompasses two main components: the hiding module (HM) and the extracting module (EM), as illustrated in Fig. 2. All components are jointly trained through an optimization strategy, the details of which are elaborated below.

A. Hiding Module

In the hiding module, two primary objectives are addressed. The first is to convert the watermark format for more effective feature extraction. The second is to directly encode the temporal-domain waveform and watermark, thereby generating the watermarked signal. To achieve the first objective, the dense block $\mathbf{DB}$ is designed, whereas the integrated content-driven encoder (ICDE) $\mathbf{E}$ is constructed to fulfill the second. The detailed architectures and watermarking process are presented as follows.

Architectures: The dense block consists of two fully connected (FC) layers interleaved with a ReLU activation function, enabling it to process the watermark of varying lengths effectively. MobileNetV2 [28] indicates that ReLU-like activation functions can result in substantial information loss, particularly for low-dimensional features. Given the critical role of low-dimensional features in waveform reconstruction, the proposed ICDE architecture eschews normalization layers and activation functions. Instead, it employs a downsampling block consisting of four one-dimensional convolutional (Conv1d) layers, followed by an upsampling block with four transpose Conv1d layers and four additional Conv1d layers. This fully convolutional design not only preserves low-dimensional features but also ensures the lightweight nature of the ICDE.

Watermarking Process: Given a watermark $\mathbf{w} \in \{0,1\}^l$, where l is the length of the watermark. $\mathbf{DB}$ is utilized to transform the watermark into the latent variable $\sigma_{\mathbf{w}}$:

$$\sigma_{\mathbf{w}} = \mathbf{DB}(\mathbf{w}) \in \mathbb{R}^{B\times C\times L}, \tag{1}$$

where B represents the batch size, C denotes the channel of the waveform, and L is the length of the waveform. Then, the latent variable $\sigma_{\mathbf{w}}$ is concatenated with the waveform $\mathbf{s}$ to acquire the final input σ of the encoder:

$$\sigma = \mathbf{s} \oplus \sigma_{\mathbf{w}} \in \mathbb{R}^{B\times C\times L}, \tag{2}$$

where $\oplus$ represents channel concatenation. The encoder $\mathbf{E}(\cdot)$ takes σ as input to reconstruct the watermarked waveform $\hat{\mathbf{s}}$. The complete watermarking process can be formalized as:

$$\hat{\mathbf{s}} = \mathbf{E}(\mathbf{s} \oplus \mathbf{DB}(\mathbf{w})). \tag{3}$$

B. Extracting Module

The extracting module consists of a convolutional block (ConvBlock) $\mathbf{D}$ and an extraction block $\mathbf{EB}$. The $\mathbf{D}$ aims to isolate the watermark features from the watermarked waveform, while the $\mathbf{EB}$ reconstructs the extracted watermark. Specific architectures of the module are available here.

Architectures: Gating mechanism [29] has demonstrated its effectiveness in feature extraction for language processing tasks [30] and classification tasks [31]. To fully exploit this advantage, we have designed the gating mechanism to capture temporal domain features, resulting in a Temporal-aware Gated Convolutional Network (TGC), which incorporates the ConvBlock in the extracting module. The TGC follows a dual-branch structure, where the main branch consists of a sequential arrangement of a Conv1d layer, batch normalization, and a sigmoid activation to implement the gating mechanism. In contrast, the shortcut branch contains only a Conv1d layer for acquiring fine-grained features from the input. The outputs of these two branches are merged using the Hadamard Product to produce the final output. This efficient design contributes to the lightweight nature of the ConvBlock.

Attack simulator: As outlined in the Introduction, ensuring strong robustness relies heavily on the attack simulator (AS), which serves as a crucial component of the overall architecture (illustrated in Fig. 2). Hence, the AS also plays an integral role

TABLE I
ROBUSTNESS RESULTS WITH/WITHOUT ATTACK SIMULATOR.

Method	GN 10dB	GN 20dB	LP 3k	BP 0.8-5k	TSI 2×
TAWM [22]					
w/o. AS	0.6012	0.7849	0.9176	0.9129	0.8348
w. AS	0.6335	0.8154	0.9934	0.9983	0.9448
True (Ours)					
w/o. AS	0.4994	0.5314	0.4960	0.6087	0.5006
w. AS	0.7922	0.9358	0.9096	0.9687	0.9678

TABLE II
COMPARISON OF FIDELITY WITH HANDCRAFT (HC) WATERMARKING METHODS AND DEEP-LEARNING-BASED (DL) METHODS.

	Method (bps)	Dataset	STOI↑	PESQ↑	SSIM↑	ACC↑
HC	FSVC (32) [16]	LJSpeech	0.9984	3.9977	0.9803	**1.0000**
	Normspace (32) [17]		0.9646	2.5506	0.8868	**1.0000**
	PBML (100) [19]		0.9861	3.7866	0.9560	**1.0000**
DL	AudioSeal (16) [21]	LJSpeech	0.9985	4.5893	0.9811	0.9214
	WavMark (32) [25]		**0.9997**	4.4628	0.9690	**1.0000**
	TAWM (100) [22]		0.9853	4.0353	0.9388	0.9998
	GROOT (100) [33]		0.9605	3.3871	0.9088	0.9969
	True (32) (Ours)	LJSpeech	0.9986	4.5748	**0.9833**	0.9986
	True (100) (Ours)	LJSpeech	0.9987	**4.6380**	0.9819	0.9973
	True (100) (Ours)	LibriTTS	0.9967	4.6290	0.9889	1.0000
	True (100) (Ours)	LibriSpeech	0.9985	4.6218	0.9753	0.9992
	True (100) (Ours)	M4Singer	0.9799	4.6193	0.9939	0.9992
	True (100) (Ours)	Opencpop	1.0000	4.6400	0.9971	0.9978

in the effectiveness of our proposed method by incorporating nine voice post-processing operations: Gaussian noise (GN), low-pass filtering (LP), band-pass filtering (BP), high-pass filtering (HP), time stretching and interpolation (TSI), suppression (SPS), resampling (ReS), echo, and dither. In practice, the AS is utilized exclusively during the training phase, with a single attack applied to each waveform.

Extracting & Verification: The procedure for extracting the watermark $\hat{\mathbf{w}}$ can be formalized as follows:

$$\hat{\mathbf{w}} = \mathbf{EB}(\mathbf{D}(\hat{\mathbf{s}})). \tag{4}$$

Inspired by [32], test hypothesis is employed for watermark verification. Assuming that the errors in the watermark bits are independent and taking into account the previously defined watermark bit length l, the number of matching watermark bits κ is calculated using the binomial distribution $Pr(X = \kappa) = \sum_{i=\kappa}^{l} \binom{l}{i} \xi^i (1-\xi)^{l-i}$, where $\xi = 0.5$ is the probability that needs to be tested under hypotheses.

C. Optimizing Strategy

The ultimate objective of jointly optimizing the HM, AS, and EM is to balance the trade-off between watermark extraction accuracy and the fidelity of the watermarked voice. To ensure high fidelity, the mel-spectrogram loss is initially employed to minimize the distance between the natural waveform $\mathbf{s}$ and the watermarked waveform $\hat{\mathbf{s}}$, which is formulated as:

$$\mathcal{L}_{MEL} = ||\psi(\mathbf{s}) - \psi(\hat{\mathbf{s}})||_1, \tag{5}$$

where $||\cdot||_1$ is L_1 norm and ψ represents the function of mel transformation. Then, the logarithmic STFT magnitude loss is utilized as an additional measure to further enhance fidelity:

$$\mathcal{L}_{MAG} = ||\log(\mathbf{STFT}(\mathbf{s})) - \log(\mathbf{STFT}(\hat{\mathbf{s}}))||_1, \tag{6}$$

where $\mathbf{STFT}(\cdot)$ denotes the transformation of STFT magnitude. The total loss for preserving the fidelity is defined as:

$$\mathcal{L}_{WAV} = \lambda_1 \mathcal{L}_{MEL} + \lambda_2 \mathcal{L}_{MAG}, \tag{7}$$

where λ_1 and λ_2 are hyper-parameters of the mel-spectrogram loss and the logarithmic STFT magnitude loss.

TABLE III
CAPACITY OF THE PROPOSED METHOD UNDER VARIOUS DATASETS.

Dataset		Capacity (*bps*)				
		32	100	300	500	600
LJSpeech	PESQ↑	4.5748	4.6380	4.6429	4.6000	4.6200
	ACC↑	0.9997	0.9973	0.9956	0.9210	0.8985
Opencpop	PESQ↑	4.5598	4.6400	4.6137	4.2753	4.6227
	ACC↑	0.9873	0.9978	0.9735	0.9716	0.8956

Binary cross-entropy is leveraged to ensure the accurate extraction of the watermark:

$$\mathcal{L}_{WM} = -\sum_{i=1}^{k} w_i \log \hat{w}_i + (1 - w_i) \log(1 - \hat{w}_i). \tag{8}$$

The overall training loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{WAV} + \alpha \mathcal{L}_{WM}, \tag{9}$$

where α serves a hyper-parameter to balance auditory quality and watermark recovery accuracy.

III. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents experiments conducted to evaluate the proposed True method based on three primary assessment criteria: fidelity, capacity, and robustness. Additionally, comprehensive experiments and analyses are provided to compare its performance against SOTA methods, further demonstrating its effectiveness.

A. Experimental Setup

Datasets and Baseline: In the context of voice datasets, other than conventional speech voice datasets like LJSpeech [34], LibriTTS [35], and Aishell3 [36], singing voice datasets like M4Singer [37] and Opencpop [38] are also considered. Moreover, comprehensive comparisons against SOTA watermarking methods are conducted, including handcrafted (HC) methods such as FSVC [16], Normspace [17], and PBML [19], as well as deep learning-based (DL) methods like AudioSeal [21], WavMark [25], TAWM [22], and Groot [33].

Evaluation Metrics: Two auditory objective metrics, Short-Time Objective Intelligibility (STOI) [39] and Perceptual Evaluation of Speech Quality (PESQ) [40], are used to evaluate the auditory performance of watermarked voice. Besides, the Structural Similarity Index Measure (SSIM) [41] is employed to assess the watermarked voice from the perspective of the visualized spectrogram.

Models & Training Settings: In the downsampling block of ICDE, Conv1d layers use a kernel size of 3, a stride of 2, and padding of 2, except for the last layer, which has padding of 1. In the upsampling block, Conv1d layers maintain a kernel size of 3, a stride of 1, and padding of 1, while transposed Conv1d layers have a kernel size of 3, a stride of 2, padding of 2, and output padding of 1, except for the first layer, which uses padding of 1. For the TGCs, both Conv1d layers in the two branches have a kernel size of 3, a stride of 2, and padding of 1. During training, AdamW optimizer [42] is employed with a learning rate of 2e-4. The training is conducted for 40 epochs with a batch size of 16. The hyper-parameters λ_1, λ_2, and α are set to 0.8, 0.1, and 0.3, respectively.

B. Indispensable AS for Frequency-domain Watermarking

The observation presented in the Introduction section highlights the significance of AS in frequency-domain watermarking techniques. To validate this, we select the TAWM

TABLE IV
ROBUSTNESS OF THE PROPOSED METHOD IN TERMS OF ACCURACY UNDER VARIOUS DATASETS.

	Dataset	GN			PN	LP	BP	HP	SPS	ReS	Echo	TSI	Dither
		10 dB	15 dB	20 dB	0.5	3k	0.5-8k	1k	behind	44.1k	default	0.5×	default
Singing Voice	M4Singer	0.9254	0.9783	0.9961	0.9645	0.9903	0.9992	0.9841	0.9992	0.9992	0.9896	0.9992	0.9992
	Opencpop	0.8260	0.9298	0.9782	0.9625	0.8305	0.9954	0.9872	0.9960	0.9961	0.9802	0.9940	0.9959
Speech Voice	LJSpeech	0.7922	0.8844	0.9358	0.9964	0.9096	0.9687	0.8417	0.9679	0.9699	0.9216	0.9678	0.9970
	LibriTTS	0.8285	0.9117	0.9557	1.0000	0.9471	0.9857	0.9084	0.9856	0.9857	0.9396	0.9852	1.0000
	Aishell3	0.9257	0.9763	0.9935	0.9992	0.8919	0.9991	0.9978	0.9991	0.9992	0.9876	0.9990	0.9992

TABLE V
COMPARISON OF ROBUSTNESS WITH SOTA METHODS IN ACCURACY.

Method (bps)	GN		PN	BP	SPS	Echo
	10 dB	20 dB	0.5	0.5-8k	behind	default
AudioSeal (16) [21]	0.6086	0.6600	0.6571	0.9764	0.8925	0.7277
Normspace (32) [17]	0.5856	0.6208	0.4733	0.4796	0.6596	0.5630
FSVC (32) [16]	0.7312	0.8835	0.8164	0.8263	0.7188	0.7976
WavMark (32) [25]	0.5295	0.6523	0.6924	0.9995	0.9713	0.8668
True (32) (Ours)	**0.9673**	**0.9986**	**0.9868**	**0.9998**	**0.9996**	**0.9963**
PBML (100) [19]	0.6060	0.7176	0.7060	0.7504	0.6365	0.6995
TAWM (100) [22]	0.6335	0.8154	0.7282	**0.9983**	0.9814	0.9471
Groot (100) [33]	**0.9929**	**0.9953**	0.9861	0.9947	0.9718	**0.9833**
True (100) (Ours)	0.8285	0.9557	**0.9964**	0.9857	**0.9856**	0.9396

[22], which serves as a representative example of frequency-domain watermarking methods for our analysis. Table I clearly demonstrates the difference in robustness of TAWM with and without the AS when subjected to various attacks. In the absence of AS, extraction accuracy can decrease by nearly 8% under certain attacks. These findings reaffirm that, while frequency-domain watermarking offers robust protection, the inclusion of AS is crucial for achieving a higher level of robustness.

C. Fidelity and Capacity

Fidelity refers to the imperceptibility of embedded watermarks, measured by the extent to which audio quality is preserved after watermarking. Table II reports both audio quality (for speech and singing voice) and the corresponding watermark extraction accuracy achieved by the proposed True. Additionally, it presents a fidelity comparison against SOTA methods on the LJSpeech dataset. All metrics are computed by comparing the watermarked waveform to its natural counterpart. The results indicate that True achieves excellent speech quality while maintaining reasonable watermark extraction accuracy on speech datasets. Similarly, it demonstrates strong fidelity and extraction performance on singing voice datasets. In comparative experiments with SOTA methods, True ranks first in both PESQ and SSIM metrics, highlighting its superior speech quality.

Capacity which reflects the length of watermarks that can be embedded, is as critical as fidelity in evaluating watermarking performance. Table III illustrates the results of the proposed method on the LJSpeech and Opencpop datasets under varying watermark capacities. Experimental results demonstrate that True is well-suited for high-capacity scenarios, supporting up to 500 bps. On the LJSpeech dataset, the method achieves an average watermark extraction accuracy of 97.84% while maintaining high fidelity, with an average PESQ score of 4.6140. Similarly, for the Opencpop dataset, it achieves an average PESQ of 4.5222 and a recovery accuracy of 98.26%. However, when the capacity exceeds 600 bps, both fidelity and extraction accuracy degrade significantly.

D. Robustness

We evaluated the robustness of the proposed method across various datasets at a watermark capacity of 100 bps. GN, Pink noise (PN), LP, BP, HP, SPS, ReS, Echo, TS, and Dither were employed for validation. Table IV illustrates the watermark extraction accuracy under these attacks for each dataset. The results show that the proposed method exhibits strong robustness, particularly on singing voice datasets. Specifically, it achieves extraction accuracies of 99.61% and 97.82% under a noise level of 20 dB. Under remanent attacks, True further attains average recovery accuracies of 99.45% and 96.85%, respectively. Although performance on speech datasets is slightly lower than that on singing voice, True still maintains desirable robustness, with average extraction accuracies of 92.33%, 94.85%, and 91.91% across the evaluated conditions.

The proposed True was further compared with SOTA methods on the LJSpeech dataset to evaluate its robustness against various voice post-processing operations, as illustrated in Table V. To account for the capacity limitation of the compared methods, two sets of experiments were conducted at capacities of 32 bps and 100 bps, respectively. At 32 bps, True achieved higher watermark extraction accuracy than all SOTA methods. Notably, it exhibited strong resilience, maintaining an average extraction accuracy of 99.14% even after undergoing six different types of signal attacks. The observed robustness gains can be attributed to ICDE's deep feature embedding and TGC's gated decoding, which jointly preserve and extract watermark signals from high-level temporal features. This design makes the method more resilient to typical signal-level distortions. Under the 100 bps configuration, True continued to demonstrate robust performance, particularly in resisting PN and SPS attacks. Although its accuracy under 10 dB GN was slightly lower than that of Groot, True still significantly outperformed the remaining two methods.

IV. CONCLUSION

In this study, we propose *True*, a temporal-aware robust watermarking method designed to proactively protect the copyrights of diverse waveform types, including both speech and singing voice. To enable seamless watermark embedding, we introduce a content-driven encoder that directly integrates the watermark into the temporal representation of the waveform and reconstructs the watermarked signal end-to-end. For extraction, we develop a temporal-aware gated convolutional network that effectively captures fine-grained features from attacked waveforms, thereby enhancing watermark recovery accuracy. The proposed True surpasses baseline methods in fidelity, supports high-capacity embedding of up to 500 bps, and exhibits strong robustness against a wide range of common waveform distortions.

REFERENCES

[1] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234–4245, 2024.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[3] Y. Zhang, R. Huang, R. Li, J. He, Y. Xia, F. Chen, X. Duan, B. Huai, and Z. Zhao, "Stylesinger: Style transfer for out-of-domain singing voice synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 597–19 605.

[4] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.

[5] T. Micro, "Unusual CEO fraud via Deepfake audio steals US$243,000 from UK company," 2019. [Online]. Available: https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/unusual-ceo-fraud-via-deepfake-audio-steals-us-243000-from-uk-company

[6] W. Economic Forum, "The US is drafting new laws to protect against AI-generated deepfakes," 2024. [Online]. Available: https://www.weforum.org/agenda/2024/02/ai-deepfakes-legislation-trust/

[7] "Chinese AI governance rules," 2023. [Online]. Available: https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[8] "The EU Artificial Intelligence Act," 2023. [Online]. Available: https://artificialintelligenceact.eu/

[9] "Ensuring Safe, Secure, and Trustworthy AI," 2023. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[10] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 448–14 457.

[11] N. Yu, V. Skripniuk, D. Chen, L. S. Davis, and M. Fritz, "Responsible disclosure of generative models using scalable fingerprinting," in *International Conference on Learning Representations*, 2022.

[12] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 466–22 477.

[13] V. Asnani, J. Collomosse, T. Bui, X. Liu, and S. Agarwal, "Promark: Proactive diffusion watermarking for causal attribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 802–10 811.

[14] G. Zhang, L. Zheng, Z. Su, Y. Zeng, and G. Wang, "M-sequences and sliding window based audio watermarking robust against large-scale cropping attacks," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1182–1195, 2023.

[15] J. Zhao, T. Zong, Y. Xiang, L. Gao, G. Hua, K. Sood, and Y. Zhang, "Ssvs-ssvd based desynchronization attacks resilient watermarking method for stereo signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 448–461, 2022.

[16] J. Zhao, T. Zong, Y. Xiang, L. Gao, W. Zhou, and G. Beliakov, "Desynchronization attacks resilient watermarking method based on frequency singular value coefficient modification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2282–2295, 2021.

[17] S. Saadi, A. Merrad, and A. Benziane, "Novel secured scheme for blind audio/speech norm-space watermarking by arnold algorithm," *Signal Processing*, vol. 154, pp. 74–86, 2019.

[18] Z. Liu, Y. Huang, and J. Huang, "Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1171–1180, 2018.

[19] I. Natgunanathan, Y. Xiang, G. Hua, G. Beliakov, and J. Yearwood, "Patchwork-based multilayer audio watermarking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2176–2187, 2017.

[20] DeepMind, "Identifying ai-generated content with synthid," 2024, https://deepmind.google/technologies/synthid/.

[21] R. S. Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, "Proactive detection of voice cloning with localized watermarking," in *Proceddings of the 41st International Conference on Machine Learning*, 2024, pp. 43 180–43 196.

[22] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, "Detecting voice cloning attacks via timbre watermarking," in *Proceedings of the 31th Network and Distributed System Security (NDSS) Symposium 2024*. The Internet Society, 2024.

[23] C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang, and N. Yu, "Dear: A deep-learning-based audio re-recording resilient watermarking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 201–13 209.

[24] X. Qu, X. Yin, P. Wei, L. Lu, and Z. Ma, "Audioqr: deep neural audio watermarks for qr code," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 6192–6200.

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

[26] 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, p. 103381, 2022.

[27] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Maskmark: Robust neural-watermarking for real and synthetic speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4650–4654.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[29] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[30] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[31] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?" in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4484–4496.

[32] D. Lin, B. Tondi, B. Li, and M. Barni, "A cyclegan watermarking method for ownership verification," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 2, pp. 1040–1054, 2025.

[33] W. Liu, Y. Li, D. Lin, H. Tian, and H. Li, "Groot: Generating robust watermark for diffusion-model-based audio synthesis," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3294–3302.

[34] K. Ito, "The lj speech dataset," 2017, https://keithito.com/LJ-Speech-Dataset/.

[35] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech*, 2019, pp. 1526–1530.

[36] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," in *Interspeech*, 2021, pp. 2756–2760.

[37] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen *et al.*, "M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6914–6926, 2022.

[38] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, "Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis," in *Interspeech*, 2022, pp. 4242–4246.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[40] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.