Audio Model Watermarking: Methods, Metrics, and Unresolved Challenges

LIAORAN XU, East China Normal University, China
ZHAOXIA YIN, East China Normal University, China
WENWU WANG, University of Surrey, United Kingdom
XINPENG ZHANG, Fudan University, China

The increasing deployment of audio models has highlighted critical security concerns, particularly those related to model originality and intellectual property protection. The risks of unauthorized copying and misuse have made watermarking an important and rapidly growing area of research. This study presents a comprehensive review of recent advances in audio model watermarking, systematically categorizing existing approaches, evaluating their strengths and weaknesses, and identifying key gaps in current research. In addition, the paper reviews watermarking techniques specific to audio models and outlines a set of metrics for evaluating their performance. The study concludes by discussing promising directions for future research in this area.

CCS Concepts: • **Security and privacy → Digital rights management**.

Additional Key Words and Phrases: Audio Model Watermarking, Survey, Generative Watermarking, Model Security

ACM Reference Format:
Liaoran Xu, Zhaoxia Yin, Wenwu Wang, and Xinpeng Zhang. 2025. Audio Model Watermarking: Methods, Metrics, and Unresolved Challenges. 1, 1 (December 2025), 33 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As the application of audio models expands into various tasks, they not only enhance the efficiency of human-computer interaction, but also provide users with tailored experiences [88]. However, this widespread adoption has raised significant issues related to intellectual property protection and security [84, 90]. Unauthorized replication and misuse lead to substantial economic losses for model developers and companies [14, 30, 74]. Consequently, watermarking technology has become increasingly important as an effective measure for model protection [103]. Watermarking for audio models is a technical measure that ensures security by embedding identification information in the model [4]. As shown in Figure 1, research on watermarking technologies for audio models has demonstrated steady growth over the past five years, both volume and publication quality showing consistent improvement. This trend reflects the field's growing impact and promising research potential (literature cutoff: 24:00 CST, 10 December, 2025). Information related to the review of the literature, such as the main libraries searched for primary studies, inclusion and exclusion criteria, and keywords and synonyms for the search is presented in Tables 1, 2 and 3.

Digital watermarking [13] includes multimedia watermarking [23] for digital media such as image [38, 68], video [89], text [3, 48], and audio [28], as well as watermarking for AI models [20, 58, 101]. Among these, multimedia watermarking

Authors' addresses: Liaoran Xu, East China Normal University, Shanghai, China, 51285904077@stu.ecnu.edu.cn; Zhaoxia Yin, East China Normal University, Shanghai, China, zxyin@cee.ecnu.edu.cn; Wenwu Wang, University of Surrey, Guildford, United Kingdom, w.wang@surrey.ac.uk; Xinpeng Zhang, Fudan University, Shanghai, China, zhangxinpeng@fudan.edu.cn.

Manuscript submitted to ACM

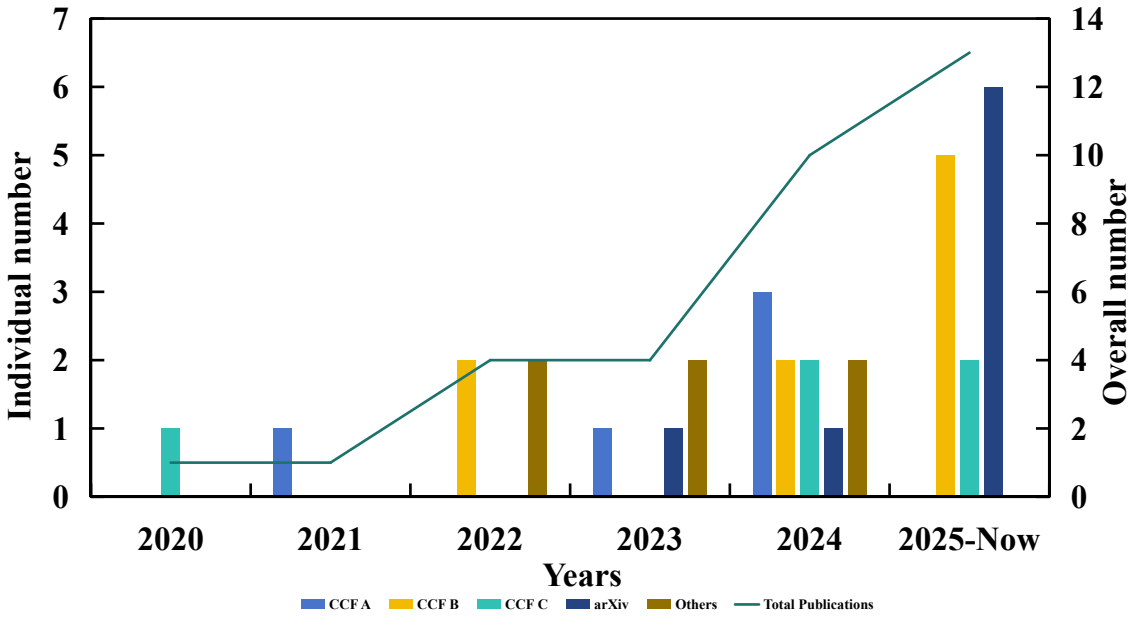

Fig. 1. Publication of audio model watermarking.

focuses on digital media such as image, video, text, and audio as the carrying content, embedding watermark information using media signals, including information hiding [32, 96]. Watermarking for AI models, on the other hand, targets AI models by embedding and verifying watermark information through methods, such as model parameter modification or trigger sets [107]. In recent years, some studies have proposed post-hoc watermarking methods [7, 34, 49, 60, 61, 75, 78], however, they largely focus on the modification of digital media content (e.g., images, text, and audio) generated by AI models, rather than the models themselves. Technically, these approaches still follow the ideas of conventional approaches designed for multimedia content watermarking, rather than the emerging methods for model watermarking. Since post-hoc audio watermarking is essentially an adaptation of the techniques for audio content watermarking, it is excluded from the analysis of watermarking techniques for audio models reviewed in this paper.

Depending on the type of models, the techniques designed to watermark the models can be divided into those for classification models [26] and those for generation models [70]. In terms of access rights to the model during watermark embedding and extraction, model watermarking can be classified into white-box watermarking [8], black-box watermarking [10], gray-box watermarking [76] and no-box watermarking [104]. Depending on the level of fragility, model watermarking can also be classified into robust watermarking [98], fragile watermarking [77], and semi-fragile watermarking [1].

In the field of audio model watermarking, Chen et al. [8] presented the first white-box watermarking method for audio models. Jia et al. [29] presented the first watermarking method for multi-modal models that include audio models. Cho et al. [12] optimized the watermarking method originally designed for an image generation model and applied it to the audio generation model. Rathi et al. [69] proposed the application of adversarial samples as a trigger set. Lv et al. [51] proposed a multi-modal watermarking method against model extraction attacks. Wu et al. proposed a

watermarking scheme for autoregressive speech generation models [99]. These innovative methods lay the groundwork for future research in this field and underscore the growing interest in safeguarding audio models from misuse and unauthorized copying.

This paper reviews key advances in audio model watermarking during 2020-2025 (literature cutoff: 24:00 CST, 10 December, 2025). We classify existing methods, analyze their strengths/weaknesses, and identify research gaps. It includes a functional analysis of audio model watermarking techniques with evaluation metrics, and concludes with future research directions. The contributions of this paper are as follows.

- We are the first to conduct a thorough analysis of audio model watermarking methods, identifying their benefits and limitations, and offering a comprehensive survey.
- We provide a set of metrics for performance evaluation, compare the performance advantages and disadvantages of different schemes, and analyze technical gaps in this field.
- We propose future research directions for audio model watermarking by analyzing the shortcomings and important missing pieces of existing work.

The subsequent sections of the paper are arranged as follows. Section 2 defines audio model watermarking and analyzes the attacker's capabilities. Section 3 summarizes the existing methods. Section 4 presents evaluation metrics for watermarking functions and robustness evaluation against watermark attacks. Section 5 explores future directions. Section 6 concludes the work.

Primary repositories	
Google Scholar	https://scholar.google.com
IEEExplore	http://ieeexplore.ieee.org
ACM Digital Library	http://dl.acm.org
Web of Science	https://webofscience.clarivate.cn/
Springer Link	https://link.springer.com
arXiv	https://arxiv.org/
Secondary repositories	
Semantic Scholar	https://www.semanticscholar.org
ScienceDirect	https://www.sciencedirect.com
Scopus	http://www.scopus.com

Table 1. Main libraries searched for primary studies.

Keyword	**Synonyms**
audio	speech, voice, sound, music
model	models
watermark	watermarked, watermarking, watermarks

Table 2. Set of keywords and synonyms for search.

Inclusion criteria
Papers on methods, metrics and challenges for audio model watermarking
Papers published in general conferences/journals
Papers published from 2020 to 2025 (literature cutoff: 24:00 CST, 10 December, 2025)
Papers on generative watermarking for audio model watermarking
Exclusion criteria
Papers not about audio model watermarking
Papers about audio watermarking
Papers on post-hoc watermarking for audio model watermarking
Publications not in English

Table 3. Inclusion and exclusion criteria.

2 DEFINITION OF AUDIO MODEL WATERMARKING AND ATTACKING

A typical watermarking method for audio models is illustrated in Figure 2 and Table 4. The watermark information can be embedded in the output m of the intermediate layer by changing the parameters of the intermediate layer L or the structure S. For the audio classification model, the goal is to classify the input audio x, with the output y being the classification result. The watermarking techniques focus more on adding watermarks in the intermediate layer L and its output m, since these layers determine the final classification result of the model [46]. For the audio generation model, the goal is to generate new audio y. A common method for embedding watermarks in this type of model is to add a watermark to the intermediate layer L, which is called generative watermarking [70]. This layer determines the characteristics of the final generated audio [39], allowing the model owner to embed copyright information directly into the generation process itself.

Symbol	Description
x	Initial input to the model
S	Structure of the model
L	Intermediate layer structure for the audio model
m	Intermediate outputs generated by each layer
y	Final output
p	Model parameter
w	Original watermark
w_e	Extracted watermarks
F_w	Watermark embedding functions
E_w	Watermark extraction function

Table 4. Parameter description of Figure 2.

Although various watermarking mechanisms offer protection for different models, these techniques also face threats from various attack scenarios. For the attack of audio watermarking models, the attacker's capabilities can be divided into black-box and white-box.

- Black-Box [21]: In a black-box attack, the attacker can only access the target model through the input and output interfaces. The attacker can construct specific input samples and observe the output results to detect the existence of the watermark.

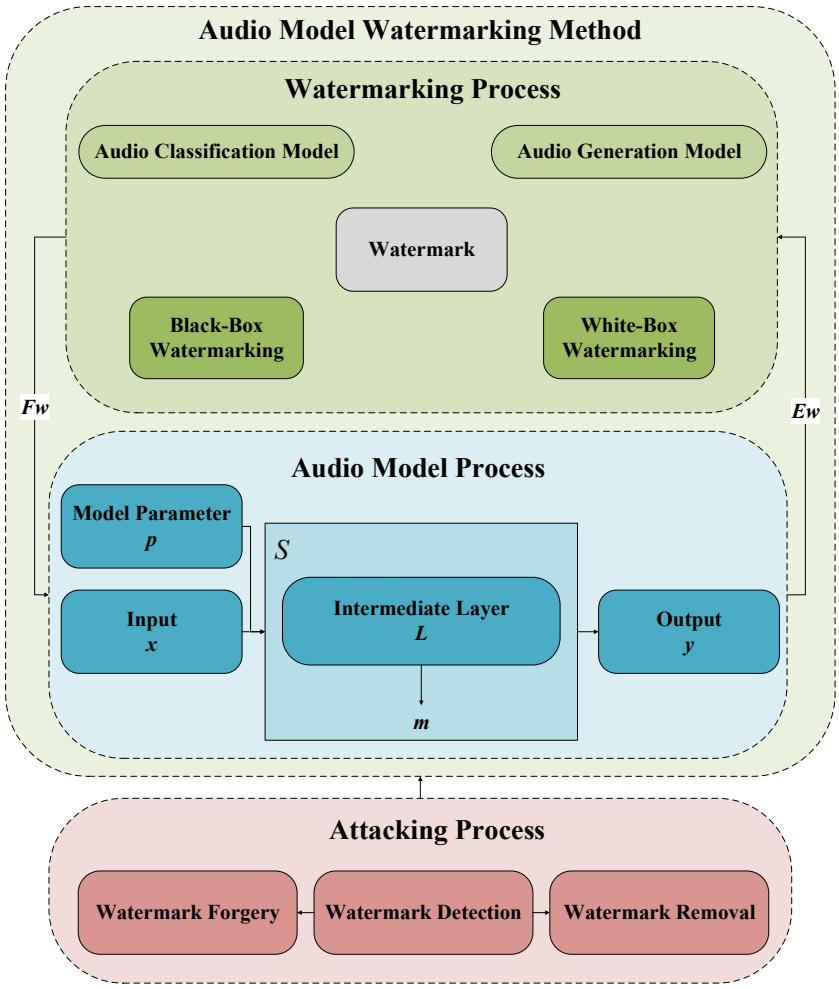

Fig. 2. General process of a typical audio model watermarking method: F_w is the watermark embedding process, E_w is the watermark extraction process.

- White-Box [59]: In a white-box attack, the attacker has full access to the internal structure and parameter information of the target model. The attacker can analyze the watermark embedding method within the model and then modify or remove the watermark directly.

3 WATERMARKING METHODS FOR AUDIO MODELS

To systematically review audio model watermarking advances (2020-2025), we first categorize methods by model type and task (see Figure 3 and Table 5). These approaches fall into two categories: parameter-modification based (Section 3.1) and trigger-set based (Section 3.2). The datasets commonly used in the scheme are summarized in Table 6. Some datasets lack specific species counts so they are denoted by '-' here.

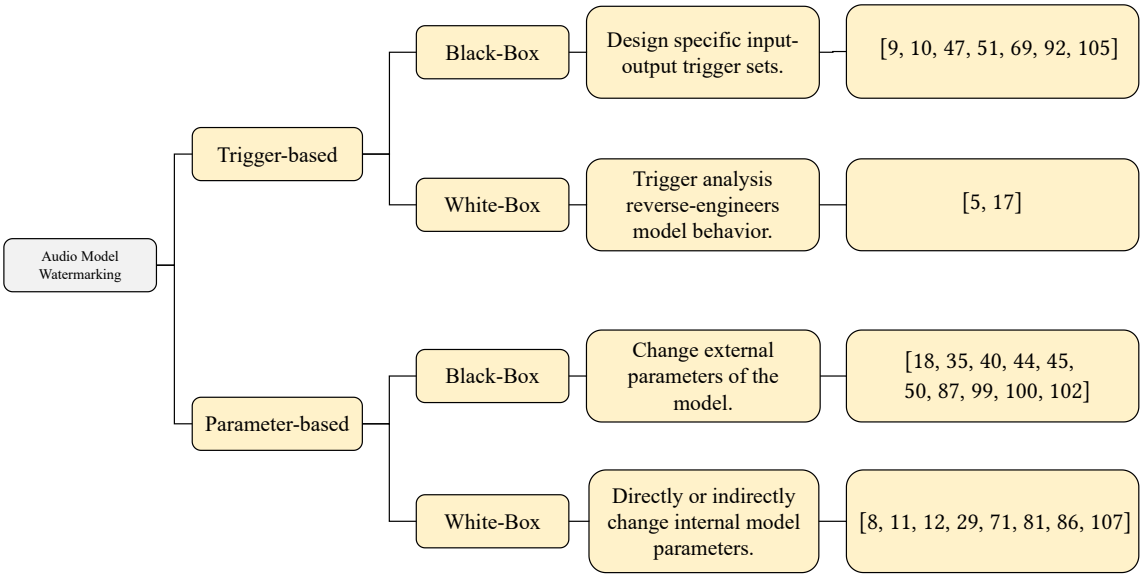

Fig. 3. Classification and implementation of audio model watermarks.

Audio Model Tasks		Number of Works	Watermarking Methods
Audio Classification Model	Speech Recognition	7	Based on model parameter modification. Based on trigger sets.
	Speaker Recognition Identification	3	Based on trigger sets.
	Speech Emotion Recognition	0	/
	Audio Event Detection	0	/
	Music Information Retrieval	0	/
Audio Generation Model	Speech Translation	0	/
	Text-to-Speech	17	Based on model parameter modification. Based on trigger sets.

Table 5. Numbers of watermarking methods for audio models.

3.1 Audio Model Watermarking Based on Model Parameter Modification

This type of technology embeds watermarks by directly or indirectly modifying the model parameters, usually under white-box conditions [8]. As for black-box conditions, the model parameters here refer to the external variables inputted into the model [50]. However, an attacker can reverse the watermark embedding logic by analyzing the model parameters to remove or forge the watermark. Improvements can be made by introducing encryption or obfuscation techniques to prevent attackers from cracking the watermark by reverse engineering [10].

Dataset	Number of classes	Number of samples	References
Speech Command	35	96783	[5, 9, 17, 29, 51]
Command Voice (en)	-	1805365	[8, 69]
TIMIT	630	6300	[9, 47, 92, 105]
LJSpeech	7	13100	[11, 12, 18, 44, 45, 50]
LibriSpeech	-	5559	[8, 10, 18, 45, 50, 86, 100, 105]
LibriTTS	-	9957	[18, 44, 45, 50, 71, 78, 107]
AN4	-	1078	[8]
Slakh2100	34	2100	[86]
MUSDB18-HQ	-	150	[87]
Football Keywords	18	30259	[17]
SC09	10	39405	[12]
VCTK	110	44000	[81]
MTG-Jamendo	195	55000	[81]
GTZAN	10	1000	[102]
MusicCaps	-	5521	[40]

Table 6. Statistics of audio datasets used in audio model watermarking methods.

3.1.1 **Parameter Modification under White-Box.** Under white-box conditions, where full access to the model is available, watermark authentication is typically achieved by directly extracting or comparing explicit watermark information embedded within the model's internal weights [59].

In speech recognition, Chen et al. [8] proposed a white-box parameter watermarking scheme. This scheme employed Discrete Cosine Transform (DCT) and Spread Spectrum technology to embed watermarks into the prominent spectral components of model weights. It achieved efficient embedding without requiring model retraining and demonstrates strong robustness against model pruning, fine-tuning, and transfer learning. Its primary limitation is its reliance on white-box scenarios, where verifiers must have direct access to the model's internal weight parameters to extract the watermark. This approach cannot be validated solely through the API input-output behavior, unlike black-box trigger set solutions. Jia et al. [29] embedded the watermark during the model training process, where the model is guided to exhibit specific behaviors for specific inputs. This scheme constituted a parameter-modifying watermarking technique, fundamentally altered how model parameters represent data features. This caused the features of watermarked data and normal task data to become entangled within the parameter space. The scheme effectively resisted model extraction attacks. However, this method incurred significant computational overhead, requiring approximately 1.5 to 2 times training data and computational resources. Furthermore, maintaining the balance between watermark robustness and model accuracy becomes increasingly challenging when scaling to deeper network architectures or more complex tasks. Cheng et al. [11] proposed a proactive strategy that combines an audio watermarking technique with HiFi-GAN vocoder. It incorporated a pre-trained watermark extraction network as part of the loss function, utilizing both watermark extraction loss and speech quality loss to guide the adjustment of generator weights. This approach demonstrated versatility and compatibility with other vocoder-based speech synthesis models. It maintained high imperceptibility while effectively resisting multiple signal post-processing attacks. However, the scheme required an additional watermark extraction network as a component of the loss function, increasing the complexity of the system. Ren et al. [71] proposed a plug-in watermark fusion method, which is based on parameter-level watermarking. It embedded watermarks into published model weights with minimal computational overhead and without compromising

original model performance, thereby addressing the challenge of model intellectual property protection in open-source scenarios. Furthermore, by embedding watermarks within the parameter space, it exhibited high robustness against attacks such as model pruning, fine-tuning, and transfer learning. However, as a white-box solution, these verifiers must access and inspect the model's internal parameters to extract the watermark, precluding ownership verification through simple black-box API queries. In addition, its applicability to different types of generative model may require tailored adjustments. Tang [86] proposed a watermarking algorithm for a diffusion model, which is based on noise distribution perturbation and authorization mechanism. The generated audio carried a traceable watermark, and the model produced low-quality output when unauthorized, thereby safeguarding both the model weights and generated content. This approach is a white-box solution, relying on intervention during the model training process, making it difficult to apply to deployed black-box generative models.

The aforementioned approaches involved modification of static parameters. There are also methods for dynamically modifying the parameters. For example, Singh et al. [81] proposed blocking unauthorized deep forgeries and enabling user tracking in the generation model by combining key-based model authentication with watermarking techniques. This approach combined key authentication with watermark tracking to protect the intellectual property of generative models and trace deepfake content, constituting a watermarking technique based on model parameter modification. It achieved this by providing a model parameter uniquely bound to the user's key, which was then applied as an additional input condition during model inference to alter the output. This mechanism altered the behavior of the model parameters under specific keys, enabling control over both the model and its generated content. Even if attackers possess the complete model parameters, they cannot produce usable content without a valid key. Providing a robust deterrent against unauthorized model replication and distribution. However, the key's role as an input during inference may introduce an additional attack surface prone to reverse engineering or side-channel attacks.

Regarding the attribution of generative models, Cho et al. [12] improved the traceability of the watermark to the source model by embedding watermarks indirectly in the synthesized content, and this helps mitigate the problem of malicious impersonation and theft of intellectual property. This method achieved a specific user-end model by fine-tuning the default generator to embed a unique user key into its parameters, thus endowment of the generated speech with watermark characteristics recognizable by classifiers. The core strength and innovation of this paper is the proposal of an angle loss that aligns the generated model's output distribution with the user key's direction. This resolved the issue of low attribution accuracy when directly applying image domain methods and achieved robust training against post-processing attacks. Its primary drawback is the trade-off between attributable quality and generated speech fidelity. Enhancing watermarking and robustness inevitably degrades speech quality, with increased robustness in training further reducing output fidelity. Zhou et al. [107] proposed a text-to-speech generation model for the direct generation of watermarked audio to form a closed-loop information embedding and extraction process. During training, it modified the model's parameters by jointly optimizing a speech quality-based loss and a watermark-based loss, enabling the output speech to carry watermark features from the outset. The core innovation is achieving frame-level watermark embedding and extraction, alongside model-embedded watermark generation. This made the watermark with increased inaudibility and robustness, particularly demonstrating strong resistance to resplicing attacks. This method addresses the issues of degraded quality and limited flexibility inherent in traditional post-processing watermarking schemes. However, implementing this framework necessitates extensive modifications and joint optimization of the entire text-to-speech model architecture, resulting in limited flexibility in applying it to deployed, non-retrainable black-box text-to-speech systems. Furthermore, watermark extraction requires a dedicated and robust extractor, increasing system complexity.

3.1.2 **Parameter Modification under Black-Box.** Black-box methods aim to address the practical constraint of inaccessibility to internal model weights. By analyzing specific inference behaviors during interface invocations, they indirectly detect watermark features concealed within parameter fine-tuning [21].

When the specific parameters of the model are unknown, watermark embedding can be achieved by equipping the model with a dedicated encoder. Liu et al. [50] indirectly modified the input parameters of the diffusion model by adding latent variables generated by the watermark encoder to the latent space. This is achieved by integrating a dedicated watermark encoder into the diffusion model and fixing the diffusion model parameters during training, focusing solely on training the encoder. This encoder mapped watermark information to the noise term within the diffusion process, thereby modifying the internal workings of the generative model to embed the watermark within the output audio. This approach fell into the category of modifying the parameters of the generative model. This design made the watermark robust against various post-processing attacks, particularly denoising attacks. However, the approach relied on modifying the diffusion model architecture and employing specially trained encoder/decoder components, precluding straightforward application to deployed black-box audio generation services. They also proposed to use low-rank adaptation to fine-tune speech diffusion models and embed watermarks indirectly [44], where the watermark is converted into latent variables via an encoder, which are then combined with the model input, to ultimately achieve efficient parameter updates through low-rank adaptation (LoRA) based fine-tuning [25].

By combining existing classical audio watermarking techniques, Tralie et al. [87] proposed an embedding method based on the classical echo watermarking technique to implicitly label training data in the generative model. This method implanted specialized echoes within the training data. As the model learned the data distribution, it encoded these echo patterns into its weight parameters. The echoes resisted attacks such as full model training, fine-tuning, audio mixing/separation, and pitch shifting. However, the information capacity of this watermark is limited, rendering it unsuitable for precise tracking applications. Furthermore, its effectiveness is highly dependent on the model's sensitivity to specific acoustic features. Juvela et al. [35] also proposed an audio watermarking method based on multi-scale feature enhancement to improve the robustness of the watermark by introducing coding perturbations. This approach modified model parameters by incorporating audio codec enhancement during the training phase for the speech synthesis models, integrating watermark extraction loss into the loss function for model training. The core approach is utilizing differentiable neural network codecs or waveform domain proxy losses to simulate non-differentiable traditional black-box codecs, integrating this technique into the training of collaborative watermarks. This overcame the drawback of conventional watermarks being easily removed by decoding operations. However, incorporating codec enhancement increases both model training time and computational resource requirements.

Considering the specific watermark embedding process, different embedding formats can be considered. Feng et al. [18] employed a trainable masking mechanism and a kernel weight normalization technique to embed watermarks directly within the convolutional layers of the generative model. This method involved embedding watermarks within the weights of the convolutional layer, constituting a technique that directly modifies the model parameters to influence the output features. The scheme embedded the watermark in the convolutional kernel weights of the model, and employed a normalization step to ensure that the statistical properties of the modified weights align with those of the original model. This preserved the resilience of the watermark while having a negligible effect on the quality of the generated audio. However, it necessitated verifiers having access to the model's internal parameters to extract the watermark. Furthermore, the approach required a dedicated encoder-decoder architecture to support both training and extraction processes, thereby increasing the overall system complexity. Lee et al. [40] embedded a barcode-shaped 4-QAM modulated watermark in the initial noise of the audio diffusion model, achieving strong robustness against

attacks such as cut-and-paste manipulation alongside high detection accuracy without requiring retraining. However, the modulation scheme limits the watermark capacity, and the robustness against some composite attacks remains insufficient.

In the context of autoregressive speech generation, Wu et al. [99] identify new challenges associated with watermarking autoregressive models and motivate the development of specialized solutions. They propose a watermarking scheme based on statistical re-weighting to embed detectable watermarks while maintaining high audio quality; however, the approach suffers from reduced detection accuracy due to re-encoding mismatch. To address this limitation, Wu et al. [100] introduce a distortion-free watermarking method that leverages clustering and alignment-based inverse sampling, effectively alleviating the re-encoding mismatch problem and substantially improving detection performance without compromising audio quality. Nevertheless, the method remains vulnerable to time-shifting attacks.

While Wu et al.'s approach focuses on adapting to retokenization mismatches through clustering, Jovanović et al. [33] adopt a more proactive robustness-oriented strategy. Although primarily developed for image models, their work includes exploratory studies on autoregressive audio models. By addressing the reverse cyclical consistency issue through targeted fine-tuning and introducing a watermark synchronization layer capable of detecting and compensating for geometric attacks, the method significantly enhances watermark robustness under complex manipulations, offering stronger protection against tampering.

As for the attribution of generative models, Yang et al. [102] used a dual-watermark embedding mechanism coupled with a consistency loss function to simultaneously embed model and data source watermarks within audio generation models. This achieved high-precision dual attribution with robust performance, though it suffered from limited watermark capacity, unoptimized codec effects on watermarks, and invalidated generalization capabilities. Li et al. [45] proposed a robust method for watermarking a speech generation model based on efficient fine-tuning of the diffusion model, by indirectly modifying the model parameters via pre-embedding watermarks into training data. However, its robustness may deteriorate under conditions such as strong noise attacks, compound desynchronization processing, and adversarial training attacks against the model.

3.2 Audio Model Watermarking Based on Trigger Sets

Watermarking based on trigger sets operates by injecting a specific behavior into a model during training, where the model learns to produce a designated output when presented with a secret trigger input, while maintaining normal performance on standard data. This technology requires the construction of specific input-output pairs, and it allows copyright holders to prove ownership simply by querying the model API without needing access to internal parameters [41]. However, this may lead to an insufficient stealthiness of the watermark, which can be easily recognized and removed by attackers. Improvements can be made by introducing more sophisticated algorithms for generating the trigger set [51]. A variety of examples can be found in the literature, as discussed below.

3.2.1 **Trigger Sets under White-Box.** Under white-box conditions, the deep mapping relationship between trigger patterns and predefined labels is typically verified by examining anomalous activation states within the model's neural layers or by leveraging gradient features [59].

For watermark embedding based on trigger sets under white-box conditions, solutions typically balance concealment and robustness by optimizing noise design and adaptive embedding strategies. Cao et al. [5] proposed a combination of standard Gaussian noise and watermarked noise to achieve embedding. This method can resist potential interference and attempts to remove the watermark while maintaining a high recognition success rate. However, the effectiveness

Method	Dataset	Fidelity ($\lvert\Delta\rvert \downarrow$)		
		WER	CER	Model ACC
Interspeech 2020 [8]	AN4	0.000	0.000	-
	Command Voice	0.000	0.010	-
	LibriSpeech	0.810	0.000	-
USENIX Security 2021 [29]	Speech Command	-	-	0.610
APPL ARTIF INTELL 2022 [69]	Command Voice	-	-	-
Symmetry 2022 [92]	TIMIT	-	-	0.072
ICASSP 2022 [10]	LibriSpeech	0.347	0.914	-
Electronics 2023 [105]	LibriSpeech	-	-	0.001
	TIMIT	-	-	0.001
TAI 2024 [9]	Speech Command	0.263	0.356	-
	TIMIT	0.005	0.040	-
S&P 2024 [51]	Speech Command	-	-	-
Vis. Intell. 2024 [47]	TIMIT	-	-	0.012

Table 7. Fidelity metrics for audio classification model watermarking schemes.

of watermarks may be affected by different audio qualities and complexities. Fei et al. [17] proposed a dynamic audio watermarking framework that attaches robust and adaptive triggers at arbitrary locations of the audio signal, and integrates boundary sample selection driven by oblivious events. This scheme employed an adaptive embedding strategy to embed watermarks in regions of low sensitivity within the loss function, thereby maximizing the preservation of the model's original performance. Boundary sampling positions trigger set samples near the model's decision boundary, enhancing watermark robustness. However, this approach may struggle against state-of-the-art adaptive watermark removal attacks.

3.2.2 **Trigger Sets under Black-Box.** Under black-box conditions, trigger sets rely solely on the model's input-output interfaces. Confirming the presence of watermarks by observing the model's predictive responses to specific trigger patterns, which offers considerable flexibility [21].

For speech recognition models in black-box settings, Chen et al. [10] introduced a watermarking framework that generates trigger audio by extending the model owner's speech segments to span the entire input. Nevertheless, such watermarks can be neutralized through model extraction attacks. Rathi et al. [69] proposed a watermarking approach based on deep recurrent neural networks that employs adversarial examples of generated target audio as trigger sets, exploiting their imperceptibility to improve watermark concealment. Despite this, the watermark remains vulnerable to removal via limited fine-tuning or the creation of a clean surrogate model using knowledge distillation. Similarly, Chen et al. [9] utilized hidden adversarial audio samples as trigger sets, designing user-specific adversarial triggers for model authentication. While effective for ownership verification, this method is unable to monitor or track the normal outputs of models that have been extracted or compromised.

For speaker recognition identification, Wang et al. [92] designed a black-box watermarking method by using triggered audio samples in the frequency domain, but the watermarks cannot resist sophisticated attacks such as frequency-domain filtering, adversarial audio perturbations and model extraction attack. To improve the robustness, Zhang et al. [105] also proposed a black-box speech recognition model protection framework that combines the active and passive protection measures. However, the robustness of this watermarking method can still be improved for more sophisticated attacks such as adversarial audio perturbations and model extraction attack. To test different kinds of Gaussian noise on the trigger, Liao et al. [47] designed a black-box watermarking method using audio with added Gaussian noise as trigger sets, but the performance of the model is greatly affected to some extent.

For a watermarking method that accommodates triggering across different modalities, Lv et al. [51] proposed a method by injecting a robust watermark against the model extraction attack. Watermark samples are generated by combining two samples from different source categories and assigning them a new label. This construction method made it difficult for extraction models to mimic the decision boundary of the original watermark model, thereby significantly enhancing the watermark's robustness against model extraction attacks. However, the paper notes that there are no theoretical guarantees against removal by future adaptive attacks or watermark stripping techniques.

4 METRICS FOR AUDIO MODEL WATERMARKING EVALUATION

The performance of model watermarking technology can be measured in various ways depending on different watermarking purposes, such as copyright protection, attribution, and tamper detection. The paper summarizes the reliability and fidelity data for each scheme in Tables 7, 8, 9, 10, and provides an assessment of the remaining metrics in Table 11. Based on existing model watermarking technology, we summarize the functions of existing audio model watermarking technology and provide corresponding evaluation metrics in Tables 12 and 13. The most important metrics are reliability, fidelity, and robustness. Other common metrics include capacity and efficiency. Next, we first introduce the evaluation metrics for various watermarking functions in audio models (Section 4.1), followed by an in-depth analysis of robustness evaluation against watermark attacks (Section 4.2).

Overall, although most existing audio model watermarking methods have achieved high reliability under standard conditions, the shift towards complex architectures for large-scale generation still poses reliability challenges. As demonstrated in Table 9 and Table 10, existing audio model watermarking schemes exhibit outstanding reliability. In copyright protection scenarios, as shown in Table 9, the vast majority of methods [8, 69] achieve extraction accuracy rates approaching 100% on datasets such as LibriSpeech and Speech Command. This indicates that current approaches—based on parameter modification or trigger sets—can provide robust proof of ownership in the absence of attacks. It is noteworthy that certain audio generation approaches fine-tuned for large language models achieve only 57% accuracy on specific tasks [99]. This highlights that embedding watermarks within generative large models remains a challenge, particularly in maintaining watermark stability across vast parameter spaces.

Performance evaluation tables indicate that most current audio model watermarking schemes maintain high fidelity, though a trade-off between robustness and fidelity remains necessary in generative tasks. Table 7 demonstrates the fidelity of the classification model, revealing that watermark embedding exerts a negligible impact on original task performance. For instance, the scheme [105] in Table 7 exhibits a mere 0.001% decline in model classification accuracy after embedding the watermark, attesting to the watermark's exceptional concealment. For the generative models in Table 8, fidelity assessment is more complex, typically employing metrics such as the mean opinion score (MOS), short time objective intelligibility (STOI), and perceptual evaluation of speech quality (PESQ). The data indicates that for most schemes [11, 50], the generated speech quality metrics show negligible differences compared to the original model after

Method	Dataset	Fidelity($\|\Delta\| \downarrow$)			
		MOS	STOI	PESQ	FAD
ICASSP 2022 [12]	LJSpeech	-	-	-	5.520
	SC09	-	-	-	0.360
TPS-ISA 2023 [5]	Speech Command	-	-	-	-
ACM MM 2024 [50]	LJSpeech	0.063	0.005	-	-
	LibriSpeech	0.001	0.010	-	-
	LibriTTSh	0.089	0.016	-	-
SPL 2024 [11]	LJSpeech	0.180	-	0.050	0.000
Interspeech 2024 [107]	LibriTTS	0.405	0.003	0.372	-
arXiv preprint 2024 [81]	VCTK	0.040	-	-	-
	MTG-Jamendo	0.140	-	-	-
arXiv preprint 2024 [87]	MUSDB18-HQ	-	-	-	-
ICASSP 2025 [86]	LibriSpeech	-	-	-	0.230
	Slakh2100	-	-	-	0.250
ICASSP 2025 [17]	Speech Command	-	-	-	0.500
	Football Keywords	-	-	-	0.100
ICASSP 2025 [35]	LibriTTS	0.400	-	-	-
PRL 2025 [40]	MusicCaps	-	-	-	0.610
SPL 2025 [18]	LJSpeech	0.003	0.003	-	-
	LibriSpeech	0.000	0.024	-	-
	LibriTTS	0.010	0.020	-	-
Interspeech 2025 [99]	C4 Subset	-	-	-	-
	Dolly-CW	-	-	-	-
	MMW Story	-	-	-	-
arXiv preprint 2025 [71]	LibriTTS	-	0.020	0.040	-
arXiv preprint 2025 [44]	LJSpeech	-	0.003	0.105	-
	LibriTTS	-	0.021	0.000	-
arXiv preprint 2025 [45]	LJSpeech	-	0.003	0.385	-
	LibriSpeech	-	0.024	0.152	-
	LibriTTS	-	0.020	0.029	-
arXiv preprint 2025 [100]	Dolly CW	-	-	-	-
	Librispeech	-	-	-	-
arXiv preprint 2025 [102]	GTZAN	-	-	-	-

Table 8. Fidelity metrics for audio generative model watermarking schemes.

Method	Function	Dataset	Reliability (ACC)↑
Interspeech 2020 [8]	Copyright Protection	AN4	1.000
		Command Voice	1.000
		LibriSpeech	1.000
USENIX Security 2021 [29]	Copyright Protection	Speech Command	0.964
APPL ARTIF INTELL 2022 [69]	Copyright Protection	Command Voice	1.000
Symmetry 2022 [92]	Copyright Protection	TIMIT	0.950
ICASSP 2022 [10]	Copyright Protection	LibriSpeech	1.000
Electronics 2023 [105]	Copyright Protection	LibriSpeech	0.960
		TIMIT	0.970
TPS-ISA 2023 [5]	Copyright Protection	Speech Command	0.858
TAI 2024 [9]	Copyright Protection	Speech Command	0.942
		TIMIT	0.978
S&P 2024 [51]	Copyright Protection	Speech Command	1.000
Vis. Intell. 2024 [47]	Copyright Protection	TIMIT	1.000
ACM MM 2024 [50]	Copyright Protection	LJSpeech	0.996
		LibriSpeech	0.994
		LibriTTS	0.995
SPL 2024 [11]	Copyright Protection	LJSpeech	0.999
arXiv preprint 2024 [87]	Copyright Protection	MUSDB18-HQ	1.000
ICASSP 2025 [86]	Copyright Protection	LibriSpeech	1.000
		Slakh2100	1.000
ICASSP 2025 [17]	Copyright Protection	Speech Command	0.991
		Football Keyword	0.992
ICASSP 2025 [35]	Copyright Protection	LibriTTS	0.997
PRL 2025 [40]	Copyright Protection	MusicCaps	0.940
SPL 2025 [18]	Copyright Protection	LJSpeech	0.998
		LibriSpeech	0.995
		LibriTTS	0.998
Interspeech 2025 [99]	Copyright Protection	C4 Subset	0.989
		Dolly-CW	0.570
		MMW Story	0.970
arXiv preprint 2025 [71]	Copyright Protection	LibriTTS	1.000
arXiv preprint 2025 [100]	Copyright Protection	Dolly CW	0.920
		LibriSpeech	0.950
arXiv preprint 2025 [44]	Copyright Protection	LJSpeech	1.000
		LibriTTS	-

Table 9. Audio model watermarking scheme types and reliability copyright protection.

Method	Function	Dataset	Reliability (ACC)↑
ICASSP 2022 [12]	Attribution	LJSpeech	0.990
		SC09	0.980
Interspeech 2024 [107]	Attribution	LibriTTS	1.000
arXiv preprint 2024 [81]	Attribution	VCTK	1.000
		MTG-Jamendo	1.000
arXiv preprint 2025 [45]	Attribution	LJSpeech	0.984
		LibriSpeech	-
		LibriTTS	-
arXiv preprint 2025 [102]	Attribution	GTZAN	0.971

Table 10. Audio model watermarking scheme types and reliability for attribution.

Method	Robustness	Fragility	Soundness	Uniqueness	Blindness	Capacity	Generality	Efficiency
Interspeech 2020 [8]	++		+	+	+		+	++
USENIX Security 2021 [29]			+	+			++	
APPL ARTIF INTELL 2022 [69]	+		++	+				++
Symmetry 2022 [92]			+	+	+			
ICASSP 2022 [12]	+		+	+			+	
ICASSP 2022 [10]	+		++	+	+	+		
Electronics 2023 [105]	+		+	+	+			
TPS-ISA 2023 [5]	+		+	+				
TAI 2024 [9]			+	+	+	+		
S&P 2024 [51]			+	+	+	+	++	
Vis. Intell. 2024 [47]			+		+	+		+
ACM MM 2024 [50]	++		+		+	++	+	
SPL 2024 [11]	+		+				+	+
Interspeech 2024 [107]	+					++	+	
arXiv preprint 2024 [81]	++		+				+	
arXiv preprint 2024 [87]	+		+				+	
ICASSP 2025 [86]	+				+		+	+
ICASSP 2025 [17]	++		+				+	
ICASSP 2025 [35]	++		+				+	
PRL 2025 [40]	++		++		++	+	+	+
SPL 2025 [18]	++		+		++		++	+
Interspeech 2025 [99]	+		++		++		+	+
arXiv preprint 2025 [71]	++		+	+				+
arXiv preprint 2025 [100]	++		++		++		+	+
arXiv preprint 2025 [44]	+			+	+	++	+	+
arXiv preprint 2025 [45]	++		+	++	+	++	+	+
arXiv preprint 2025 [102]	++		++	+	++	+	+	+

Table 11. Performance summary of audio model watermarking schemes: '+' in the table indicates that the performance was only covered in the paper, or only briefly tested, or the data is not outstanding compared to the same category. '++' indicates that the performance is the first time an experiment has been conducted, or the data are prominent in the same category.

Evaluation Metrics	Description
Reliability	Embedded watermarks can be fully extracted or watermarks can be verified with high probability
Fidelity	Embedding the watermark does not impact the model's task performance
Robustness	The watermark can be verified after watermark attacks and should not be modified by unauthorized users
Fragility	Sensitivity of the watermark to changes or attacks on watermarked content
Soundness	The adversary cannot fake watermarked content without knowing the key
Uniqueness	Watermarking methods should embed unique watermarks for each user
Blindness	How much information is needed to extract data from watermarked model
Capacity	Bits of the watermark information
Generality	Watermarking methods can be applied to different audio models
Efficiency	Speed and overhead of the watermark embedding and extraction process

Table 12. Evaluation metrics.

Function	Reliability	Fidelity	Robustness	Fragility	Soundness	Uniqueness	Blindness	Capacity	Generality	Efficiency
Copyright Protection	✓	✓	✓		✓	✓	✓	✓	✓	
Attribution	✓	✓	✓		✓	✓	✓	✓	✓	✓
Tamper Detection	✓	✓		✓				✓	✓	✓

Table 13. Evaluation metrics for different watermarking functions in audio models.

embedding the watermark. However, certain trigger-based methods may introduce a slight perceptible degradation in audio quality on specific datasets. This suggests that while pursuing high robustness, further optimization remains necessary to minimize interference with the timbre and prosody of the generated audio.

4.1 Definition and Calculation Method of Evaluation Metric

4.1.1 **Reliability.** Embedded watermarks can be fully extracted or watermarks can be verified with high probability.

- **Bit Error Rate (BER)**

$$BER = \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\left(b_i' \neq b_i\right), \tag{1}$$

where b_i' is the extracted i-th watermark bit, b_i is the original watermark bit, and T is the watermark length. The lower the BER, the stronger the reliability [27].

- **Detection Accuracy**

$$ACC_D = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

where TP (True Positive) denotes the number of samples correctly detected as containing a watermark, TN (True Negative) denotes the number of samples correctly judged as lacking a watermark, FP (False Positive) denotes the number of false alarm samples, and FN (False Negative) denotes the number of missed detection samples. ACC_D measures the ability of a watermark detector to correctly determine whether content contains a watermark. A higher ACC_D indicates greater accuracy [82].

- **False Positive Rate (FPR)**

$$FPR = \frac{FP}{TN + FP}, \tag{3}$$

where FPR represents the probability of the watermark-free content being incorrectly identified as containing a watermark. The lower the FPR, the greater the reliability [82].

4.1.2 **Fidelity.** Embedding the watermark does not impact the model's task performance.

- **Model Performance Metrics**

$$WER = \frac{Number\ of\ Word\ Errors}{Total\ Words}, \tag{4}$$

$$CER = \frac{Number\ of\ Character\ Errors}{Total\ Characters} \tag{5}$$

The lower the WER and CER, the stronger the fidelity [64].

- **Metric of Model Performance Changes**

$$\|\Delta w\| = \left\|w^* - w\right\|_F \|\Delta W\|_F \tag{6}$$

$$= \sqrt{\sum_{i,j} \left(W_{ij}^* - W_{ij}\right)^2} \tag{7}$$

Among them, w is the parameter of the original model. $\|\Delta w\|$ measures the perturbation in the parameter space. w^* is the parameter of the model after embedding of the watermark. W is the parameter matrix of a certain layer in the original model. W^* is the matrix of corresponding parameters after embedding of the watermark. $\|\cdot\|_F$ is the Frobenius norm, which is used to measure the magnitude of changes in parameters. The smaller the performance change, the stronger the fidelity [64].

- **Short-Time Objective Intelligibility (STOI)**

$$\rho_t = \frac{\sum_k (X_{env}(k) - \mu_X)(Y_{env}(k) - \mu_Y)}{\sqrt{\sum_k (X_{env}(k) - \mu_X)^2 \sum_k (Y_{env}(k) - \mu_Y)^2}}, \tag{8}$$

$$STOI = \frac{1}{T} \sum_{t=1}^{T} \rho_t, \tag{9}$$

where $X_{env}(k)$ and $Y_{env}(k)$ represent the spectral envelopes of the original and distorted speech signals, respectively, derived from the Short-Time Fourier Transform (STFT). μ_X and μ_Y are the means of these spectral envelopes, and T denotes the total number of frames over which the correlation coefficients ρ_t are averaged to compute the final $STOI$ value. The higher the $STOT$, the better the quality of the generated audio [85].

- **Perceptual Evaluation of Speech Quality (PESQ)**

$$PESQ = a_0 - a_1 D_{sym} - a_2 D_{asym}, \tag{10}$$

where D_{sym} represents the average symmetric disturbance, and D_{asym} represents the asymmetric disturbance, taking into account the distinct perceptual effects of additive noise versus signal attenuation. According to

ITU-T P.862 [72], the coefficients are set to $a_0 = 4.5$, $a_1 = 0.1$, and $a_2 = 0.0309$. The higher the *PESQ* score, the better the audio quality.

- **Mean Opinion Score (MOS)**

$$MOS = \frac{1}{N}\sum_{i=1}^{N} S_i, \tag{11}$$

where N represents the number of users participating in the evaluation, while S_i denotes the subjective rating given by the i-th user for speech or audio quality, typically ranging from 1 (poor) to 5 (excellent). The higher the *MOS* score, the better the voice quality [83].

- **Fréchet Audio Distance (FAD)**

$$FAD(\mathbf{Y}, \mathbf{Y}') = \|\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}'}\|_2^2 + \mathrm{Tr}\big(\Sigma_{\mathbf{Y}} + \Sigma_{\mathbf{Y}'} - 2(\Sigma_{\mathbf{Y}}\Sigma_{\mathbf{Y}'})^{\frac{1}{2}}\big), \tag{12}$$

where FAD measures the discrepancy between the distribution of generated audio $\mathbf{Y}'$ and that of the real audio $\mathbf{Y}$. Here, $\mu_{\mathbf{Y}}$ and $\Sigma_{\mathbf{Y}}$ denote the mean vector and covariance matrix of the true audio feature distribution, respectively; $\mu_{\mathbf{Y}'}$ and $\Sigma_{\mathbf{Y}'}$ denote the mean vector and covariance matrix of the watermarked audio feature distribution. FAD quantifies the naturalness and perceived quality of the audio. A lower FAD value indicates that the generated audio's quality and distribution are closer to the original audio, signifying greater fidelity [36].

4.1.3 **Robustness.** The watermark can be verified after watermark attacks and should not be modified by unauthorized users.

- **Bit Error Rate (BER) After Attacks**

$$BER = \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\left(b_i' \neq b_i\right), \tag{13}$$

where b_i' is the extracted i-th watermark bit after attacks, b_i is the original watermark bit, and T is the length of the watermark. The smaller the *BER*, the stronger the robustness [27].

4.1.4 **Fragility.** Fragility measures the sensitivity of the watermark to changes or attacks on watermarked content.

- **Accuracy of tamper detection**

$$ACC = \frac{Number\ of\ Detected\ Tampering}{Total\ Tampering\ Incidents}, \tag{14}$$

where the *Number of Detected Tampering* refers to the number of tampering incidents that were successfully detected. This refers to the number of times that the watermark was able to effectively identify tampering after it occurred. The *Total Tampering Incidents* refers to the total number of tampering incidents that were carried out. This includes all attempts to tamper with the model, regardless of whether the watermark successfully detected them. The higher the *ACC*, the better the fragility [73].

4.1.5 **Soundness.** The adversary cannot fake watermarked content without knowing the key.

- **Probability of Successful Forgery**

$$P_{Forge} = \frac{Number\ of\ Successful\ Forgery}{Total\ Forgery}, \tag{15}$$

$$P_{Soundness} = 1 - P_{Forge}, \tag{16}$$

where P_{Forge} is the probability of the adversary successfully forging the watermark without knowing the key. The $P_{Soundness}$ closer to 1 indicates a more soundness scheme [15].

- **Computational Complexity**

$$Soundness \propto O(f(n)). \tag{17}$$

Among them, $O(f(n))$ the time complexity required for the adversary to forge the watermark, n is the key length or the complexity of the watermarking algorithm [63]. The higher the time complexity, the better the soundness.

- **Information Entropy**

$$H(K) = -\sum_i p(k_i) \log_2 p(k_i), \tag{18}$$

where $H(K)$ is the entropy of the key, and $p(k_i)$ is the probability of key k_i. The higher the entropy value, the stronger the soundness [37].

4.1.6 **Uniqueness.** Watermarking methods should embed unique watermarks for each user.

- **Structural Similarity**

$$Similarity(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|}, \tag{19}$$

where X and Y represent the watermarks of two users. Their similarity can be measured using the cosine similarity. If the similarity is close to 0, it indicates that there is a significant difference between the watermarks, and the uniqueness is high [93].

- **Information Entropy**

$$H(K) = -\sum_i p(k_i) \log_2 p(k_i). \tag{20}$$

If the generation of watermarks depends on the randomness of the key k_i, it can be measured by calculating the information entropy of the key, and $p(k_i)$ is the probability of the key k_i. The higher the entropy value, the stronger the uniqueness of the watermark [37]..

4.1.7 **Blindness.** Blindness represents how much information is needed to extract data from the watermarked model.

- **Ratio of Needed Information**

$$P_{extract} = \frac{Bits\ of\ Needed\ Information}{Total\ Information}. \tag{21}$$

The more bits of information required to extract the watermark, the better the blindness [55].

4.1.8 **Capacity.** Bits of the watermark information.

- **Bit-level Watermark Accuracy**

$$ACC = \frac{Number\ of\ Watermark\ Bits}{Total\ Bits}. \tag{22}$$

The effective capacity of a watermark is the ratio of the number of bits predicted accurately to the total number of bits [80].

4.1.9 **Generality.** Watermarking methods can be applied to different audio models.

- **Transfer Success Rate**

$$P_{success} = \frac{Number\ of\ Applicable\ Model}{Total\ Model}. \tag{23}$$

The larger $P_{success}$ is, the more types of models the method can be transferred to, and the stronger the generality [53].

4.1.10 **Efficiency.** Efficiency is often measured in terms of the speed and overhead of the watermark embedding and extraction process.

- **Process Time Consumption**

$$T = End\ Time - Begin\ Time, \tag{24}$$

where T represents the time consumed. The smaller T is, the higher the process efficiency [43].

- **Model Trainable Parameters Overhead**

$$M_{overhead} = \frac{M_W - M_{Original}}{M_{Original}}. \tag{25}$$

This metric measures the increase in model parameters following the introduction of a watermarking mechanism. $M_{overhead}$ denotes the percentage increase in parameters. M_W represents the trainable parameters of the model after embedding the watermark. $M_{Original}$ denotes the trainable parameters of the original model. A lower value is preferable [43].

- **Runtime Overhead**

$$T_{overhead} = \frac{T_W - T_{Original}}{T_{Original}}. \tag{26}$$

This metric measures the increase in time consumed by the model during inference or generation phases following the introduction of a watermarking mechanism. $T_{overhead}$ denotes the percentage increase in runtime. T_W represents the time required for the watermarked model to complete a single processing task. $T_{Original}$ denotes the time required for the original model to complete the same task. Lower values are preferable [43].

4.2 Evaluation Metrics for Various Watermarking Functions in Audio Models

This section outlines the metrics available for different functionalities and provides corresponding usage examples from the perspective of scheme functionality. This section also presents the distribution of schemes from a functional perspective, as shown in Figures 4 and 5. Table 14 separately examines the efficiency metrics associated with each approach. Existing solutions exhibit relatively low values for Trainable Parameters (M) and Runtime Overhead (s). This represents an area requiring improvement in future design iterations, particularly in light of the practical applicability of watermarking schemes.

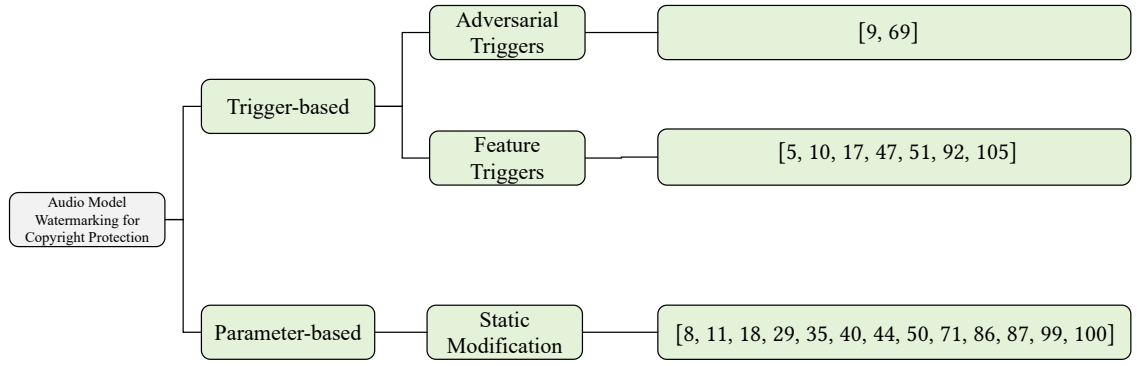

Fig. 4. Taxonomy of Watermarking for Audio Model Copyright Protection.

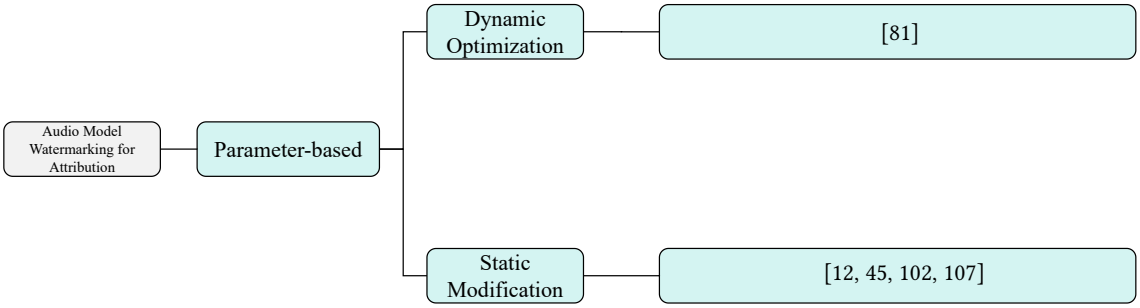

Fig. 5. Taxonomy of Watermarking for Audio Model Attribution.

Method	Efficiency	Model	Trainable Paramete (M)	Runtime Overhead (s)
Interspeech 2020 [8]	✓	Embedder	-	9.767×10^{-2}
		Detector	-	1.098×10^{-2}
S&P 2024 [51]	✓	Embedder	-	2.07×10^{-3}
		Detector	-	-
arXiv preprint 2025 [44]	✓	Embedder	2.25	-
		Detector	0.25	-

Table 14. Efficiency metrics for audio model watermarking solutions.

4.2.1 **Copyright Protection.** Model copyright protection is a common function of watermarking technologies for audio models. By extracting watermarks from the media, owners can verify their ownership and thus protect the audio model copyrights. Metrics such as reliability, fidelity, robustness [94], and uniqueness [52] are the most commonly used.

4.2.2 **Attribution.** Attribution is the model watermarking technique that has the function of indicating its origin and is based on multiclass classification to trace the corresponding models [19]. The attribution can be categorized into model attribution and user attribution. Cho et al. investigated a solution for model attribution, which discussed algorithmic improvements for audio models that empirically achieved high attribution accuracy while maintaining

high generation quality [12]. Li et al. proposed to implement multiple traceability at the content, model, and user levels [45]. The most popular metrics for assessing this function are reliability, fidelity, robustness, and efficiency. Yang et al. employed a dual-watermark embedding mechanism and a consistency loss function to simultaneously embed model and data source watermarks within audio generation models, thereby achieving high-precision dual attribution [102].

4.2.3 **Tamper Detection.** Tamper detection has two aspects, namely, determining whether a sample protected by a watermark has been tampered with and identifying specific areas of the sample that have been tampered with [62]. Fragile model watermarking for tamper detection is one of the critical topics in model watermarking research. While numerous fragile watermarking solutions exist for visual model tamper detection [20], no fragile watermarking techniques have been specifically developed for audio model tamper detection. This notable research gap constitutes a significant omission in audio model watermarking research and represents a key focus for future studies.

Based on fragility [2], model watermarks are classified as robust watermarking, fragile watermarking, and semi-fragile watermarking. Robust watermarking is usually used when resistance to various model attacks is required. Fragile watermarking is used in cases that require high-sensitivity detection, allowing for quick identification of tampering behavior. Semi-fragile watermarking combines the properties of both and is suitable for retaining validity after a certain level of attack, while failing when tampering is evident. Apart from fragility, the most popular metrics for assessing this function are reliability, fidelity, generality, and efficiency.

Regarding computational efficiency and real-time performance in practical deployments, current research papers generally lack systematic quantification and reporting. Only a handful—fewer than ten—provide key efficiency metrics such as trainable parameter counts and model sizes. Given the distinct technical approaches employed, significant variations exist in parameter efficiency across different solutions. SOLIDO's [44] parameter-efficient fine-tuning technique results in substantially lower parameter counts and model sizes compared to other generative watermarking approaches, demonstrating its outstanding advantages in model lightweighting and computational cost reduction. Regarding inference speed, a critical real-time performance metric [11] provides concrete embedding time data, reporting an average embedding time of 2.07 milliseconds per audio clip. This figure, compared to other post-processing watermarking methods listed in the literature (such as Timbre's 7.25 milliseconds [49] and notably PatchMultilayer's 1254 milliseconds [56]), highlights the absolute speed advantage of neural network-based watermarking methods and suggests the unsuitability of traditional algorithms for real-time scenarios. However, few other papers report latency or throughput metrics for their watermark embedding or extraction processes, preventing a comprehensive assessment of real-time performance across end-to-end pipelines.

In summary, within generative speech watermarking, Li et al. [44] represents a technically viable path to high parameter efficiency through advanced training strategies, making it well-suited for computationally and storage-constrained environments. HiFi-GANw [11] demonstrates the fastest inference speed among disclosed data, potentially better suited for synthesis scenarios that require real-time performance. This pervasive data gap profoundly reflects current research's emphasis on feasibility, while generally neglecting usability shortcomings and failing to evaluate computational efficiency.

4.3 Robustness Evaluation Against Watermark Attacks

With the increasing capability of attackers, audio model watermarking technology faces various security challenges. We have also summarized the corresponding metrics for robustness evaluation against watermark attacks, as shown in Table 15.

Watermark Attack	Robustness Evaluation Against Watermark Attacks								
	Reliability	Fidelity	Robustness	Soundness	Uniqueness	Blindness	Capacity	Generality	Efficiency
Watermark Detection	✓	✓						✓	
Watermark Forgery	✓	✓	✓	✓	✓	✓	✓	✓	✓
Watermark Removal	✓	✓				✓		✓	✓

Table 15. Evaluation metrics for robustness against watermark attacks.

4.3.1 **Watermark Detection.** In watermark detection, the goal of the attacker is to detect whether the content contains any hidden watermark or not. One method uses offset learning to isolate the effects of watermarking by comparing watermarked and non-watermarked models to identify the embedded watermark [106]. Reliability, fidelity, efficiency, and blindness are usually the metrics used for performance evaluation. Watermark detection, as the core functionality of the system, employs a comprehensive evaluation logic. It demands an optimal balance between high reliability and high uniqueness, which collectively form the foundation of detection accuracy. Blindness defines the functional application paradigm, whilst efficiency directly determines the feasibility of the detection process in practical implementation. Furthermore, fidelity during initial embedding ensures the imperceptibility of the watermark, and capacity determines its information-carrying capability. Together with generality, these elements underpin the functional breadth and practicality of the watermarking system.

In the paper [106], reliability is specifically quantified as the watermarking success rate (WSR). Experimental data indicate that under in-distribution settings, the WSR exceeds 80%, demonstrating the mechanism's efficacy as proof of ownership. Should the WSR prove inadequate, the entire copyright protection system would become virtually ineffective. Currently, uniqueness ensures the exclusivity of watermark detection, preventing false positives when no Watermark Trigger is present. The paper employs a dedicated classifier trained to recognize the category of generated watermarked audio. This design inherently imposes stringent uniqueness requirements, guaranteeing the singularity and authority of copyright claims while preventing erroneous attribution. Copyright verification is considerably more practical in real-world settings because the detection process is fully blind, requiring only a secret key and no access to the original training data or models. With respect to fidelity, the watermark embedding must not degrade the quality of audio generated under normal conditions. The paper uses objective metrics to show that the generation quality of the watermarked model remains largely unaffected in benign scenarios. Additionally, the audio produced during the detection process should itself maintain high perceptual quality.

4.3.2 **Watermark Forgery.** In a watermark forgery attack, a watermark is forged, and the content containing the faked watermark is confirmed to be authentic. Usually, the attacker knows the embedding method and embeds their watermark in the same way [105]. There is also a way to forge by reversing the watermark embedding method through the synthesized results or intermediate outputs with watermarks. Performance evaluation typically uses metrics such as blindness, fidelity, reliability, and efficiency. For watermark forgery attacks, the focus of the evaluation framework has shifted from survivability to anti-counterfeiting capability. While security remains central, the emphasis here lies more on the cryptographic robustness and authentication mechanisms of the system, preventing key compromise or replication of watermark patterns. Blindness similarly serves as a typical prerequisite for such attacks, where the system's anti-counterfeiting capability faces its true test in the absence of reference to the original carrier. The most direct and critical quantitative metric for evaluating anti-forgery capability is uniqueness, representing the system's

false positive rate. A system with high uniqueness implies a low probability of counterfeit watermarks being mistaken for genuine ones. Similar to watermark removal attacks, fidelity also acts as a constraint here, as forgers similarly wish their counterfeit watermarked carriers to remain undetected.

In [105], watermark forgery takes the form of an ambiguity attack. In this attack, an adversary who has acquired the protected model attempts to challenge the uniqueness of the original owner's copyright claim by embedding a forged watermark, thereby creating ambiguity and dispute over attribution. Uniqueness in this scenario directly determines whether a watermarking scheme can resist forgery attacks. Experiments demonstrate that the verification success rate for the owner's watermark remains above 90%, whilst the attacker's watermark fails to achieve comparable verification efficacy. This indicates that the watermarking system effectively distinguishes genuine from counterfeit watermarks, with the owner's watermark possessing unambiguous uniqueness. Consequently, even when subjected to forgery attacks, it continues to provide clear and credible proof of ownership.

4.3.3 **Watermark Removal.** Watermark removal attacks are a class of attacks in which the attacker tries to remove or destroy the watermark information embedded in the model, and the specific existing techniques include regeneration attacks [24], adversarial attacks [31], and editing attacks [24]. Regeneration attacks attempt to destroy the watermark information inside the model by performing various transformation operations on the original model. Adversarial attacks make the watermark detection algorithm unable to recognize the watermarked signal correctly by adding imperceptible perturbations to the input and intermediate layers. The editing attack removes the watermark signal by non-destructive editing of the model. Performance evaluation usually relies on metrics including blindness, reliability, fidelity, and efficiency. When evaluating a system's resilience against watermark removal attacks, the core logic lies in verifying the survivability of the watermark. This directly and primarily relates to the Security metric, as removal attacks are inherently malicious and targeted. Blindness defines the most prevalent attack scenario here, wherein the attacker cannot obtain the original carrier, significantly increasing the attack's feasibility. We ultimately quantify attack success by measuring the post-attack Reliability of the watermark. Throughout this process, Fidelity serves as a critical constraint, as a successful removal attack must not only substantially reduce reliability but also ensure that the processed carrier remains visually or aurally viable. Efficiency, meanwhile, pertains more to the practicality of the watermarking scheme itself rather than its direct resistance to removal.

In [40], watermark removal is concretized as a series of malicious operations designed to disrupt or eliminate watermarks. The core attack scenario is the cut-and-paste attack, supplemented by multiple signal processing attacks. The attacker's fundamental objective is to render watermark detectors ineffective, thereby enabling the unauthorized use of audio content while falsely claiming authorship. Robustness stands as the foremost objective and direct measure of this scheme's defense against watermark removal attacks. It quantifies the watermark's capacity to retain intact information and remain correctly detectable after enduring diverse malicious manipulations. The paper quantifies robustness through the core metric Bit Error Rate (BER). The scheme specifically addresses the low BER observed in targeted cut-and-paste attacks, validating the effectiveness of its barcode-shaped embedding strategy in the one-dimensional Fourier domain. This approach exploits the stability of frequency distribution along the audio signal's time axis, thereby resisting localized deletion. Reliability in this context is directly correlated with the credibility of the watermark detection function, ensuring the detection system consistently produces accurate judgments. The paper evaluates reliability through detection accuracy and Area Under the Curve (AUC). High reliability forms the foundation for a watermark system to serve as valid legal evidence in real-world disputes, guaranteeing the success rate of copyright assertion. Furthermore, fidelity, as a fundamental constraint metric, plays a crucial role in the logical

chain against watermark removal. Attackers aim to remove watermarks without significantly compromising audio quality. Consequently, an excellent watermarking scheme must strike a balance between high robustness and high fidelity. The paper demonstrates through Fresche Audio Distance (FAD) metric data that its watermark embedding process exerts a negligible perceptual impact on the generated audio.

4.4 Trade-off Analysis in Watermark Performance Evaluation

Within the domain of audio model watermarking, evaluating the merits of a solution typically involves balancing three core and mutually constraining dimensions: reliability, fidelity, and robustness [42]. A watermarking scheme rarely achieves optimal performance across all three simultaneously [48], but a strategic trade-off based on the target application scenario.

Reliability vs. Fidelity: Increasing watermark capacity often comes at the expense of model performance. More complex embedded watermarks, while offering higher reliability, introduce greater interference with the original model's weights.

Robustness vs. Fidelity: To enhance resistance against attacks, embedded watermark signals must be stronger and more dispersed, inevitably introducing greater perceptual distortion.

Robustness vs. Reliability: Certain highly robust watermarks may necessitate statistical detection through extensive querying, thereby compromising the reliability of the detection process.

Taking the paper [18] as an example, we conduct an in-depth analysis of its trade-off strategy across metrics. This approach falls under parameter-modification-based methods, aiming to protect model intellectual property by embedding watermarks directly into the convolutional layer weights of the generative model. To achieve high fidelity, the approach incorporates a normalization step to calibrate the modified kernel weights, ensuring that they are aligned with the statistical characteristics of the original model. This strategy successfully minimizes audio quality degradation, guaranteeing the imperceptibility of the watermark. However, to ensure the watermark's robustness against common audio attacks, the scheme must selectively embed the watermark within the convolutional layers of the generative model, rather than merely in shallow layers.

Watermark extraction relies on a dedicated encoder-decoder architecture, which falls under the category of white-box detection. This design ensures nearly perfect recovery reliability in the absence of attacks. Crucially, as the watermark is deeply encoded within the model's core convolutional parameters, it exhibits high resistance to structural attacks and audio signal manipulations. However, this high-intensity embedding inevitably increases design and training complexity, imposing greater demands on model training stability and convergence speed. This approach strategically leverages the strengths of white-box techniques. Through meticulous parameter modulation and normalization techniques, it achieves balance between robustness and fidelity. This demonstrates that watermarking research, while prioritizing practicality, must rely on the characteristics of deep learning architectures to perform refined trade-offs between key metrics.

Based on the performance data consolidated in Tables 7 and 8, a detailed analysis is conducted on the trade-off between reliability, fidelity, and robustness across various watermark embedding strategies.

4.4.1 **Comparison Analysis Based on Parameter Modification.** Dynamic parameter modification techniques deeply embed watermark information into model weights, thereby conferring enhanced robustness. As evidenced by Table 13, the dynamic parameter modification approach demonstrates strong resilience against structural attacks on adversarial models. In contrast, while static parameter modification guarantees high fidelity, its embedded information

proves relatively vulnerable when subjected to weight pruning. Static parameter modification typically achieves the highest fidelity among the four methods, reflecting a minor trade-off in the original model's generative capability due to the coupling within the training process.

4.4.2 **Comparison Analysis Based on Trigger Sets.** The intrinsic benefit of trigger set-based methods is their capacity for black-box detection and control over the results of content creation. Adversarial sample-based trigger sets emphasize stealthiness, whereas feature-based trigger set techniques typically show reduced capacity. They trigger watermark output through minor perturbations imperceptible to the human ear. Table 8 demonstrates that such triggers remain relatively vulnerable to channel attacks. The watermark's detection accuracy quickly decreases if little disturbances in the input signal are disrupted.

5 THE FUTURE OF AUDIO MODEL WATERMARKING

5.1 Method Analysis Summary

According to the analysis of the various methods, we have the following observations. Watermarking based on the modification of model parameters generally shows strong robustness. These methods can effectively resist common attacks such as fine-tuning and pruning, but usually require white-box access permissions [8, 81]. In contrast, trigger-set watermarking shows more flexibility and is particularly suitable for watermark verification in black-box environments [51]. However, their robustness is relatively weaker compared to the former, and they are prone to attacks such as watermark forgery and watermark removal [47, 92, 105]. Since parameter-based watermarks are often embedded in the low-significance bits or redundant subspaces of model weights, they are highly sensitive to network quantization and weight pruning. In trigger-set methods, the primary failure mode is catastrophic forgetting during fine-tuning. In summary, if security and copyright traceability are the core objectives, watermarking based on model parameter modification can be selected [12, 45, 102, 107]. If method flexibility is prioritized, watermarking based on trigger set methods may be more practical [47, 51]. The combination of multiple methods is also emerging as an important trend to enhance robustness and generality.

5.1.1 **Shortcomings of the Existing Approach in Audio Models.** For the category of models with fewer existing watermarking solutions, each faces its own distinct challenges [6]. The main obstacle for speech translation models is cross-lingual acoustic interference [97]. It is challenging for the watermark to remain resilient and language-agnostic due to differences in pronunciation, prosody, and linguistic structures among several languages. The complexity is increased by making sure that the watermark does not affect the quality of the translation or the consistency of the semantics.

In Music Information Retrieval, the diversity of musical genres and the presence of complex harmonic structures introduce high variability in spectral and temporal features [66], which can easily mask or distort the embedded watermark. Maintaining both model fidelity and retrieval accuracy after watermark insertion is particularly challenging.

Watermarking for speech emotion recognition must deal with extremely sensitive emotional signals that are embedded in minute acoustic characteristics, such as energy, tone, and pitch [95]. Even minor perturbations may alter perceived emotion or reduce classification accuracy, posing a trade-off between watermark imperceptibility and model reliability.

In Audio Event Detection [67], the difficulty arises from non-stationary and overlapping environmental sounds, which lead to inconsistent feature representations. Ensuring watermark detectability under varying noise conditions and mixed-signal scenarios continues to pose a significant research challenge. Model categories with fewer alternative

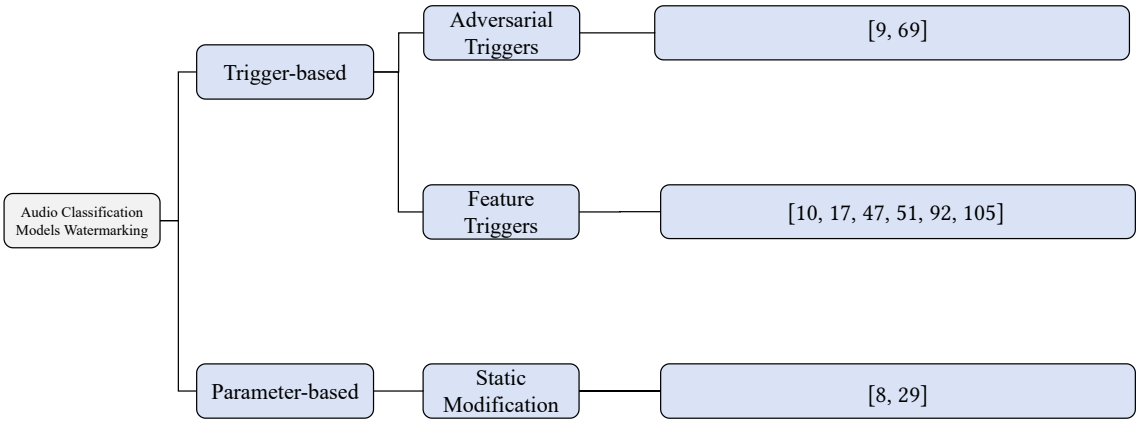

Fig. 6. Taxonomy of Watermarking for Audio Classification Models.

solutions encounter challenges similar to those outlined above. Addressing these issues is critical to the development of effective and practical watermarking schemes.

For speech translation models, watermarking technology faces the dual challenge of cross-lingual acoustic feature shifts and semantic-prosodic consistency [16]. Specifically, speech translation entails not merely textual conversion but the reconstruction of acoustic features from the source language to those of the target language. If the watermark is embedded at the encoder stage, it must survive complex cross-lingual attention mechanisms and decoding processes that often reshape the prosodic structure of audio, potentially causing watermark information loss [57]. Conversely, embedding watermarks at the decoder stage requires ensuring that the watermark signal does not interfere with semantic expression of the target language. For instance, in tonal languages such as Mandarin, fine-tuning fundamental frequency for watermark embedding may erroneously alter word tones, resulting in severe semantic errors or unnatural machine-generated speech. Consequently, future research must explore semantically aware watermarking mechanisms that adaptively embed information within the redundancy space of acoustic features without compromising the target language's unique prosodic rules or semantic integrity.

Moreover, Table 5 indicates that models with a greater number of existing methods are typically those with well-established application markets. In contrast, audio event detection and music information retrieval, while promising modeling tasks, currently have no associated watermarking methods. Similarly, speech translation models have very limited watermarking approaches, despite the fact that this area has been developing rapidly and attracting increasing attention in recent years [79]. The paper also summarizes relevant solutions by model type, as illustrated in Figure 6 and Figure 7.

5.1.2 **Comparison with Other Modalities.** Compared to image model watermarking, audio model watermarking presents unique challenges and opportunities, primarily stemming from the one-dimensional temporal nature of audio data and the sensitivity of the human auditory system. Firstly, in terms of data dimension and structure, images constitute two-dimensional spatial data with high local correlation, permitting substantial information concealment within texturally complex regions without detection [54]. In contrast, audio signals are one-dimensional time series with strong temporal correlations, and contemporary audio models depend extensively on long-range context. As a result, even minor perturbations introduced during watermark embedding in model parameters or intermediate layers may accumulate over time, producing audible phase distortion or pop noise that disrupts waveform continuity.

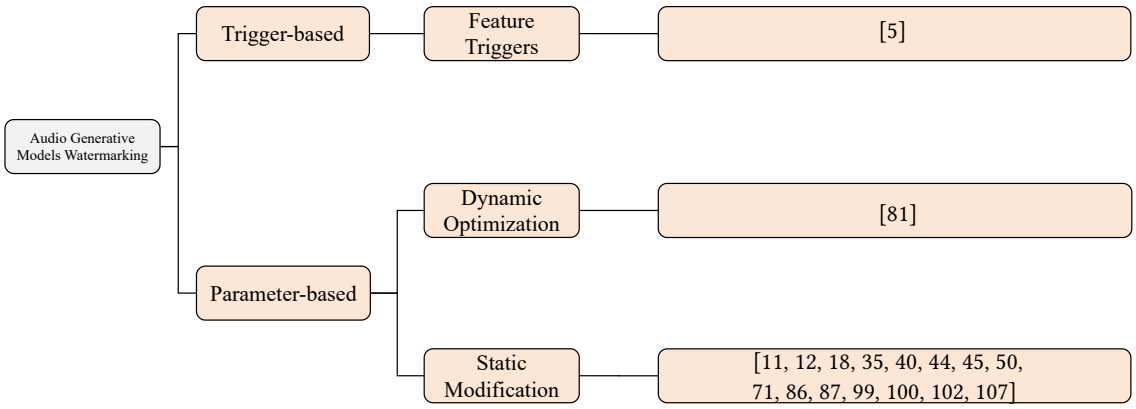

Fig. 7. Taxonomy of Watermarking for Audio Generative Models.

Regarding perceptual masking effects, the human visual system exhibits low sensitivity to high-frequency noise, whereas the auditory system is highly sensitive to temporal desynchronization and frequency perturbations [22]. This substantially limits the embedding capacity of audio watermarks by requiring them to function under incredibly tight frequency or temporal masking thresholds.

Regarding robustness against attacks, audio faces fundamentally different desynchronization threats [40] compared to images. While images can be cropped and scaled, audio-specific methods like time-stretching [86], pitch-shifting [100] can seriously interfere with the watermark signal's synchronization. Particularly for trigger-based approaches, designing a trigger pattern that remains precisely recognizable by the model after temporal displacement proves considerably more challenging than creating a cropping-resistant image trigger. Consequently, designing audio model watermarks often necessitates more difficult trade-offs between capacity, imperceptibility, and resistance to desynchronization attacks.

5.2 Future Work Direction

As shown in Table 5, some aspects of current audio model watermarking have had significant success, while others have not received enough attention because they are new or lack widespread application in the market. We summarize several research directions as follows:

- Developing fragile watermarking techniques for audio models to enable tamper detection is essential and represents a core challenge in model watermarking research. Although there are numerous fragile watermarking solutions for visual model tamper detection [20], no fragile watermarking techniques have been specifically developed for audio model tamper detection. For fragile audio model watermarks, the blurred boundary between benign editing and malicious tampering represents a critical issue that requires resolution. This notable research gap constitutes a significant omission in audio model watermarking research and represents a key focus for future studies.
- Watermarking methods are needed for the models used in music information retrieval, audio event detection, speech emotion recognition, and speech translation. They may be misused to illegally replicate or profit in faking music albums, faking sound events, and cloning voice. Each model type faces unique challenges, including cross-lingual acoustic interference in speech translation, which threatens watermark robustness, and high feature variability in music information retrieval, which complicates the balance between fidelity and retrieval

performance [65]. In addition, speech emotion recognition and audio event detection models must embed watermarks within sensitive or non-stationary acoustic features without affecting perceptual quality or task accuracy.

- Studies are required to improve the robustness of the watermark methods against a more diversified attack, such as combining adversarial training and watermarking techniques. Existing watermarking schemes demonstrate reasonable resilience against rudimentary attacks, yet their robustness remains constrained when confronted with sophisticated, targeted assaults. Future research should concentrate on developing diverse countermeasures that are seamlessly integrated into the training process.
- Methods are also required for the attribution of audio watermarking methods combined with biometric features. In deepfake scenarios, a reliable deepfake detection model should simultaneously satisfy transferability, interpretability, and robustness [91].

  Future watermarking attribution methods are expected to be integrated with the biometric features that are inherent to the audio content itself to achieve simultaneous tracing of the model, the generated data, and the user.
- Another promising direction for future research lies in the development of cross-modal generic embedding architectures and the design of cross-modal watermark consistency verification algorithms. The challenge lies in designing a watermark based on functional abstraction that does not rely on the specific layer structure of the model, but rather on its generative properties, thereby enabling cross-architecture detection.

6 CONCLUSION

This paper has provided a comprehensive analysis of the major developments in audio model watermarking technology in the past five years, with a focus on a functional analysis of audio model watermarking technology and a discussion of performance evaluation metrics. Based on the results of the existing literature review, we identify gaps in existing work and propose research directions that can be explored, providing inspiration for the future development for watermarking audio models. In-depth exploration of these emerging directions, particularly innovations in fragile watermarking, semantic preservation, and cross-modality aspects, will be key to advancing audio model watermarking towards practical implementation.

REFERENCES

[1] Agit Amrullah, Ferda Ernawan, Anis Farihan Mat Raffei, and Liew Siau Chuin. 2025. TDSF: Two-Phase Tamper Detection in Semi-Fragile Watermarking using Two-Level Integer Wavelet Transform. *Engineering Science and Technology, an International Journal* 61 (2025), 101909.

[2] Md Asikuzzaman and Mark R Pickering. 2017. An Overview of Digital Video Watermarking. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2017), 2131–2153.

[3] Flavio Bertini, Alessandro Benetton, and Danilo Montesi. 2025. Distributed Ledger and Text Watermarking for Fine-Grain Provenance Checking of Textual Content. In *Companion Proceedings of the ACM on Web Conference 2025*. Association for Computing Machinery, New York, NY, USA, 2626–2633.

[4] Franziska Boenisch. 2021. A Systematic Review on Model Watermarking for Neural Networks. *Frontiers in Big Data* 4 (2021), 729663.

[5] Xirong Cao, Xiang Li, Divyesh Jadav, Yanzhao Wu, Zhehui Chen, Chen Zeng, and Wenqi Wei. 2023. Invisible Watermarking for Audio Generation Diffusion Models. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, Atlanta, GA, USA, 193–202.

[6] S. Chandrakala and S. L. Jayalakshmi. 2019. Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance: A Survey and Comparative Studies. *ACM Comput. Surv.* 52, 3, Article 63 (June 2019), 34 pages. doi:10.1145/3322240

[7] Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. 2023. WavMark: Watermarking for Audio Generation. arXiv:2308.12770 [cs.SD] https://arxiv.org/abs/2308.12770

[8] Huili Chen, Bita Darvish Rouhani, and Farinaz Koushanfar. 2020. SpecMark: A Spectral Watermarking Framework for IP Protection of Speech Recognition Systems.. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*. ISCA, ShanghaiChina, 2312–2316.

[9] Haozhe Chen, Jie Zhang, Kejiang Chen, Weiming Zhang, and Nenghai Yu. 2023. Model Access Control Based on Hidden Adversarial Examples for Automatic speech Recognition. *IEEE Transactions on Artificial Intelligence* 5, 3 (2023), 1302–1315.

[10] Haozhe Chen, Weiming Zhang, Kunlin Liu, Kejiang Chen, Han Fang, and Nenghai Yu. 2022. Speech Pattern Based Black-Box Model Watermarking for Automatic Speech Recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 3059–3063.

[11] Xiangyu Cheng, Yaofei Wang, Chang Liu, Donghui Hu, and Zhaopin Su. 2024. HiFi-GANw: Watermarked Speech Synthesis via Fine-Tuning of HiFi-GAN. *IEEE Signal Processing Letters* 31 (2024), 2440–2444.

[12] Yongbaek Cho, Changhoon Kim, Yezhou Yang, and Yi Ren. 2022. Attributable Watermarking of Speech Generative Models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 3069–3073.

[13] Ingemar Cox, Matthew Miller, Jeffrey Bloom, and Chris Honsinger. 2002. Digital Watermarking. *Journal of Electronic Imaging* 11, 3 (2002), 414–414.

[14] Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Zhe Peng, Qian Wang, and Chao Shen. 2025. A Survey of Defenses Against AI-Generated Visual Media: Detection, Disruption, and Authentication. *ACM Comput. Surv.* 58, 5, Article 123 (Nov. 2025), 35 pages. doi:10.1145/3770916

[15] Lei Fan, Chenhao Tang, Weicheng Yang, and Hong-Sheng Zhou. 2024. Two Halves Make a Whole: How to Reconcile Soundness and Robustness in Watermarking for Large Language Models. Cryptology ePrint Archive, Paper 2024/2062. https://eprint.iacr.org/2024/2062

[16] Muhammad Umar Farooq and Thomas Hain. 2022. Investigating the Impact of Crosslingual Acoustic-Phonetic Similarities on Multilingual Speech Recognition. In *Interspeech 2022*. ISCA, Incheon, Korea, 3849–3853. doi:10.21437/Interspeech.2022-10916

[17] Hao Fei, Hewang Nie, Siqi Sun, Songfeng Lu, Ting Luo, Ling Qian, Dunbo Cai, Zhiguo Huang, and Runqing Zhang. 2025. Optimized Dynamic Watermarking for Audio DNNs with Adaptive Embedding and Boundary Sampling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5.

[18] Yan Feng, Xuebin Zhang, Fuyuan Feng, Guanglin Zhang, and Longting Xu. 2025. Robust and Imperceptible Watermarking Framework for Generative Audio Models. *IEEE Signal Processing Letters* 32 (2025), 3196–3200.

[19] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Paris, France, 22466–22477.

[20] Zhenzhe Gao, Yu Cheng, and Zhaoxia Yin. 2026. A Survey of Fragile Model Watermarking. *Signal Processing* 238 (2026), 110088.

[21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing SSurveys (CSUR)* 51, 5 (2018), 1–42.

[22] Charles F. Hall and Ernest L. Hall. 1977. A Nonlinear Model for the Spatial Characteristics of the Human Visual System. *IEEE Transactions on Systems, Man, and Cybernetics* 7, 3 (1977), 161–170. doi:10.1109/TSMC.1977.4309680

[23] F. Hartung and M. Kutter. 1999. Multimedia Watermarking Techniques. *Proc. IEEE* 87, 7 (1999), 1079–1107.

[24] Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. 2024. Can Watermarks Survive Translation? On the Cross-Lingual Consistency of Text Watermark for Large Language Models. arXiv:2402.14007 [cs.CL] https://arxiv.org/abs/2402.14007

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-Rank Adaptation of Large Language Models. *ICLR* 1, 2 (2022), 3.

[26] Guang Hua and Andrew Beng Jin Teoh. 2023. Deep Fidelity in DNN Watermarking: A Study of Backdoor Watermarking for Classification Models. *Pattern Recognition* 144 (2023), 109844.

[27] M. Jeruchim. 1984. Techniques for Estimating the Bit Error Rate in the Simulation of Digital Communication Systems. *IEEE Journal on Selected Areas in Communications* 2, 1 (1984), 153–170. doi:10.1109/JSAC.1984.1146031

[28] Shengpeng Ji, Ziyue Jiang, Jialong Zuo, Minghui Fang, Yifu Chen, Tao Jin, and Zhou Zhao. 2025. Speech Watermarking with Discrete Intermediate Representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, Philadelphia, PA, USA, 24239–24247.

[29] Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. 2021. Entangled Watermarks as a Defense Against Model Extraction. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Vancouver, B.C., Canada, 1937–1954.

[30] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. 2023. Evading Watermark Based Detection of AI-Generated Content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, USA, 1168–1181.

[31] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. 2023. Evading Watermark Based Detection of AI-Generated Content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, USA, 1168–1181.

[32] Neil F Johnson, Zoran Duric, and Sushil Jajodia. 2001. *Information Hiding: Steganography and Watermarking-Attacks and Countermeasures*. Vol. 1. Springer Science & Business Media, USA.

[33] Nikola Jovanović, Ismail Labiad, Tomas Soucek, Martin Vechev, and Pierre Fernandez. 2025. Watermarking Autoregressive Image Generation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., San Diego, CA, USA. https://openreview.net/forum?id=hVdD72iom4

[34] Lauri Juvela and Xin Wang. 2024. Collaborative Watermarking for Adversarial Speech Synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Seoul, Korea, Republic of, 11231–11235.

[35] Lauri Juvela and Xin Wang. 2025. Audio Codec Augmentation for Robust Collaborative Watermarking of Speech Synthesis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5.

[36] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. arXiv:1812.08466 [eess.AS] https://arxiv.org/abs/1812.08466

[37] Li Kong, Hao Pan, Xuewei Li, Shuangbao Ma, Qi Xu, and Kaibo Zhou. 2019. An Information Entropy-Based Modeling Method for the Measurement System. *Entropy* 21, 7 (2019). doi:10.3390/e21070691

[38] Feifei Kou, Yuhan Yao, Siyuan Yao, Jiahao Wang, Lei Shi, Yawen Li, and Xuejing Kang. 2025. IWRN: A Robust Blind Watermarking Method for Artwork Image Copyright Protection Against Noise Attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, Philadelphia, PA, USA, 370–378.

[39] Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and Björn W. Schuller. 2023. Sparks of Large Audio Models: A Survey and Outlook. arXiv:2308.12792 [cs.SD] https://arxiv.org/abs/2308.12792

[40] Kyungryeol Lee, Seongmin Hong, and Se Young Chun. 2025. Robust Watermarks for Audio Diffusion Models by Quadrature Amplitude Modulation. *Pattern Recognition Letters* 198 (2025), 22–28.

[41] Suyoung Lee, Wonho Song, Suman Jana, Meeyoung Cha, and Sooel Son. 2023. Evaluating the Robustness of Trigger Set-Based Watermarks Embedded in Deep Neural Networks. *IEEE Transactions on Dependable and Secure Computing* 20, 4 (2023), 3434–3448. doi:10.1109/TDSC.2022.3196790

[42] Jingyang Li and Guoqiang Li. 2025. Triangular Trade-off between Robustness, Accuracy, and Fairness in Deep Neural Networks: A Survey. *ACM Comput. Surv.* 57, 6, Article 140 (Feb. 2025), 40 pages. doi:10.1145/3645088

[43] Tao Li and Lizy Kurian John. 2003. Run-Time Modeling and Estimation of Operating System Power Consumption. In *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (San Diego, CA, USA) *(SIGMETRICS '03)*. Association for Computing Machinery, New York, NY, USA, 160–171. doi:10.1145/781027.781048

[44] Yue Li, Weizhi Liu, and Dongdong Lin. 2025. SOLIDO: A Robust Watermarking Method for Speech Synthesis via Low-Rank Adaptation. arXiv:2504.15035 [cs.CR] https://arxiv.org/abs/2504.15035

[45] Yue Li, Weizhi Liu, and Dongdong Lin. 2025. TriniMark: A Robust Generative Speech Watermarking Method for Trinity-Level Attribution. arXiv:2504.20532 [cs.MM] https://arxiv.org/abs/2504.20532

[46] Yue Li, Hongxia Wang, and Mauro Barni. 2021. A Survey of Deep Neural Network Watermarking Techniques. *Neurocomputing* 461 (2021), 171–193.

[47] Junpei Liao, Liang Yi, Wenxin Shi, Wenyuan Yang, Yanmei Fang, and Xin Yang. 2024. Imperceptible Backdoor Watermarks for Speech Recognition Model Copyright Protection. *Visual Intelligence* 2, 1 (2024), 23.

[48] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024. A Survey of Text Watermarking in the Era of Large Language Models. *ACM Comput. Surv.* 57, 2, Article 47 (Nov. 2024), 36 pages. doi:10.1145/3691626

[49] Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. 2024. Detecting Voice Cloning Attacks via Timbre Watermarking. In *Network and Distributed System Security Symposium*. ISOC, San Diego, CA, USA. doi:10.14722/ndss.2024.24200

[50] Weizhi Liu, Yue Li, Dongdong Lin, Hui Tian, and Haizhou Li. 2024. Groot: Generating Robust Watermark for Diffusion-Model-Based Audio Synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 3294–3302.

[51] Peizhuo Lv, Hualong Ma, Kai Chen, Jiachen Zhou, Shengzhi Zhang, Ruigang Liang, Shenchen Zhu, Pan Li, and Yingjun Zhang. 2024. MEA-defender: A Robust Watermark Against Model Extraction Attack. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 2515–2533.

[52] Yihan Ma, Zhengyu Zhao, Xinlei He, Zheng Li, Michael Backes, and Yang Zhang. 2023. Generative Watermarking Against Unauthorized Subject-Driven Image Synthesis. arXiv:2306.07754 [cs.CV] https://arxiv.org/abs/2306.07754

[53] John McCarthy. 1987. Generality in Artificial Intelligence. *Commun. ACM* 30, 12 (Dec. 1987), 1030–1035. doi:10.1145/33447.33448

[54] Han Meng and Fangru Guo. 2021. Image Classification and Generation Based on GAN Model. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE, Taiyuan, China, 180–183. doi:10.1109/MLBDBI54094.2021.00042

[55] Seung-Min Mun, Seung-Hun Nam, Haneol Jang, Dongkyu Kim, and Heung-Kyu Lee. 2019. Finding Robust Domain from Attacks: A Learning Framework for Blind Watermarking. *Neurocomputing* 337 (2019), 191–202. doi:10.1016/j.neucom.2019.01.067

[56] Iynkaran Natgunanathan, Yong Xiang, Guang Hua, Gleb Beliakov, and John Yearwood. 2017. Patchwork-Based Multilayer Audio Watermarking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 11 (2017), 2176–2187.

[57] H. Ney. 1999. Speech Translation: Coupling of Recognition and Translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99*, Vol. 1. IEEE, Phoenix, AZ, USA, 517–520 vol.1. doi:10.1109/ICASSP.1999.758176

[58] Hong-Hanh Nguyen-Le, Van-Tuan Tran, Thuc Nguyen, and Nhien-An Le-Khac. 2025. A Survey on Proactive Deepfake Defense: Disruption and Watermarking. *ACM Comput. Surv.* 58, 5, Article 126 (Nov. 2025), 37 pages. doi:10.1145/3771296

[59] Srinivas Nidhra and Jagruthi Dondeti. 2012. Black Box and White Box Testing Techniques-A Literature Review. *International Journal of Embedded Systems and Applications (IJESA)* 2, 2 (2012), 29–50.

[60] Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. 2025. Deep Audio Watermarks are Shallow: Limitations of Post-Hoc Watermarking Techniques for Speech. arXiv:2504.10782 [cs.SD] https://arxiv.org/abs/2504.10782

[61] Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. 2024. Maskmark: Robust Neuralwatermarking for Real and Synthetic Speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Seoul, Korea, Republic of, 4650–4654.

[62] Minzhou Pan, Zhenting Wang, Xin Dong, Vikash Sehwag, Lingjuan Lyu, and Xue Lin. 2024. Finding Needles in a Haystack: A Black-Box Approach to Invisible Watermark Detection. arXiv:2403.15955 [cs.CV] https://arxiv.org/abs/2403.15955

[63] Christos H. Papadimitriou. 2003. *Computational complexity*. John Wiley and Sons Ltd., GBR, 260–265.

[64] Chanho Park, Hyunsik Kang, and Thomas Hain. 2024. Character Error Rate Estimation for Automatic Speech Recognition of Short Utterances. In *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, Lyon, France, 131–135. doi:10.23919/EUSIPCO63174.2024.10715433

[65] Geoffroy Peeters, Zafar Rafii, Magdalena Fuentes, Zhiyao Duan, Emmanouil Benetos, Juhan Nam, and Yuki Mitsufuji. 2025. Twenty-Five Years of MIR Research: Achievements, Practices, Evaluations, and Future Challenges. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5.

[66] Nikki Pelchat and Craig M. Gelowitz. 2020. Neural Network Music Genre Classification. *Canadian Journal of Electrical and Computer Engineering* 43, 3 (2020), 170–173. doi:10.1109/CJECE.2020.2970144

[67] Jose Portelo, Miguel Bugalho, Isabel Trancoso, Joao Neto, Alberto Abad, and Antonio Serralheiro. 2009. Non-Speech Audio Event Detection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Taipei, Taiwan, 1973–1976. doi:10.1109/ICASSP.2009.4959998

[68] Changyu Rao, Gaozhi Liu, Sheng Li, Xinpeng Zhang, and Zhenxing Qian. 2025. DynMark: A Robust Watermarking Solution for Dynamic Screen Content with Small-size Screenshot Support. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 7463–7471. doi:10.1145/3746027.3754897

[69] Pulkit Rathi, Saumya Bhadauria, and Sugandha Rathi. 2022. Watermarking of Deep Recurrent Neural Network using Adversarial Examples to Protect Intellectual Property. *Applied Artificial Intelligence* 36, 1 (2022), 2008613.

[70] Kui Ren, Ziqi Yang, Li Lu, Jian Liu, Yiming Li, Jie Wan, Xiaodi Zhao, Xianheng Feng, and Shuo Shao. 2024. SoK: On the Role and Future of AIGC Watermarking in the Era of Gen-AI. arXiv:2411.11478 [cs.CR] https://arxiv.org/abs/2411.11478

[71] Yong Ren, Jiangyan Yi, Tao Wang, Jianhua Tao, Zheng Lian, Zhengqi Wen, Chenxing Li, Ruibo Fu, Ye Bai, and Xiaohui Zhang. 2025. P2Mark: Plug-and-play Parameter-Level Watermarking for Neural Speech Generation. arXiv:2504.05197 [cs.SD] https://arxiv.org/abs/2504.05197

[72] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. 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*, Vol. 2. IEEE, Salt Lake City, UT, USA, 749–752 vol.2. doi:10.1109/ICASSP.2001.941023

[73] Preston K. Robinette, Thuy Dung Nguyen, Samuel Sasaki, and Taylor T. Johnson. 2026. Trigger-Based Fragile Model Watermarking for Image Transformation Networks. In *Computer Security – ESORICS 2025*, Vincent Nicomette, Abdelmalek Benzekri, Nora Boulahia-Cuppens, and Jaideep Vaidya (Eds.). Springer Nature Switzerland, Cham, 346–365.

[74] Rowena Rodrigues. 2020. Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities. *Journal of Responsible Technology* 4 (2020), 100005.

[75] Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. 2024. Proactive Detection of Voice Cloning with Localized Watermarking. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. JMLR.org, Vienna, Austria, Article 1759, 17 pages.

[76] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. 2019. Deepsigns: an End-To-End Watermarking Framework for Protecting the Ownership of Deep Neural Networks. In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, Vol. 3. Association for Computing Machinery, New York, NY, USA, 1.

[77] Aditya Kumar Sahu, M Hassaballah, Routhu Srinivasa Rao, and Gulivindala Suresh. 2023. Logistic-Map Based Fragile Image Watermarking Scheme for Tamper Detection and Localization. *Multimedia Tools and Applications* 82, 16 (2023), 24069–24100.

[78] Robin San Roman, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel. 2025. Latent Watermarking of Audio Generative Models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5.

[79] Mohammad Sarim, Saim Shakeel, Laeeba Javed, Jamaluddin, and Mohammad Nadeem. 2025. Direct Speech to Speech Translation: A Review. arXiv:2503.04799 [cs.CL] https://arxiv.org/abs/2503.04799

[80] S.D. Servetto, C.I. Podilchuk, and K. Ramchandran. 1998. Capacity Issues in Digital Image Watermarking. In *Proceedings 1998 International Conference on Image Processing. ICIP98*, Vol. 1. IEEE, Chicago, IL, USA, 445–449 vol.1. doi:10.1109/ICIP.1998.723521

[81] Mayank Kumar Singh, Naoya Takahashi, Wei-Hsiang Liao, and Yuki Mitsufuji. 2024. LOCKEY: A Novel Approach to Model Authentication and Deepfake Tracking. arXiv:2409.07743 [cs.CR] https://arxiv.org/abs/2409.07743

[82] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI 2006: Advances in Artificial Intelligence*, Abdul Sattar and Byeong-ho Kang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1015–1021.

[83] Robert Streijl and David Hands. 2016. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems* 22 (03 2016), 213–227. doi:10.1007/s00530-014-0446-1

[84] Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-Language Models for Audio-Centric Tasks: A Survey. arXiv:2501.15177 [cs.SD] https://arxiv.org/abs/2501.15177

[85] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Dallas, TX, USA, 4214–4217. doi:10.1109/ICASSP.2010.5495701

[86] Yi Tang. 2025. Poisoning The Diffusion: A Simple and Robust Watermarking Method for Audio Generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5.

[87] Christopher J. Tralie, Matt Amery, Benjamin Douglas, and Ian Utz. 2024. Hidden Echoes Survive Training in Audio To Audio Generative Instrument Models. arXiv:2412.10649 [cs.SD] https://arxiv.org/abs/2412.10649

[88] Luís Vilaça, Yi Yu, and Paula Viana. 2025. A Survey of Recent Advances and Challenges in Deep Audio-Visual Correlation Learning. *ACM Comput. Surv.* 57, 12, Article 299 (July 2025), 46 pages. doi:10.1145/3696445

[89] Heng Wang, Hongxia Wang, Mingze He, Fei Zhang, and Jinghong Xia. 2025. Robust Blind Video Watermarking Against Digital Editing and Camcording. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 11 (2025), 11068–11082.

[90] Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Dayong Ye, Wanlei Zhou, and Philip Yu. 2025. Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey. *ACM Comput. Surv.* 58, 4, Article 83 (Oct. 2025), 36 pages. doi:10.1145/3764113

[91] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. 2024. Deepfake Detection: A Comprehensive Survey from the Reliability Perspective. *ACM Comput. Surv.* 57, 3, Article 58 (Nov. 2024), 35 pages. doi:10.1145/3699710

[92] Yumin Wang and Hanzhou Wu. 2022. Protecting the Intellectual Property of Speaker Recognition Model by Black-Box Watermarking in the Frequency Domain. *Symmetry* 14, 3 (2022), 619.

[93] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

[94] Zihan Wang, Olivia Byrnes, Hu Wang, Ruoxi Sun, Congbo Ma, Huaming Chen, Qi Wu, and Minhui Xue. 2023. Data Hiding with Deep Learning: A Survey Unifying Digital Watermarking and Steganography. *IEEE Transactions on Computational Social Systems* 10, 6 (2023), 2985–2999.

[95] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. 2021. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* 9 (2021), 47795–47814. doi:10.1109/ACCESS.2021.3068045

[96] Steffen Wendzel, Luca Caviglione, Wojciech Mazurczyk, Aleksandra Mileva, Jana Dittmann, Christian Krätzer, Kevin Lamshöft, Claus Vielhauer, Laura Hartmann, Jörg Keller, Tom Neubert, and Sebastian Zillien. 2025. A Generic Taxonomy for Steganography Methods. *ACM Comput. Surv.* 57, 9, Article 233 (May 2025), 37 pages. doi:10.1145/3729165

[97] Peter Wu, Jiatong Shi, Yifan Zhong, Shinji Watanabe, and Alan W Black. 2021. Cross-Lingual Transfer for Speech Processing Using Acoustic Language Similarity. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, Cartagena, Colombia, 1050–1057. doi:10.1109/ASRU51503.2021.9688276

[98] Shaowu Wu, Wei Lu, Xiaolin Yin, and Rui Yang. 2025. Robust Watermarking Against Arbitrary Scaling and Cropping Attacks. *Signal Processing* 226 (2025), 109655.

[99] Yihan Wu, Ruibo Chen, Georgios Milis, Junfeng Guo, and Heng Huang. 2025. A Watermark for Auto-Regressive Speech Generation Models. In *Proc. Interspeech 2025*. ISCA, Rotterdam, The Netherlands, 3474–3478.

[100] Yihan Wu, Georgios Milis, Ruibo Chen, and Heng Huang. 2025. Robust Distortion-Free Watermark for Autoregressive Audio Generation Models. arXiv:2510.21115 [cs.SD] https://arxiv.org/abs/2510.21115

[101] Qiang Xu, Wenpeng Mu, Jianing Li, Tanfeng Sun, and Xinghao Jiang. 2025. Advancements in AI-Generated Content Forensics: A Systematic Literature Review. *ACM Comput. Surv.* 58, 3, Article 60 (Sept. 2025), 36 pages. doi:10.1145/3760526

[102] Xuefeng Yang, Jian Guan, Feiyang Xiao, Congyi Fan, Haohe Liu, Qiaoxi Zhu, Dongli Xu, and Youtian Lin. 2025. DualMark: Identifying Model and Training Data Origins in Generated Audio. arXiv:2508.15521 [cs.SD] https://arxiv.org/abs/2508.15521

[103] Peigen Ye, Huali Ren, Zhengdao Li, Anli Yan, Hongyang Yan, Shaowei Wang, and Jin Li. 2025. Securing Large Language Models: A Survey of Watermarking and Fingerprinting Techniques. *ACM Comput. Surv.* (Dec. 2025). doi:10.1145/3773028 Just Accepted.

[104] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. 2021. Deep Model Intellectual Property Protection via Deep Watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4005–4020.

[105] Jing Zhang, Long Dai, Liaoran Xu, Jixin Ma, and Xiaoyi Zhou. 2023. Black-Box Watermarking and Blockchain for IP Protection of Voiceprint Recognition Model. *Electronics* 12, 17 (2023), 3697.

[106] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. 2024. Invisible Image Watermarks Are Provably Removable Using Generative AI. In *Advances in Neural Information Processing Systems*, Vol. 37. Curran Associates, Inc., New Orleans, Louisiana, 8643–8672.

[107] Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, Ye Bai, Chu Yuan Zhang, Yong Ren, and Zhengqi Wen. 2024. TraceableSpeech: Towards Proactively Traceable Text-To-Speech with Watermarking. arXiv:2406.04840 [cs.SD] https://arxiv.org/abs/2406.04840