European Commission

Technical Solutions for Marking and Detecting AI-generated Audio Content in the Context of Article 50(2) AI Act

Final Study Report

Written by
Xavier Serra [1], R. Oguz Araz [1], Roser Batlle Roca [1], Lauri Juvela [2], David López [1], Martín Rocamora [1]
[1] Universitat Pompeu Fabra, Barcelona, Spain
[2] Aalto University, Espoo, Finland
contact: xavier.serra@upf.edu

EUROPEAN COMMISSION
Directorate-General for Communications Networks, Content and Technology (CNECT)
Directorate A — Artificial Intelligence Office
Unit A.2 — Regulation and Compliance
Contact: Nandi Robijns
E-mail: nandi.robijns@ec.europa.eu
European Commission
B-1049 Brussels

Technical Solutions for Marking and Detecting AI-generated Audio Content in the Context of Article 50(2) AI Act

Final study report

Directorate-General for Communications Networks, Content and Technology (CNECT)

More information on the European Union is available on the Internet (http://www.europa.eu).
Luxembourg: Publications Office of the European Union, 2026
KK-01-25-157-EN-N
ISBN 978-92-68-34421-7
doi:10.2759/0784462

Table of contents

1. SUMMARY

This report reviews the current state of technical solutions and research on **marking and detecting AI-generated audio** within the framework of **Article 50 of the EU AI Act**. It aims to support the development of **Codes of Practice** addressing transparency obligations by providing an informed and comprehensive analysis of existing methods, their capabilities, and limitations. Specifically, the report: (1) examines the principal mechanisms for audio marking and detection, (2) proposes a set of evaluation criteria aligned with the AI Act's requirements, (3) assesses state of the art approaches against these criteria, (4) identifies their main shortcomings and technical gaps, and (5) outlines strategic directions for future research and standardization.

The report distinguishes between **marking**, **detection**, and **identification**, which are often used interchangeably but serve distinct roles. *Marking* refers to the intentional embedding or attachment of information that signals AI generation (e.g., metadata or watermarks). *Detection* involves verifying whether such marks are present or have been altered. *Identification* extends further, aiming to infer the origin of the content — such as the generative model used — even when no explicit mark exists. This distinction underpins the comparative analysis and evaluation presented in the report.

The analysis is based on an extensive review of **technical literature, existing standards, and industrial practices**, complemented by **expert consultations** involving researchers, industry specialists in watermarking and synthetic media detection, and policymakers.

The evaluation shows that **no single technology currently fulfils all requirements** of Article 50, namely effectiveness, interoperability, robustness, and reliability. **Metadata**, particularly when combined with cryptographic protection, ensures transparency and ease of integration but is vulnerable to being stripped or altered. **Audio watermarking** enables the embedding of persistent identifiers but faces challenges in robustness and interoperability. **Audio fingerprinting** effectively verifies known content but cannot identify new or modified AI-generated material. **Generative model identification** can detect AI-specific artefacts without prior marking, yet it lacks generalization and is vulnerable to manipulation.

Given these limitations, the report advocates for a **multi-layered approach** that combines complementary methods. A robust framework should integrate **cryptographically protected metadata** for transparency, **imperceptible watermarking** for persistence, and **forensic detection techniques** for unmarked content. **Fingerprinting** remains useful in specific, controlled scenarios where reference databases are available.

Ensuring reliable deployment and regulatory compliance will require **continued research and standardization**, particularly to enhance robustness, interoperability, and adaptability to evolving generative models. Achieving these objectives calls for **coordinated efforts** among academia, industry, and regulators, supported by the **European Commission's role** in promoting open standards, funding public infrastructures (e.g., reference datasets and benchmarks), and fostering collaborative ecosystems.

Ultimately, this report contributes to the European Commission's ongoing initiative to define **Codes of Practice** that guide the practical implementation of transparent, accountable, and technically sound solutions for the marking and detection of AI-generated audio.

2. INTRODUCTION

To promote the uptake of human-centric and trustworthy artificial intelligence, the AI Act requires providers of certain generative AI systems to mark their output in a machine-readable format that makes it identifiable as artificially generated or manipulated. Providers should also ensure that their technical solutions are effective, interoperable, robust, and reliable as far as this is technically feasible. These obligations are specified in Article 50(2) (marking of AI-generated content).

The primary objective of this report is to assess audio marking solutions in terms of their effectiveness in fulfilling the obligations outlined in the AI Act. We (1) study their effectiveness with respect to several proposed criteria, (2) highlight areas where existing approaches fall short of these criteria, and (3) propose research and development efforts to improve them.

For clarity, this report uses the following terminology throughout: **Marking** designates any *proactive* mechanism by which a creator or system embeds or attaches information to an audio signal indicating its AI-generated nature. Typical examples include metadata or digital watermarks. **Detection** refers to the *verification or discovery* of such marks, including determining their validity or detecting their removal or manipulation. **Identification** denotes a *forensic analysis process* that seeks to determine whether a given audio recording was generated by AI — even in the absence of explicit marking — often by recognizing characteristic artefacts of specific generative models. These three processes are complementary but conceptually distinct: marking enables transparency at the source, detection ensures their persistence and verifiability, and identification provides accountability when provenance information is missing. This conceptual framework guides the technical and regulatory analysis developed in the following sections.

A typical end-to-end process of generating and using audio with generative AI comprises several stages. First, an AI model is developed using specific training data, which may include audio recordings, symbolic music scores, text, or other relevant data modalities. The AI model is then integrated into a generation system capable of producing audio signals based on user-provided controls, such as text prompts or other interactive inputs. A user employs this system to create specific audio content, which can then be distributed through digital platforms where consumers can search, select, and listen to the content.

Marking, detection, and identification mechanisms can be integrated at different stages of this generative AI pipeline. During model training or audio generation, proactive **marking** techniques — such as embedding metadata or imperceptible watermarks — can be applied to ensure that AI-generated content carries persistent identifiers from its origin. At the point of distribution, platforms can **detect** and preserve such identifiers to maintain provenance information. Post hoc **identification** methods

can be used to analyze the generated audio to determine its origin, for example, by matching audio fingerprints against a trusted reference database, or by detecting characteristic artifacts left by a specific generative model. Such identification can be carried out by distribution platforms prior to making content available, or by third parties, including regulators and rights holders, to verify authenticity and provenance. Thus, marking supports transparency from the moment of creation, while detection and identification enable ongoing verification throughout the content's lifecycle.

While AI-driven music generation encompasses both symbolic representations (such as MIDI files) and audio synthesis, this report does not explicitly cover the marking of symbolic music generation. Given that MIDI and other symbolic formats are more akin to text in their structure, they are better addressed using metadata-based or text-based transparency techniques rather than those designed for audio signals. Our focus remains on marking methods applicable to waveform-based audio content, aligning with the primary concerns of the AI Act regarding AI-generated media that is directly consumed in audio format. The transparency of other modalities of content, such as text and image, is examined in other reports commissioned by the Commission.

Article 50(2) of the AI Act exempts AI systems that serve an assistive function for standard editing or do not substantially alter input data from transparency marking requirements. However, the lack of clear criteria introduces challenges for enforcement and compliance, as minor yet perceptually significant modifications to audio could be excluded from regulation depending on interpretation. Given this ambiguity, our report focuses on AI-generated content where the transformation is substantial, while acknowledging the need for further legal and technical clarity on this exemption. Determining what constitutes a **substantial alteration** versus **standard editing** is central to this discussion. The former implies a fundamental change to the content's meaning or authenticity, such as creating a deepfake voice that attributes new statements to an individual. The latter refers to common enhancements like noise reduction, equalization, or adding reverb, which clean or improve the existing audio without misleading the listener about its original source or message. This distinction hinges on the intent and effect of the AI tool, counterbalancing the benefits of technological assistance with the need to prevent deceptive manipulation.

In this report, we first describe the **methodology** adopted for the study, combining a comprehensive review of existing literature and standards with expert consultations. We then provide an **overview of the state of the art**, highlighting the main technologies available for marking and detecting AI-generated audio. Building on this foundation, we propose a **set of evaluation criteria** aligned with the requirements of the AI Act and apply them to assess the identified solutions. The report then presents a **comparative analysis** of these technologies, identifying their principal strengths and shortcomings and offering a **European perspective** on their relevance and maturity. Finally, we outline **future research directions** and summarize the key takeaways that can inform forthcoming policy and standardization initiatives.

3. METHODOLOGY

This report is based on a comprehensive review of existing technical solutions and qualitative insights gathered from various stakeholders. The methodology was designed to ensure a thorough and balanced assessment, integrating documented evidence and practical experience.

A thorough review of technical solutions was conducted by surveying existing academic literature, industry reports, and technical standards. Sources were selected based on their credibility, relevance to the topic, publication recency, the maturity level of the described solutions, and contribution to standardization efforts.

To complement the technical review, qualitative insights were gathered from stakeholders actively engaged in leading initiatives and developing technical solutions for the audio and music sector. Inputs were collected through a combination of semi-structured interviews and an online questionnaire. Please refer to Appendix A for the online questionnaire and an indicative list of topics and questions discussed during the interviews. Additionally, roundtable discussions and panel sessions were held to gather insights and concerns from other stakeholders.

Stakeholders were selected based on expertise, professional experience, and organizational representation. Participants included academic researchers, industry practitioners and researchers, policymakers and regulatory experts, as well as representatives from non-governmental organizations (NGOs) and community groups. The selection aimed to ensure a diversity of perspectives, covering both technical and socio-economic dimensions of the topic.

Semi-structured interviews were conducted with more than a dozen participants via video conferencing or in-person, depending on their availability and preference. The semi-structured interview format allowed flexibility in exploring individual expertise areas while maintaining consistency across discussions. Participants were informed of the study's purpose and consented to their insights being used and their names being added to the report.

The consultations focused on current technical solutions for marking, detecting, and identifying AI-generated audio, and their assessment in terms of (1) technical effectiveness, (2) integration and practicality, (3) information security and adaptability, and (4) compliance and standardization support.

Public consultation was conducted by distributing an open questionnaire to a broader group of stakeholders. The questionnaire was designed to capture structured feedback on the specific technologies that were previously identified: (1) Marking with Audio **Metadata**, (2) Marking with Audio **Watermarking**, (3) Identification with Audio **Fingerprinting**, and (4) **Identification** of Generative Audio Models. Practical examples of each technology were provided for assessment, and participants were asked to express their opinion on their adequacy and answer questions related to their shortcomings and potential research directions that could help improve them. Additionally, they were asked to report any other solution they were aware of that was not considered in the questionnaire. Participants were informed that their responses

to the questionnaire would remain confidential and would be used solely for the purpose of writing the report.

In addition, on 4 September 2025, the European Commission organized an online workshop on *Transparency in AI-generated Content*, which gathered a wide range of stakeholders from academia, industry, and policy. During this workshop, we presented the overall conclusions of the first draft of this report. Feedback was collected both verbally during the discussions and subsequently in written form. This input has been integrated into the present version of the report, ensuring that the conclusions reflect a broader consensus and address the concerns raised by the participants.

Insights from stakeholders were analyzed in conjunction with the technical review findings. Where subjective assessments (e.g., perceived feasibility, stakeholder confidence) were involved, stakeholder input was triangulated with available literature to ensure consistency and robustness. Where appropriate, expert judgment from the authors supplemented stakeholder inputs, particularly in areas where empirical evidence was limited. In these cases, assumptions were explicitly stated to maintain transparency and accountability.

4. STATE OF THE ART

In response to the growing need for reliable methods to verify the authenticity and provenance of audio content, a variety of technologies have been proposed and developed. This section reviews the current state of the art across complementary approaches, providing the basis for evaluating technical solutions against the regulatory and practical requirements of the AI Act.

While Article 50(2) of the AI Act focuses specifically on the marking of AI-generated content as such—implying a legally mandated binary classification of unknown content as AI-generated or not—this report adopts a broader perspective. We consider marking, detection, and identification as part of a continuum of capabilities relevant to transparency and provenance. Marking techniques such as metadata or watermarking can carry richer information beyond the binary label, enabling more detailed authenticity checks and provenance tracking once detected. Moreover, even when content has not been deliberately marked, identification methods—such as audio fingerprinting or the detection of distinctive artifacts left by specific AI models—can help determine origin and establish whether the content has been artificially generated or manipulated. By incorporating these broader methods, the report addresses not only the AI Act's core marking and detection requirements but also additional capabilities that can enhance the integrity, accountability, and traceability of audio content in practical deployment.

We first examine **audio metadata**, which has long been essential for organizing, managing, and securing audio content. Traditional metadata standards and recent advancements in **cryptographic** protection and provenance frameworks are discussed to highlight their role in ensuring content integrity and traceability.

Next, we address marking techniques based on **audio watermarking**, where information is imperceptibly embedded within the audio signal itself. We review both

classical signal processing-based approaches and emerging deep learning methods, analyzing their trade-offs in robustness, perceptibility, and standardization potential.

We then explore **audio fingerprinting**, a technique that focuses on identifying audio content without marking it, through distinctive signal features, without relying on embedded metadata. Both traditional signal processing-based methods and modern neural network approaches are considered, with a focus on their robustness, scalability, and applicability to AI-generated content verification.

Finally, we survey recent work on the **identification of generative audio models**, an area that leverages the intrinsic artifacts left by AI generation processes to trace content back to its source. Techniques for detecting decoder and vocoder artifacts are discussed, with an emphasis on their relevance for content attribution without explicit marking.

Together, these four areas provide a comprehensive overview of the technical landscape for AI-generated audio content marking and identification, informing the design and evaluation of solutions aligned with the requirements of the AI Act.

4.1 AUDIO METADATA

Metadata is structured information used to describe the properties of content. When applied to AI-generated content, this metadata acts as a mark attached to the content itself. We distinguish between two metadata-based techniques: **plain metadata**, which simply attaches the AI-generated label; and **cryptographic methods**, which use a secure **digital signature** as metadata. This signature is computed using the generative AI provider's private key and is nearly impossible to forge, ensuring that the content's origin can be reliably verified with the corresponding public key.

Audio metadata is structured, encoded data that provides detailed information about the content and representation characteristics of an audio file, which facilitates the discovery, assessment, interpretation, management, generation, manipulation, and distribution of the described audio (Kriechbaum, 2009). The National Information Standards Organization (NISO) from the United States establishes that metadata can be classified into four different types (Riley, 2017):

1) Descriptive metadata, which provides information about the content of the audio file to aid in discovery and understanding, such as the title, artist, album, genre, and release date,
2) Administrative metadata, which includes information necessary to manage the audio file or understand its creation, including technical specification, preservation metadata for long-term management, and rights metadata with intellectual property (IP) details,
3) Structural metadata, which captures the relationships between parts of the audio file, such as the arrangement of tracks in an album or the sequence of sections in a composition, and
4) Mark-up languages, which combine metadata and content within the same file, using embedded tags or markers to denote specific features, such as tempo changes, lyrics, or annotations.

Audio metadata plays a crucial role across various domains, playing a key role in the organization, distribution, and preservation of audio content. For instance, music streaming platforms and digital service providers (DSPs), such as Spotify, Apple Music, and YouTube, rely on descriptive metadata to categorize music (e.g., by genre, artist, and mood), enabling music discovery, appropriate attribution and credit, as well as IP management.[1,2] Furthermore, in the broadcasting industry, metadata especially supports royalty management by linking audio content to creators and rights holders, ensuring accurate compensation through systems such as the International Standard Recording Code (ISRC) (IFPI, 2021).

The implementation of audio metadata involves a range of established formats and standards. For instance, ID3 tags (Nilsson and Sundstrom, 1999) are commonly used in MP3 files to encode audio information, while Broadcast Wave Format (BWF) supports professional workflows by embedding technical metadata, such as time codes and loudness data (EBU, 2011). Such standards ensure that metadata is embedded, stored, and exchanged across different platforms and devices, allowing consistent audio content management and distribution.

Cryptographic techniques are essential to ensure audio metadata's security, authenticity, and integrity. These techniques help prevent tampering, forgery, and unauthorized modifications. Common methods to protect audio metadata include digital signatures, hash functions, and public key infrastructure (PKI) (Qadir and Varol, 2019; Hardjono, 2019). Digital signatures authenticate audio metadata by generating a unique cryptographic signature based on the metadata's content and a secret key (Ali, 2017). The signature is verifiable by third parties without revealing the signer's private key, ensuring that any alteration to the metadata invalidates the signature. Advanced methods, such as audio digital signature algorithms based on Principal Component Analysis (PCA), leverage wavelet packet decomposition and eigenvector extraction to enhance robustness against tampering (Yang, 2009). Alternatively, hash functions generate fixed-length metadata representations, acting as digital fingerprints that detect even minor modifications. Since these functions are collision-resistant, they ensure that any unauthorized changes to the metadata result in a completely different hash value, making tampering easily detectable. Then, PKI systems use asymmetric encryption to authenticate and protect metadata, ensuring that only authorized parties can access or modify it. Public-key cryptosystems, such as RSA (Rivest, Shamir & Adleman, 1978) further enhance metadata security by enabling digital signatures and encrypted transmission, thereby safeguarding metadata integrity in distributed environments.

Recent advancements in metadata technologies have introduced new methodologies for enhancing the authenticity and origin of content. One significant development is the Coalition for Content Provenance and Authenticity (C2PA): an open technical specification designed to embed cryptographic signatures and detailed data provenance into the metadata of audio files. This standard helps verify the origin and integrity of audio content, addressing challenges related to copyright infringement and the misuse of deepfake audio (C2PA, 2024). By providing verifiable information about the creation and modification of audio files, C2PA supports trust and transparency in

[1] https://distromono.com/how-to/music-metadata-dna-distribution/

[2] https://sonosuite.com/en/blog/what-is-music-metadata-and-why-is-important-to-digital-music/

the digital content ecosystem. Another relevant example is the Secure Evidence Attribution Label (SEAL) specification (Krawetz, 2024). SEAL is an open solution for media attribution that cryptographically signs a file to enable decentralized verification of authorship without requiring a central authority. SEAL verifies the file's signature against a public key retrieved via DNS (Domain Name System). While this method supports attribution regardless of whether content is original, AI-generated, or modified, it does not track semantic integrity or transformation history and does not constitute legal proof of originality.

4.2 MARKING WITH AUDIO WATERMARKING

Audio watermarking refers to information embedded in an audio signal in such a way that it is imperceptible to the human ear but can still be reliably detected or extracted by specific algorithms. It is a form of steganography designed for audio content and can be used to embed metadata such as content origins, ownership details, timestamps, or unique IDs. In practice, capacity, perceptibility, and robustness are interdependent. For instance, increasing robustness often requires embedding the watermark with higher signal energy or in more resilient frequency bands, which can make the watermark audible and reduce perceptual transparency. Conversely, ensuring imperceptibility may limit the strength of the watermark, making it easier to remove through compression or filtering. Applications differ in the balance they require: in broadcast monitoring or forensic tracing, robustness is prioritized even if a very slight audible distortion is tolerated, since the watermark must survive multiple transmission steps; in consumer music distribution, perceptibility constraints are stricter, and fidelity preservation takes precedence over extreme robustness; in secure archival use cases, higher payload capacity may be emphasized, so that detailed provenance information can be embedded even if robustness is moderate. Finding the right balance among these factors is therefore highly application-dependent (Bender et al., 1996). Traditionally, the motto for watermarking has been to "Keep Honest People Honest" by providing a moderate level of security that, in most applications, is not comparable to cryptographic security guarantees (Cox et al. 2006).

Established methods for audio watermarking are based on digital signal processing (DSP), and use techniques such as echo hiding (Gruhl et al., 1996) and spread spectrum watermarking (Cox et al., 1996). DSP-based techniques typically rely on manually designed transformations in the time or frequency domain, where human expertise is used to select features of the audio signal that can be modified without perceptible distortion. Technology based on DSP is mature and used in commercial solutions such as Digimarc (Alattar et al., 2020), based on spread spectrum watermarking. Benefits of DSP-based methods include interpretability, computational efficiency, and potential for backward compatibility. However, technical details of commercial methods are often not publicly available, and it is difficult to compare the robustness of commercial solutions due to a lack of unified evaluation benchmarks.

More recently, there has been increased research activity in deep-learning–based audio watermarking techniques (Deep Watermarks) from major industry research teams, for example, WavMark (Microsoft; Chen et al., 2023), MaskMark (Adobe; O'Reilly et al., 2024), AudioSeal (Meta; San Roman et al., 2024), SilentCipher (Sony; Singh et al., 2024), and SynthID (Google; Dathathri et al., 2024). Unlike DSP methods, deep learning approaches automatically learn where and how to embed information

by optimizing over large datasets, potentially discovering more robust and imperceptible watermarking strategies. While deep watermarking methods show great promise for improved robustness and adaptability, they are computationally more expensive and less standardized than DSP approaches. The emerging deep watermarking methods have no established standards yet, but there are recent research efforts to establish benchmarks for robustness (Liu et al., 2024). Furthermore, current work has focused on singular systems and developing cross-compatible solutions requires future research.

Some standards in DSP-based audio watermarking already exist, but these methods were not initially designed for public disclosure purposes and require further evaluation for robustness against tampering and fitness for purpose. The Advanced Television Systems Committee (ATSC) has proposed a Standard A/334: Audio Watermark Emission (ATSC, 2024). This method is based on modifying the least-significant bit in sub-band autocorrelation signals. The standard does not discuss the scope, limitations, and robustness in detail, but it appears to be intended for detecting tampering in broadcast audio signals. Society of Motion Picture and Television Engineers Standard ST 2112-10:2018 (SMPTE, 2020) describes a watermarking method based on frequency-domain phase modulation and provides recommendations for adjusting the watermark strength to balance robustness with perceptibility.

The Digital Cinema Initiatives (DCI) Specification (DCI, 2024) sets out general requirements for forensic watermarking in professional cinema workflows. It specifies conditions for watermark location, robustness (or “survivability” under common transformations), and the types of metadata (payload) that can be embedded, such as distributor identifiers or session information. Unlike other standards, however, the DCI specification deliberately avoids prescribing implementation details of particular watermarking algorithms, leaving flexibility for different vendors to develop proprietary solutions as long as they meet the survivability and payload requirements. As a result, it serves as a framework for interoperability and compliance in digital cinema rather than a technical standardization of watermarking techniques.

4.3 IDENTIFICATION WITH AUDIO FINGERPRINTING

Audio fingerprinting methods generate compact and distinctive representations from audio clips. These methods enable the identification of audio content even after common signal alterations such as audio degradations (compression, EQ, additive background noise, etc.) or transformations (pitch shifting, time stretching, etc.). Unlike methods that rely on embedding metadata in the audio (Section 4.1) or watermarking (Section 4.2), audio fingerprinting methods depend solely on the unmodified audio signal. No additional information is embedded in the file, nor is any additional information required to accompany it. All metadata is stored only with the reference recordings (known content) in the database, while the unknown audio clip to be identified carries nothing but its raw audio.

Historically, audio fingerprinting employed landmark-based methods, which involve detecting energy peaks within a signal’s time-frequency representation to characterize the audio content. Work by Cano et al. (2002), Wang (2006), Six and Leman (2014), and Sonnleitner and Widmer (2016) highlighted how these peaks, due to their higher

energy relative to surrounding frequencies, withstand various signal distortions. The spatial and temporal coordinates of these peaks are encoded as binary hashes, enhancing retrieval efficiency. Rather than comparing individual peaks, the method records the relative positions of three or more peaks to form a peak constellation. This unique arrangement is referred to as a 'fingerprint.' Traditionally, the extraction of these constellations was governed by rule-based algorithms derived from signal properties.

Recent advancements have seen the introduction of neural networks in generating these fingerprints, thus eliminating the need for manually designed rule-based algorithms. Modern approaches utilize time-frequency representations like mel-spectrograms and deploy neural networks such as CNNs (Agüera y Arcas et al., 2017), LSTMs (Báez-Suárez, 2020), or Transformers (Singh et al., 2023) to embed the audio content into fingerprints. While some methods generate real-valued fingerprints (Arcas et al., 2017; Yu et al., 2020; Chang et al., 2020; Singh et al., 2022), offering greater expressiveness but requiring more storage and slower matching, others continue to produce binary-valued fingerprints, keeping with the tradition (Báez-Suárez, 2020; Singh et al., 2020; Wu et al., 2022).

Neural network-based methods for audio fingerprinting require the simulation of targeted audio degradations and musical transformations during training. While a comprehensive list of common real-life audio degradations exists, currently no neural network approach comprehensively addresses a broad spectrum of these issues. For instance, the study by Arcas et al. (2017) does not account for any degradation, whereas Báez-Suárez (2020) focuses specifically on time stretching and pitch shifting. Yu et al. (2020) expand the scope to include high-pass and low-pass filtering, additive white noise, echo effects, equalization, and MP3 compression. On the other hand, Chang et al. (2020) and Singh et al. (2023) explore the impacts of additive background noise, room impulse responses, and microphone impulse responses.

Using a fingerprint extraction algorithm, a database of fingerprints is created from an original audio collection and stored with accompanying metadata about the original audio. When the identity of an unknown audio clip is needed, fingerprints of the query audio are extracted and compared against all the fingerprints in the database using retrieval algorithms. To facilitate the identification process, a fingerprint extraction algorithm should be paired with a good retrieval algorithm. Since the database can grow to contain millions of entries, scalability becomes a central consideration: efficient indexing and search strategies (e.g., locality-sensitive hashing, approximate nearest neighbor search) are required to ensure that identification remains fast and reliable at a large scale.

4.4 IDENTIFICATION OF GENERATIVE AUDIO MODELS

This approach, which falls into the field of media forensics, focuses on analyzing audio files to identify telltale signs or artifacts indicative of specific generative AI models used in their creation. By examining patterns inherent to specific AI models, it is possible to attribute audio to its generative source (Wolff et al., 2022). These artifacts, often introduced by the neural decoders or vocoders (the model components responsible of converting low-dimensional features into audio waveforms), can act as an unintentional *signature*: subtle, systematic traces in the signal that result from the

particular architecture, training data, or synthesis method of the model. For example, certain vocoders may leave characteristic spectral smoothing, temporal discontinuities, or statistical irregularities in the waveform that differ from those of other models. Such recurring patterns are not deliberately embedded, as in watermarking, but emerge naturally from the way the model generates audio and can therefore be exploited to identify the generative model used (Li et al., 2024).

Detection of generated speech has long been an active research area in the context of anti-spoofing (Wu et al. 2015) for automatic speaker verification (ASV). Unified evaluation benchmarks are regularly tested in challenges, such as the bi-annual ASVspoof challenge (Liu et al. 2023) or the Audio Deep synthesis Detection (ADD) challenge (Yi et al., 2022). Examples of well-performing models include RawNet2 (Tak et al., 2020) and the graph neural network based AASIST (Jung et al., 2022).

Recent advancements in identifying AI-generated music focus on leveraging the inherent characteristics of neural decoders to establish a link between generated songs and the models that created them. These decoders, particularly in autoencoders (AEs), introduce unique artifacts—such as checkerboard patterns from transposed convolutions—that act as unintentional signatures embedded in the audio (Afchar et al., 2025). By transforming audio into various input representations, such as raw waveforms, the complex STFT, and spectrograms, researchers found that spectrograms most effectively highlight these subtle decoder-specific features. Using a convolutional neural network (CNN) with six convolutional layers and two linear layers, the method detects these unique characteristics.

Leveraging decoder artifacts provides a way to link AI-generated music samples to their generative models, enhancing traceability without relying on explicit watermarking or metadata, which are prone to removal or tampering. These decoder-specific artifacts act as unique *signatures*, inherently embedded in the output of generative models and connecting generated samples to their source. However, maintaining this traceability poses challenges, as artifacts can be degraded or removed through common audio manipulations such as pitch shifting, time stretching, or re-encoding (Zang et al., 2024). Ensuring these artifacts persist despite such transformations is critical for reliable attribution. Additionally, the evolving nature of generative models demands continuous updates to detection pipelines to account for new architectures and artifact patterns. Therefore, despite the fragility of these artifacts under manipulation and the need for continuous adaptation to new models, the underlying principle of using intrinsic decoder artifacts for content attribution remains a significant area of investigation.

Another approach to identifying AI-generated audio sources involves detecting artifacts introduced by neural vocoders, which are essential components in most deep-learning-based speech synthesis models (Sun et al., 2023). Neural vocoders synthesize waveforms from intermediate representations, such as mel spectrograms, and their processing introduces subtle yet consistent distortions. By leveraging these vocoder-specific artifacts, researchers designed a multi-task learning framework based on a modified RawNet2 model. This model jointly performs binary classification—distinguishing real from synthetic voices—while also identifying the specific vocoder used to generate the audio. Including the vocoder identification task alongside binary classification compels the model's feature extraction layers to learn

representations that are highly sensitive to these subtle vocoder-specific characteristics, effectively leveraging them as unintentional signatures of the generative model. The experiments demonstrated that this approach achieves high accuracy and remains robust even against common audio manipulations, such as resampling and background noise, making it a promising tool for AI-generated audio attribution (Sun et al., 2023).

5. EVALUATION CRITERIA FOR AI AUDIO MARKING

Article 50 of the AI Act specifies that technical solutions for marking and detection of AI-generated content should be assessed according to four main criteria: **effectiveness**, **interoperability**, **robustness**, and **reliability**. While these concepts are not formally defined in the Act, they can be interpreted from the broader legislative context as follows:

- **Effectiveness** is the capacity of a solution to fulfil its intended purpose—in this case, to mark or detect AI-generated content accurately and consistently—under normal operating conditions.
- **Interoperability** refers to the ability of the solution to function across diverse technical environments, formats, and platforms, enabling seamless integration into existing ecosystems.
- **Robustness** is the resilience of the solution to intentional attacks, technical transformations, or other adverse conditions, while maintaining its intended performance.
- **Reliability** concerns the consistency and predictability of the solution's performance over time and across different scenarios, ensuring that the results can be trusted by stakeholders.

To provide a more comprehensive and operational basis for assessment, we propose a few more criteria and organize them into four broader categories:

1. **Technical Effectiveness**, encompassing accuracy, robustness, fidelity preservation, and efficiency.
2. **Potential for Integration**, covering interoperability, disruptiveness to existing workflows, and accessibility for diverse stakeholders.
3. **Information Management Capabilities**, addressing information capacity, resilience to technical evolution, and tamper-detection capability.
4. **Compliance and Standardization Support**, including understandability, potential for standardization, and reliability.

Within this framework, robustness is treated as a dimension of Technical Effectiveness rather than as a standalone category. This reflects the view that robustness is best assessed alongside other technical performance attributes—such as accuracy, fidelity preservation, and efficiency—since they are interdependent in determining whether a system can deliver its intended function under real-world conditions.

This organization enables a more granular, technology-driven evaluation, while still mapping each sub-criterion back to the legislative requirements. In this way, the approach ensures both regulatory alignment and a richer understanding of the strengths and limitations of the solutions under review.

5.1 TECHNICAL EFFECTIVENESS

Technical Effectiveness is here understood as the ability of the mechanism to achieve its intended marking or detection objectives under real-world conditions, while maintaining technical performance and output quality.

Accuracy: The degree to which the mechanism correctly marks or detects AI-generated content, minimizing both false positives (incorrectly marking human-generated content) and false negatives (failing to mark AI-generated content).

Robustness: The ability of the mechanism to maintain functionality when subjected to common audio transformations (e.g., compression, noise addition, format conversion) and to resist intentional attempts to remove or alter the mark.

Fidelity Preservation: The extent to which the marking process avoids introducing perceptible degradation or unwanted artifacts, thereby maintaining the original audio quality and characteristics.

Efficiency: The capacity of the mechanism to achieve the intended marking or detection results effectively while optimizing the use of computational, storage, and financial resources, in line with cost-related considerations in Article 50(2) of the AI Act.

5.2 POTENTIAL FOR INTEGRATION

Evaluates how readily the mechanism can be adopted and integrated into existing audio production, distribution, and consumption workflows.

Interoperability: The ability of the mechanism to be compatible across different file formats, platforms, and ecosystems, enabling seamless operation within existing workflows.

Disruptiveness: The degree to which adoption of the mechanism alters current audio production, distribution, and consumption pipelines, including any modifications needed to existing tools or processes.

Accessibility (in this report's operational sense): The ease with which different stakeholders (e.g., creators, platforms, regulators) can adopt and use the mechanism, considering technical skills, infrastructure requirements, and cost barriers. [*Note:* This operational definition differs from the AI Act's legal notion of **accessibility**, which refers to **accessibility for persons with disabilities**, as established in Article 3(7) and Recital 60 of the AI Act. The latter concerns ensuring that AI systems and related interfaces are perceivable, operable, and understandable by users with disabilities.]

5.3 INFORMATION MANAGEMENT CAPABILITIES

Assesses the overall ability of the mechanism to embed, preserve, adapt, and safeguard provenance-related information throughout the content's lifecycle, even as technical conditions evolve or tampering attempts occur.

Information Capacity: The specific quantity and type of provenance or descriptive metadata the mechanism can embed and convey, such as content origin, model used, or generation timestamp.

Resilience to Technical Evolution: The ability of the mechanism to maintain its information-carrying function as AI generation methods evolve and production environments change.

Tamper-detection capability: The ability of the mechanism to detect and flag attempts to remove, alter, or forge the embedded information. While *robustness* concerns the mark's capacity to withstand such interference and remain functional, *tamper-detection capability* focuses on signaling when interference has occurred, even if the mark has been degraded or removed.

5.4 COMPLIANCE AND STANDARDIZATION SUPPORT

Evaluates attributes that facilitate the mechanism's oversight, adoption, and harmonization with legal and industry frameworks.

Understandability: The extent to which the mechanism's principles, design, and limitations can be clearly communicated and evaluated by relevant stakeholders for legal, ethical, and operational oversight.

Standardization Potential: The feasibility of formalizing the mechanism through recognized technical or industry standards to enable broad adoption and regulatory harmonization.

Reliability: The consistency and predictability of the mechanism's performance across time, contexts, and use cases, ensuring that outputs remain trustworthy for all stakeholders.

6. MARKING WITH AUDIO METADATA

As discussed in the State of the art section (Section 4.1), marking audio with metadata involves embedding structured and encoded information into or alongside audio content to describe its characteristics and origin. Metadata may encompass a wide range of contextual information, including the time and place of creation, authorship, and ownership. The specific categories and granularity of metadata used can vary depending on the intended use case, application, and type of audio (i.e., speech, sound, or music). However, defining these categories or evaluating their relative importance is beyond the scope of this report.

Audio metadata is embedded into an audio file through cryptographic techniques (e.g., digital signatures, public-key encryption, hashing). These methods aim to ensure the integrity and reliability of the metadata, preventing potential tampering as the audio is distributed across platforms. Two recent notable examples of these techniques are C2PA (Coalition for Content Provenance and Authenticity) and SEAL (Secure Evidence Attribution Label).

On the one hand, the C2PA specification (C2PA, 2024; NIST AI, 2024) provides a structured system for embedding verifiable provenance data into media assets, including audio, through structured assertions grouped into a cryptographically signed claim. These claims contain metadata describing the content's origin, capture context, editing history, ownership, and chain of provenance, and are bundled into a digitally signed C2PA Manifest using a claim generator. The integrity and authenticity of this manifest are ensured using public-key cryptography, digital certificates, and trust lists, enabling consumers and automated systems to validate the content's provenance throughout its distribution. C2PA supports both embedded and externally referenced metadata. The architecture was designed with scalability, interoperability, privacy, and harm mitigation in mind; however, some external reviews have flagged concerns regarding the complexity of the methodology and its underlying security assumptions. In particular, the robustness of C2PA depends on the security of the cryptographic primitives it employs (e.g., public-key encryption), the trustworthiness of certificate authorities managing keys, and the reliability of trust lists. If any of these assumptions are compromised—for instance, if a certificate authority is breached or cryptographic algorithms become obsolete—the integrity of provenance claims could be undermined. C2PA also supports documenting actions performed by AI systems, including the use of generative models, through dedicated assertions.

On the other hand, the SEAL specification (Krawetz, 2024) is an open solution that ensures media attribution, including audio, by cryptographically signing a file. It aims to provide proper attribution to media, whether the content is original, AI-generated, deep fake, or an altered original. SEAL verifies a file's signature, stored in the metadata, against a public key retrieved via DNS, following an authentication mechanism inspired by DKIM (DomainKeys Identified Mail) (Kucherawy, 2011) used in email authentication. This method allows media files to be distributed with verified authorship, enabling decentralized trust without requiring a central authority. However, SEAL does not verify the semantic integrity or originality of the content, nor does it maintain a record of its historical information or transformation chain. As such, SEAL signatures serve as technical evidence of authorship, but do not prove originality or legal ownership unless supported by legal or social frameworks.

In this section, we focus on evaluating the use of audio metadata marking approaches that rely on cryptographic techniques to enhance the authenticity and traceability of audio content, specifically assessing the C2PA and SEAL technical solutions.

6.1 TECHNICAL EFFECTIVENESS

In the context of marking audio with metadata using cryptographic approaches, such as C2PA and SEAL, **accuracy** does not derive from direct content analysis but from the integrity and reliability of the metadata provided by the signer. Neither mechanism verifies whether the content is from an original human source, has been altered, or is AI-generated. Instead, metadata assertions are bound through cryptographic signatures and depend on the trustworthiness of the individual or entity who signs the embedded metadata. In C2PA, metadata is registered into a claim and digitally signed. While this process confirms the integrity of the manifest and the identity of the signer against trust lists, it does not provide whether the actual data is true or accurate. Similarly, SEAL enables attribution authentication through a cryptographic signature over the file and its metadata, where the signer is expected to vouch for the accuracy

of the metadata upon signature. However, a digital signature only assures that the metadata has not been altered since signing, rather than verifying if the metadata is accurate or the file is authentic or altered from the original content. Thus, for both C2PA and SEAL, the trust in the accuracy of the metadata relies solely on the credibility of the signer, and there is no objective technical metric within these specifications to assess this subjective trust.

In addition, this approach is **robust** in terms of tamper evidence as cryptographic signatures bind the metadata to the audio content, ensuring that any modification to the signed data invalidates the previous signature. However, because the signature is linked to the original file, this mark is lost when audio content is modified by any standard processing, such as compression, noise addition, or format conversion, which alters the underlying audio bytes and thus breaks the cryptographic binding. Hence, C2PA and SEAL are effective for tampering detection, but they are not robust in preserving the mark across transformations or adversarial attempts to remove it.

Furthermore, marking audio with metadata **preserves** content **fidelity**, as techniques such as C2PA and SEAL are non-invasive, attaching cryptographically signed metadata without altering the audio signal itself. Thus, this approach ensures the perceptual quality and characteristics of the original audio remain intact and does not compromise audio fidelity.

In terms of **efficiency**, marking audio with metadata is highly suitable. Even if cryptographic operations such as hashing, signing, and verification may require additional computational resources, these tasks involve relatively simple mathematical operations (e.g., modular exponentiation, hashing over file bytes) that can be executed quickly on modern processors, often in milliseconds per file. Unlike computationally intensive AI training or deep watermarking algorithms, cryptographic metadata signing and verification scale linearly with file size and are well within the capacity of commodity hardware and cloud-based infrastructures. This makes them suitable for large-scale and automated pipelines, such as media platforms handling thousands of uploads per minute, or regulatory compliance checks running in batch or streaming modes. Verification is designed to be lightweight and can be integrated into client-side applications or server-side validation without significant latency. However, metadata storage efficiency depends on the volume of the information itself. While SEAL typically requires a few hundred bytes (less than 500 B), C2PA can reach one or more megabytes for complex provenance records in large files. Overall, the method is computationally efficient and scalable; however, storage costs may vary depending on metadata complexity and the use case.

6.2 POTENTIAL FOR INTEGRATION

Interoperability is one of the core design principles behind both C2PA and SEAL. Both approaches aim to ensure compatibility across different formats, platforms, and ecosystems by relying on open specifications, leveraging standardized metadata embedding techniques, and incorporating cryptographic authentication methods to ensure the integrity of metadata. This provides a shared foundation for verification across tools. However, achieving interoperability in practice remains a challenge. The consistent adoption of marking methodologies may be hindered by the coexistence of multiple approaches, the diversity of metadata requirements across platforms, and the

varying use of protocols and encryption, which can complicate implementation in real-world use cases.

Metadata is well established in the audio domain and widely used to accompany files with relevant information, considering the provenance of the file. Nonetheless, marking audio with metadata using cryptographic-based approaches, such as C2PA and SEAL, introduces a new level of complexity that can be considered **disruptive** to current workflows. These systems require the integration of cryptographic processes beyond the standard audio production and distribution pipelines, which can lead to significant modifications to existing tools and infrastructure. This can be particularly challenging for organizations and individuals that currently lack the appropriate setup to handle this type of authenticated metadata. Thus, while marking with metadata is aligned with existing practices, authenticated metadata marking mechanisms introduce challenges that require consistent support and coordinated implementation. It is worth noting that these disruptions may disproportionately affect small creators and independent labels that lack dedicated technical infrastructure.

Accessibility is a core intention of metadata marking systems, such as C2PA and SEAL. Both approaches aim to foster inclusive and widespread adoption by relying on open standards, supporting open source tools, and providing implementation guidance for multiple stakeholders. However, in practice, the technical complexity of cryptographic processes may pose substantial challenges for some individuals or entities. As a result, despite their open and inclusive aims, metadata marking strategies may be inaccessible to many stakeholders unless broader infrastructure development and sustained support are in place.

6.3 INFORMATION MANAGEMENT CAPABILITIES

Metadata marking approaches offer a high **information capacity**, encoding a diverse, structured, and granular range of data. Frameworks such as C2PA support detailed content information, including per-track assertions, segment-specific data, inputs used in generation (e.g., text prompts), and information about the AI tools or generative models. Alternatively, SEAL was designed to authenticate or attest to other metadata and the content itself, rather than carrying a large amount of detailed provenance information within the SEAL signature itself. In addition, while the system's capacity to record information is high, it is important to note that storage cost may increase with metadata complexity and could become a limiting factor in some use cases (see Section 6.1).

Furthermore, metadata-based marking strategies demonstrate a degree of **resilience to evolving technology** (i.e., AI generation techniques), although such resilience varies across analyzed frameworks. C2PA is designed explicitly to adapt over time, with versioned and extensible specifications that already include AI-relevant features and support generation recording inputs. In contrast, SEAL focuses on cryptographic authentication of existing metadata and content, ensuring integrity across formats without embedding extensive provenance data itself. Thus, its resilience lies in its applicability to any file format, rather than describing AI-specific transformations. Yet, its ability to support upcoming technological evolution depends on the parallel evolution of the metadata formats that it signs.

When metadata is cryptographically embedded into the content, marking systems, such as C2PA and SEAL, can offer a strong **tamper-detection capability** by enabling detection of any unauthorized changes to either the asset or the provenance data. C2PA links the structure assertions to the audio asset through cryptographic hashes. Any modification to the content or the metadata breaks the binding, triggering tamper detection. Instead, SEAL signs the entire file. If the file is altered after signing, signature validation fails, signaling tampering. A shared limitation of both systems is that, when metadata is stored externally or does not cover the whole file, it can be easily removed from the audio without triggering detection. In such instances, validators may be unaware that the content was ever marked, weakening the system's tamper-evidentness.

6.4 COMPLIANCE AND STANDARDIZATION SUPPORT

Marking audio with cryptographically embedded metadata enhances **understandability** by providing comprehensive documentation, enabling source (signer) validation, and allowing for external scrutiny and accountability. C2PA is particularly strong in this category, as it includes openly documented mechanisms that bind structured provenance data to the file and track changes over time. Its layered system allows different stakeholders to access and interpret embedded data at varying levels of detail, supporting informed decision-making. SEAL also contributes to transparency by enabling the verification of content's integrity and the signer's identity through open cryptographic standards. Although it does not maintain a history of edits, its implementation allows for direct audit. However, for both systems, effective transparency depends on how these mechanisms and verification tools are widely accessible and trusted (see Sections 6.1 and 6.2).

In terms of **standardization potential**, marking audio with metadata has a strong profile across regulatory and industry frameworks. Both C2PA and SEAL include essential features for trust, authentication, and accountability. C2PA is supported by major industry stakeholders and is currently progressing toward ISO standardization. It is based on open source tools, includes user guidelines, and its architecture is designed for global adoption. While less mature, SEAL is built on the widely established system DKIM model, and integrates standard file formats and metadata structures, ensuring its compatibility with existing systems. However, to avoid fragmentation and ensure effectiveness, efforts towards standardization must be coordinated across multiple stakeholders. The coexistence of multiple, incompatible standards would undermine the adoption and diminish the regulatory clarity and technical reliability of this approach (see Sections 6.1 and 6.2).

From a **reliability** perspective, metadata-based marking can provide consistent results when provenance information is properly embedded and preserved through standardized formats. However, its reliability is contingent on the persistence of metadata during distribution and on the absence of stripping or alteration, which can occur in uncontrolled environments. In closed or well-governed ecosystems, this approach can achieve highly predictable performance over time.

7. MARKING WITH AUDIO WATERMARKING

Based on the current state of the art (Section 4.2), we have identified two exemplary audio watermarking methods, one based on neural networks and another on traditional audio signal processing. The first method, AudioSeal (San Roman et al., 2024), is a recent neural network–based watermarking method that provides localized watermarking with *minimal perceptual impact on audio quality*. While *minimal* does not mean zero impact, the alterations introduced are designed to be imperceptible to human listeners under typical listening conditions. Perceptual impact is commonly assessed using objective metrics such as the Perceptual Evaluation of Audio Quality (PEAQ) or POLQA, as well as subjective listening tests, where human subjects rate the transparency of watermarked audio compared to the original. AudioSeal is available as open source, demonstrates state of the art robustness against removal attacks, and is computationally relatively efficient. The second method, Audiowmark (Westerfield, 2018), is a popular open source audio watermarking library based on the spectral patchwork watermarking algorithm. This method has been widely adopted as a baseline watermarking approach based on traditional signal processing in recent research on neural network watermarking (Chen et al., 2023; San Roman et al., 2024).

7.1 TECHNICAL EFFECTIVENESS

Accuracy of the example methods is typically measured as a bit error rate (BER), which quantifies the proportion of incorrectly retrieved watermark bits relative to the total number embedded. A low BER corresponds to high bit retrieval accuracy, meaning the embedded information can be recovered reliably. BER is conceptually linked to false negatives, since an elevated BER may prevent the correct recovery of the watermark and lead to the system failing to detect its presence. In contrast, false positives (incorrectly detecting a watermark when one does not exist) are typically measured separately using binary detection metrics. When the objective of the watermark is purely to mark content as AI-generated without embedding extra information, performance is reported in terms of binary detection accuracy: whether the watermark is detected (true positive) or not (true negative). In clean conditions—i.e., when no removal attacks are applied—both Audiowmark and AudioSeal achieve close to 100% accuracy, with near-zero BER and almost no false detections. However, this scenario is not representative of real-world use, since even without malicious attacks, most audio is distributed using lossy codecs that can introduce bit errors.

Robustness is a key aspect in audio watermarking. Traditional signal processing–based methods (e.g., Audiowmark) tend to be more sensitive to audio transformations because they embed the watermark by directly modifying certain spectral or temporal features of the signal. Operations such as resampling, time-stretching, or pitch-shifting alter these low-level features in ways that can distort or erase the embedded pattern, leading to higher bit error rates. For example, Chen et al. (2023) found Audiowmark's BER increased from around 3% in no-attack conditions to roughly 14% under common signal modifications. In contrast, neural network methods can be trained with these transformations included as data augmentations, allowing the model to learn embedding strategies that are invariant to such changes and thus maintain lower error rates. In contrast, this is not straightforward to do with traditional signal processing methods such as Audiowmark. Chen et al. (2023) compared their method WavMark,

directly to Audiowmark and achieved around 0.5% BER under similar attack conditions. Meanwhile, San Roman et al. 2024 found that AudioSeal outperformed WavMark in binary detection accuracy under a wider range of attacks.

Setting robustness requirements requires realistic threat modeling. Simply adding more attacks in the pool and requiring robustness against all of them can negatively affect the fidelity of the watermarked audio. When multiple attacks are present, the model has no knowledge of which attack to expect and must prepare for the worst-case scenario. Developing threat models is an ongoing work, and new vulnerabilities are being discovered. For example, Wen et al. (2025) found that all currently available audio watermarking methods, including AudioSeal and Audiowmark, are sensitive to *physical replay attacks*. In this type of attack, the watermarked audio is played through a loudspeaker in a physical space and then re-recorded with a microphone. This process introduces distortions such as room reverberation, speaker–microphone frequency response characteristics, and ambient noise. These alterations can weaken or destroy the embedded watermark signal, making it undetectable in the re-recorded version even though the audio remains perceptually intact for listeners.

Using watermarking to mark generated audio inevitably has an impact on **fidelity**. This is because watermarking techniques, by design, make slight modifications to the audio waveform to embed information. Even when these changes are engineered to be imperceptible to the average listener, they still alter the original signal in a permanent way. Such alterations can manifest as subtle noise, phase shifts, or changes in frequency balance, which may not always be audible but can reduce the exact fidelity of the original audio. For example, Audiowmark introduces a small but measurable degradation in signal quality, similar in magnitude to the artifacts caused by common lossy codecs such as MP3 or AAC. Importantly, the level of distortion can be controlled: watermarking systems usually include a strength parameter, where stronger embedding increases robustness against removal but makes the distortion more noticeable, while weaker embedding preserves audio transparency at the cost of being easier to erase. Neural methods such as AudioSeal aim to optimize this trade-off by embedding the watermark in less perceptually sensitive parts of the signal, achieving high subjective quality in listening tests. Nonetheless, even in these advanced systems, the watermarked audio can still be distinguished from an unmodified original under critical listening or with objective audio quality measures (e.g., PESQ, STOI, or SDR). San Roman et al. (2024) found that AudioSeal preserves high perceptual quality in subjective listening tests, but watermarked content is still distinguishably degraded from unmodified audio. Recent robustness evaluation has found that both DSP-based and Deep Watermarks could be removed by resynthesis using neural vocoders, codecs, or denoising methods (O'Reilly et al., 2025; Özer et al., 2025).

Computational efficiency is an important consideration. Like most traditional signal processing–based watermarking methods, Audiowmark is relatively lightweight and comparable in complexity to widely used audio coding methods (e.g., MP3 or AAC encoding). The operations typically involve Fourier transforms, filtering, and modulation in the frequency domain, which are well-optimized and require only modest CPU resources. For example, watermark embedding and detection in Audiowmark can typically be performed in real-time on a standard laptop or even a mobile phone without specialized hardware acceleration. In contrast, neural network–based

watermarking methods are computationally more demanding. They rely on deep learning architectures often convolutional or transformer-based, to analyze and modify high-dimensional audio features. Embedding typically requires multiple forward passes through a deep model, which may involve millions of parameters and billions of multiply–accumulate (MAC) operations, similar in scale to the computations needed by state of the art generative audio models themselves. For instance, methods like AudioSeal require GPU acceleration for efficient training and often for real-time embedding as well. While detection can be made lighter, it still involves running a neural network inference pipeline, which may not run in real time on low-power devices without optimization (e.g., model quantization, pruning, or hardware accelerators like NPUs).

7.2 POTENTIAL FOR INTEGRATION

Interoperability of current audio watermarking methods presents many challenges. A worst-case scenario for interoperability would be that every GenAI service provider also provides their proprietary watermark detector, and users would need to run the full array of all possible detectors to check for marked content. Even if detection is provided as a service, users would have to call multiple services on all incoming content.

In a slightly better scenario, a current or near-future method is adopted as a standard. This setting would still be incomplete, since the current methods enable only a limited watermark message to be embedded. For our example methods, Audiowmark uses 128-bit messages, while AudioSeal uses 32 bits per audio segment. The messages are constant over a relatively long period of time, typically a full song in music or one spoken utterance in speech.

With current technology, it would be possible to use short, standardized messages to mark content as generated, but having richer metadata information would require a database for retrieving metadata entries using the watermark messages as keys. To avoid users having to query all possible service providers, database queries would have to be handled in a centralized manner.

Disruptiveness of watermarking is constituted by two main factors. First, implementing watermark embedding and decoding with current technologies is relatively straightforward. Both of our example methods, Audiowmark and AudioSeal, can be applied as post-processing and require no changes to production GenAI systems themselves. Second, a more disruptive feature of watermarking is that the generated content is *permanently modified and cannot be updated*. Furthermore, implementing interoperable systems is likely to cause disruptions, and there is no clear roadmap for interoperable audio watermarking.

Accessibility of the two example systems is good due to their open source availability. Open availability is desired for accessibility but has a downside of making open systems vulnerable to targeted tampering attempts, especially using adversarial attacks on neural network systems (discussed in Section 7.3). Some of this risk can be mitigated by using asymmetric watermarking systems, where the embedding of the mark is private and only the detection tools are publicly available.

7.3 INFORMATION MANAGEMENT CAPABILITIES

Generally, the **Information capacity** of audio watermarking methods is relatively low. For our case study methods, AudioSeal uses a 32-bit message that is repeated throughout the audio segment and detected at every audio sample. Audiowmark uses 128-bit messages that are similarly distributed over several seconds of audio. The current capacity is not sufficient for including metadata directly in audio watermarks, unless the metadata is simply for binary detection of marked content, as legally mandated by Article 50(2) of the AI Act. Rich metadata with audio watermarks would require use of external databases where the watermark messages can be used as retrieval keys. This poses several challenges for interoperability, as detailed in section 7.2.

Resilience to technical evolution is another limitation in audio watermarking methods. Watermarking permanently alters the marked audio content and updating the watermark without further degrading the content is both an open research question and potential attack vector for tampering the watermarks. Some audio watermarking methods utilize invertible neural networks (such as WavMark by Chen et al., 2023) to prevent permanent modification of the content, which somewhat mitigates this issue.

Tampering audio watermarks requires modifying the audio content, and this can cause **evident** degradations in the audio quality. However, current methods only aim to be robust against tampering attempts but have no explicit mechanism for detecting tampering. Not finding a watermark can be due to a removal attack or simply the absence of a watermark in the first place. Furthermore, only removal attacks are typically considered in audio watermarking threat models, leaving out watermark modification and spoofing. Removal attacks can be targeted or non-targeted, while modification is always targeted. The type of removal attacks considered in AudioSeal (San Roman 2024) are non-targeted perturbation attacks. Liu et al. (2024) also consider targeted adversarial attacks against neural network audio watermark models.

7.4 COMPLIANCE AND STANDARDIZATION SUPPORT

Understandability and interpretability of watermarks are highly desirable for legal and ethical oversight. Comparing our two example methods, Audiowmark has interpretable rules for both embedding and detecting the watermark. This is typically true for other signal processing-based audio watermarking methods as well. In contrast, neural network methods are black boxes that work in an end-to-end manner and only output the decoded message. For example, AudioSeal watermark embedding and extraction modules both consist of multiple layers of convolutional and recurrent neural networks, whose operation is non-trivial to interpret.

In the current state of technology, audio watermarking can be used to embed low-bitrate disclosure messages in generated content, and the technology can be made robust to common low-effort and accidental removal attacks. **Standardization** of audio watermarks for the purpose of marking AI-generated content for public disclosure requires further development in the core methodology, as well as a rethinking of interoperability requirements. The classical use case for audio watermarking has been property rights management, where the watermark only needs to be readable by the

watermark embedder, and the receiver only needs to search for their own watermark. Interoperable watermarks for multiple stakeholders require more research and standardization. Another challenge for standardization is the fast-moving nature of both watermarking methods (especially those based on neural networks) and the related threat models for robustness. Furthermore, watermarking typically requires permanent modification of the marked content, and retroactively modifying previous watermarks when technology progresses remains difficult.

Audio watermarking can deliver **reliable** performance when robust embedding and detection algorithms are used, allowing consistent identification across diverse audio content and use cases. Reliability depends on the watermark's ability to survive standard processing and its interoperability with different playback and distribution environments. Properly standardized schemes can maintain predictable results across long-term deployments.

8. IDENTIFICATION WITH AUDIO FINGERPRINTING

As described in the State of the Art (Section 4.3), a prerequisite for content identification with audio fingerprinting is a database of fingerprints extracted from original audio content and stored in an organized manner. Reference examples that work with this principle include Wang (2006), and Sonnleitner and Widmer (2016). When a query is made (an unknown audio clip needs to be identified), the search is realized by comparing the query clip's fingerprints against all the fingerprints in the database. Given the 'create new content on demand' nature of generative models, and the exponentially growing number of audio content, no database can be practically complete. This limits the potential use cases of audio fingerprinting methods for AI-generated content marking and detection. While all experts consulted agree that scalability poses a challenge for using fingerprinting to identify AI-generated audio, there is some debate on whether this could be overcome in the future. On one hand, the existence of music recognition systems that can recognize songs with only a few seconds of audio across datasets of up to a billion songs suggests significant promise. On the other hand, extending such systems to handle the hundreds of billions of expected synthetic audio outputs would introduce strong barriers in engineering, retrieval speed, and operational cost, rendering the approach impractical for real-world applications.

Nonetheless, fingerprinting methods may be applied to add an additional layer of security for sensitive audio content. For instance, by changing the problem definition from 'create a database of AI generated content and mark query content as AI generated when there is a match to this database' to 'create a database of real content and mark query content as AI generated when there is no match to this database', we can find relevant practical applications of using audio fingerprinting to recognize AI generated content. For example, recordings of parliamentary discussions, which are often publicly available, could be fingerprinted. Then, if an audio clip of a parliamentary discussion emerges with its origin or authenticity in question, we could compare its fingerprint with the database of the official fingerprints. This would allow us to determine whether the clip came from the official archive or not. With this approach, instead of identifying AI-generated content, we identify real, recorded content, thus performing inverse verification.

All fingerprinting systems work by comparing a query audio clip to a reference database of authenticated recordings. When a query matches a database entry, it is flagged as positive, meaning the audio is presumed genuine (although an error in the matching process can still yield a false positive). When no match is found, the audio is flagged as negative (i.e., presumably AI-generated). A negative label may stem from imperfect matching or an incomplete database (both cases are false negatives) or from the audio's synthetic origin (a true negative), but the system does not reveal which of these explanations applies. Therefore, the system relies on a robust and distinctive fingerprint extraction algorithm, a reliable matching algorithm, and an extensive database of authenticated audio recordings to function effectively. Next, we analyze the standard fingerprinting systems, such as the ones referenced, using the evaluation criteria outlined in Section 5.

This solution is sometimes referred to as logging-based or information retrieval-based, for instance, in AI-generated text applications (Krishna et al., 2023). In this method, the generator provider maintains a private log or dataset of all content generated by the generative model. A detection tool runs an information retrieval engine on this dataset by using candidate content items as queries. In the case of audio, a fingerprint is logged instead of the audio itself, as it is a much more compact and efficient representation.

8.1 TECHNICAL EFFECTIVENESS

For well-defined and specific use cases (like the one described above), for which we can have a reference database of audio fingerprints, a fingerprinting-based system can achieve high **accuracy**, meaning that it can reliably identify matching content with very low error rates (false positives or false negatives), even under common audio degradations such as compression, noise, or equalization.

Fingerprints of most commercial systems are designed to be robust against common audio degradations and transformations. Therefore, a fingerprinting-based system can exhibit high **robustness**, in the sense that it can still correctly identify audio even after lossy compression, background noise, equalization, or moderate tempo and pitch shifts, which typically degrade the performance of less resilient identification methods. Audio fingerprinting is inherently non-invasive, as it does not introduce alterations to either the query or the reference signal. Hence, the fingerprinting systems exhibit the highest possible **fidelity preservation**.

In terms of **efficiency**, audio fingerprinting for identification depends on (1) creating a database by extracting and storing the fingerprints, (2) extracting fingerprints from the query audio during verification, (3) using a search algorithm to compare the query fingerprints with the database fingerprints. Extraction is relatively efficient, even in commercially available CPUs, meaning that the process of converting an audio clip into a fingerprint can be done quickly (often in real-time or faster) without requiring specialized hardware, such as GPUs or dedicated accelerators. Compared to computationally heavier methods like deep watermarking or model identification, fingerprint extraction involves lightweight operations (e.g., spectral analysis and peak detection) that are well within the capabilities of standard consumer devices. Storage

can be facilitated by using a solid-state disk (SSD) with fast read and write speeds. As the number of database fingerprints increases, searching for matches becomes more difficult. Nevertheless, the success of enterprises such as Shazam, which provide services with a database of millions of songs, suggests that deploying efficient systems based on fingerprinting is feasible. To facilitate large-scale applications, a large storage space is necessary, and cost-effective solutions are available on the market.

8.2 POTENTIAL FOR INTEGRATION

Fingerprint systems are designed to be compatible with a variety of audio formats (.wav, .mp3, .flac, etc.). Moreover, a fingerprinting system can operate across various platforms and ecosystems, as it requires only the audio itself and the accompanying identifier to function. Therefore, the system exhibits high **interoperability**.

The type of usage proposed for taking advantage of fingerprinting does not use or analyze AI-generated audio content; thus, the issue of **disruptiveness** does not apply here.

In terms of **accessibility**, designing robust and scalable fingerprints is an active area of research. In the industry, companies such as Audible Magic offer fee-based fingerprinting services for various use cases, while open source, research-backed solutions are also available online. Different stakeholders could build in-house solutions using open source software (if they are equipped with sufficient technical background) or simply purchase a commercial service.

8.3 INFORMATION MANAGEMENT CAPABILITIES

In terms of **information capacity**, the type of use cases proposed for fingerprinting systems would not give any information related to the AI-generated audio being analyzed; thus, this metric does not apply.
In terms of **resilience to technical evolution**, the system's performance does not depend on how the audio was produced. Its fingerprints capture intrinsic acoustic patterns rather than details of the sound source—be it a human vocal tract, guitar strings, or an AI synthesizer. Consequently, even as deep-fake quality improves, fingerprint matches should remain unaffected, rendering the system highly resilient to technical evolution.

In terms of **tamper-detection capability**, this evaluation criterion is largely irrelevant for audio fingerprinting systems because they do not embed or modify information within the AI-generated audio itself. Instead, fingerprinting operates externally by comparing the raw signal to a reference database. Since no metadata or watermark is bound to the audio file, the system cannot directly signal whether the file has been tampered with. However, a fingerprinting-based system is open to adversarial attacks, i.e., users with malicious intentions can query the system with problematic audio clips. Nonetheless, the disruptive effect such attacks can have on the system is limited.

8.4 COMPLIANCE AND STANDARDIZATION SUPPORT

In terms of **understandability**, because an audio-fingerprinting system assigns every recording a unique digital “fingerprint”—just as each person has a unique physical one—its operating principles are straightforward and can be understood and evaluated by a wide range of stakeholders, even without deep technical expertise.

In terms of **standardization potential,** the widespread use of audio fingerprinting systems across various tasks suggests that it would be feasible to adapt the system to regulatory and industry frameworks.

Audio fingerprinting is generally **reliable** for recognizing content that matches entries in a well-maintained reference database, yielding consistent identification across contexts. Its reliability is limited by the completeness and accuracy of the database, as well as potential mismatches that may occur when content has been significantly altered. Within managed ecosystems, fingerprinting can offer predictable performance over time.

9. IDENTIFICATION OF GENERATIVE AUDIO MODELS

As introduced in the State of the Art (Section 4.4), this forensic methodology concentrates on identifying subtle, often imperceptible artifacts introduced during the audio generation process.

Many contemporary AI music generators, particularly those creating raw audio waveforms, often employ a two-stage architecture. The first stage involves an autoencoder (AE) responsible for the fundamental audio synthesis. This AE learns to compress raw audio into an efficient intermediate representation – such as discrete tokens or latent vectors, using techniques like neural codecs or mel-spectrogram methods. Critically, a decoder component within the AE then reconstructs a listenable audio waveform from this compressed data. The second stage involves an internal generative module, such as a large language model or a similar architecture, which operates on the compressed representation to generate the musical structure or sequence, often based on prompts or learned patterns. While this internal module dictates the musical output, the decoder is responsible for synthesizing the final waveform.

The core hypothesis is that the decoder component invariably leaves behind specific, subtle artifacts in the generated audio signal (Afchar et al., 2025). These artifacts act like unique signatures, characteristic of the specific neural network architecture, training data, and processes used for that decoder. Different decoders, even if they generate the same musical piece from the same intermediate representation, are likely to introduce distinct technical signatures, such as the known checkerboard artifacts associated with certain operations. Specifically, deconvolution modules commonly used in generative models for music generation exhibit systematic frequency artifacts, such as distinctive spectral peaks. These subtle traces can potentially reveal the audio's synthetic origin and even point towards the specific generation model employed.

To train a detector sensitive to these decoder signatures, a carefully designed framework is necessary to avoid learning spurious correlations or confounding factors. For instance, simply comparing datasets of real music and AI-generated music might lead a model to mistakenly identify differences in musical genre, recording quality, or file compression artefacts as signs of artificial generation. Instead, to isolate the decoder artifacts, the methodology involves creating a dataset of genuine, human-made music and generating parallel reconstructions (Afchar et al., 2025). These reconstructions are generated by passing the original real audio through the decoder stages of various known AI models. In this way, the reconstructed samples share the same underlying musical content and use the same file encoding parameters as the originals.

The crucial learning task for the detection model then becomes distinguishing between an original real audio sample and its corresponding reconstructed version produced by a specific decoder. By focusing the detection model on this specific distinction, it is guided to learn the subtle differences introduced solely by the decoding process – the target generation artifact – while ignoring variations in musical content or file format. This approach, therefore, provides a way to identify AI-generated audio based on the technical evidence left by the synthesis process itself. Shallow convolutional neural networks (CNNs) can attain very high accuracy in controlled settings but exhibit limited robustness to simple audio transformations and often generalizes poorly to unseen generative models (Afchar et al., 2025). A detailed Fourier explanation of the deconvolution modules used in generative models and the spectral peaks they introduce allows the development of an interpretable and straightforward detection criterion based on a linear regressor (Afchar et al., 2025b). Other simple detectors were applied to the embeddings of an audio-language model (CLAP), including SVMs, random forests, and K-nearest neighbors (KNN), showing their vulnerability to various audio transformations (Vila et al. 2025). A more sophisticated approach, the SpectTTTra model, leverages long-range musical context through spectro-temporal tokenization of audio, outperforming traditional CNN and Transformer-based models (Rahman et al., 2025).

9.1 TECHNICAL EFFECTIVENESS

The **effectiveness** of CNNs for this specific detection task stems from their inherent ability to identify localized patterns within grid-like data structures with high efficiency. The subtle artefacts introduced by different decoders often manifest as specific textures or patterns within this time-frequency representation. CNNs, through their convolutional filters, excel at learning to detect such spatial hierarchies of features – from simple local patterns in early layers to more complex combinations in deeper layers. This makes them well-suited to recognizing the potentially unique, localized signatures or textures that distinguish a decoder's output from original audio, even when these differences are not obvious to the human ear.

In terms of **accuracy**, specifically for the controlled experimental setup where the task was to differentiate the original audio from its reconstructions generated by known decoders included in the training set, the performance achieved using this artefact-detection approach was remarkably high. Accuracies exceeding 99% were reported, particularly when using amplitude spectrograms as input to relatively basic CNN architectures, indicating that capturing these specific decoder artefacts can be

surprisingly effective under known conditions. However, subsequent analysis revealed significant challenges regarding the detector's generalization capabilities and **robustness**. While demonstrating reasonable generalization within the same decoder family (e.g., successfully identifying reconstructions from different bitrate versions of Encodec when trained on one), the models failed significantly when encountering reconstructions from entirely unseen decoder families (e.g., a model trained only on Encodec artefacts performed poorly on DAC artefacts). This suggests the detectors learn highly specific fingerprints unique to each decoder type. Furthermore, the system demonstrated limited robustness when subjected to various common audio manipulations. These manipulations could represent standard music production steps (like applying equalization or reverb, which had a lesser impact) or deliberate attempts to deceive the system. Performance degraded substantially when confronted with transformations such as pitch shifting, the addition of noise, or re-encoding the audio into different lossy formats (like low-bitrate MP3 or AAC). The experiments presented in Afchar et al. (2025) show that several accuracy scores drop to almost zero for pitch shifting and below 20.0% for noise addition. Besides, models trained to identify a specific decoder failed (with zero accuracy) to generalize to some other decoders. This highlights the detector's sensitivity to processing not encountered during its training and questioning the **fidelity preservation** of the detected artefacts under such transformations.

The performances drop drastically under pitch shifts, the addition of white noise, and re-encoding with different codecs.

9.2 POTENTIAL FOR INTEGRATION

Technically, the core detector, often a standard CNN or other simple model, allows for potential integration into audio processing pipelines. However, its functional **interoperability** is limited by its reliance on being trained specifically against known decoder artefacts. Integrating detection for a new AI model requires retraining or fine-tuning with appropriately reconstructed audio samples, which demands ongoing maintenance and access to diverse generative tools. This reliance on specific training also means it doesn't inherently interoperate well with systems that require a nuanced understanding (e.g., identifying which specific model version was used, beyond the trained families) or with human verification workflows unless supplemented by explainability techniques. The system's **disruptiveness** lies not just in adding an analysis step but also in managing its outputs; the high failure rate on manipulated or unseen audio implies a significant potential for false negatives, allowing modified or novel AI music to bypass detection. Conversely, unexpected inputs may default to a 'real' classification, further complicating trust. The need for continuous updates to counter new generators mirrors the disruptive "cat-and-mouse" dynamic seen in other security fields. Regarding **accessibility**, while the underlying architectures of the detectors are familiar to machine learning practitioners, and the release of code aids researchers, the system's demonstrated fragility makes its practical deployment challenging for platforms or end-users seeking a reliable, standalone solution. Over-reliance on high accuracy scores achieved in controlled settings would be misleading, and robust implementation requires expertise not only in deployment but also in continuous evaluation and adaptation, limiting its broad, simple accessibility.

9.3 INFORMATION MANAGEMENT CAPABILITIES

In terms of information security and adaptability, the artifact-based detection approach presents a mixed profile. Its **information capacity** is primarily limited to identifying whether an audio sample likely originated from one of the specific decoder families it was trained on, distinguishing them from real audio under controlled conditions. It doesn't inherently possess the capacity to provide granular details about the specific model version or parameters used, nor does it reliably quantify the amount of AI generation if AI-generated audio is combined with authentic audio components. This capacity is further diminished by manipulations, making the information it provides brittle. The system exhibits low resilience to technical evolution. As highlighted by the poor generalization to unseen decoder families, the detector's effectiveness is tightly coupled to the specific artefacts learned during training. Even modest modifications within the same encoder/decoder family—such as retraining on a different dataset, adjusting hyperparameters (e.g., learning rate, quantization depth, or codebook size), or fine-tuning decoder weights—can alter the statistical properties of the generated audio enough to weaken or erase the previously learned artefacts. For instance, changing the size of the latent codebook or the regularization applied during training may reduce "checkerboard" patterns or noise floor artefacts that the detector relies upon. Similarly, updates to vocoders or post-processing modules can suppress or reshape artefacts without changing the overall architecture, resulting in a detector trained on an earlier version misclassifying outputs from the updated model. This sensitivity implies that even incremental updates in training or parameterization could render detectors obsolete unless they are continually retrained with up-to-date examples, making long-term robustness difficult to guarantee. Even incremental improvements within a known model family might alter artefacts enough to reduce detection rates. Regarding **tamper-detection capability**, the approach appears highly vulnerable. The significant performance degradation under common audio manipulations, such as pitch shifting or re-encoding, indicates that relatively simple processing, not necessarily requiring sophisticated adversarial techniques, can obscure the learned artifacts and cause detection failure. The system isn't designed to flag the manipulation itself as suspicious; rather, it often defaults to misclassifying the manipulated content, potentially as 'real'. This fragility suggests a low barrier for bypassing detection, making it poorly suited for security applications requiring high confidence against intentional evasion without constant updates and potentially complementary detection methods.

9.4 COMPLIANCE AND STANDARDIZATION SUPPORT

For regulatory purposes, predictable and stable performance is often desired, but this technique operates within a constant "cat-and-mouse" dynamic. As new generative models or variations emerge, the detector requires updates to learn new artefacts, making consistent regulatory compliance difficult to guarantee over time. Furthermore, while the underlying principle relies on identifying decoder artefacts, the models used often lack inherent **understandability**, making it hard to audit why a specific piece was flagged or missed, which is crucial for accountability. The potential for standardization also faces significant challenges. While one could attempt to standardize test datasets or performance reporting for known models, the rapid

evolution of AI generators, including countless "handcrafted" models emerging from the research community and hobbyists, makes creating a comprehensive and lasting standard incredibly difficult. It's practically impossible to track and incorporate artefacts from every potential generator. A proposed method to improve robustness against known audio transformations is data augmentation – explicitly training the detector on audio subjected to pitch shifting, time stretching, noise addition, and other similar manipulations. However, while this may improve resilience to anticipated variations, it doesn't fundamentally solve the generalization problem for entirely unseen models or novel manipulation techniques. Consequently, the lack of robust **standards** and the need for continuous adaptation against a perpetually shifting landscape place a significant burden on anyone attempting to implement and maintain such a system reliably, demanding considerable ongoing effort and expertise.

Generative model identification methods can be **reliable** when model-specific artifacts are stable across outputs and not easily removed by post-processing. Reliability improves when detection algorithms are kept up to date with evolving model architectures. While results may vary with new generation techniques, continuous monitoring and algorithmic updates can help maintain consistent and trustworthy performance.

10. COMPARISON OF TECHNICAL SOLUTIONS

The following table summarizes the comparative evaluation of the different technical solutions discussed in the previous sections, emphasizing the criteria specified in the AI Act.

	Metadata-based marking	**Audio watermarking**	**Audio fingerprinting**	**Generative model identification**
Technical Effectiveness (including Effectiveness & Robustness)	Accurate and reliable in controlled ecosystems; robustness limited by metadata persistence—effectiveness lost if metadata is stripped or altered.	High accuracy when properly embedded; robust to most common transformations but vulnerable to strong compression or re-encoding.	Highly effective for verifying known content in reference databases; robust to minor modifications but fails on substantially altered audio.	Accuracy varies with model and training data; robustness limited by evolution of generation techniques and need for retraining.

Potential for Integration (including Interoperability)	Interoperable where metadata standards (e.g., C2PA, SEAL) are supported; minimal disruption in compliant systems but dependent on standard format adoption.	Interoperability contingent on common embedding/detection standards; moderate integration effort required for production workflows.	Straightforward integration if fingerprint database and infrastructure exist; minimal disruption to current workflows.	Integration depends on availability of detection tools and reference datasets; limited interoperability beyond trained model families.
Information Management Capabilities	High information capacity; can include cryptographic tamper-evidence; resilience limited when metadata can be stripped.	Moderate capacity for provenance data; resilient to common transformations; tamper-evident through failed detection.	Low capacity—identifies only pre-registered content; resilience depends on database completeness; lacks tamper detection.	No embedded data; relies on intrinsic model artefacts; resilience and reliability depend on artefact stability and retraining frequency.
Compliance and Standardization Support (including Reliability)	Reliable in closed or standardized ecosystems; strong potential for formal standardization; highly explainable and transparent.	Reliable when applied through robust and standardized schemes; strong prospects for industry standardization; explainability varies by method.	Reliable for content within comprehensive databases; standardization possible via shared protocols; operationally transparent.	Reliability decreases as models evolve; standardization challenging due to model diversity; explainability remains limited.

As shown in the table, existing technical approaches for marking and detecting AI-generated audio exhibit **complementary strengths and inherent trade-offs** that must be evaluated considering the AI Act's criteria of **effectiveness, interoperability, robustness, and reliability**.

- **Trade-offs between robustness and usability:** Metadata-based marking offers high transparency, interoperability, and information richness but remains fragile in open environments where metadata can be removed. Audio watermarking provides stronger robustness to signal transformations but requires standardized schemes and may degrade under extreme processing. Audio fingerprinting is efficient and reliable for registered content, but cannot

label novel, unregistered AI-generated audio. Generative model identification enables forensic detection without prior marking but is sensitive to post-processing and model drift, requiring continuous retraining.

- **No universal solution:** Each approach performs best under specific operational or regulatory contexts. None currently satisfies all technical and legal requirements simultaneously, particularly in open, heterogeneous distribution ecosystems.
- **Value of hybrid and layered strategies:** Combining complementary methods—such as cryptographically protected metadata for transparency, watermarking for persistence, fingerprinting for database verification, and forensic detection for unmarked content—can enhance **traceability, resilience, and regulatory compliance**. A **multi-layered approach** thus represents the most realistic path forward for ensuring the reliable identification and accountability of AI-generated audio in practice.
- **Compliance readiness:** Among the available approaches, **metadata-based marking** is already *technically feasible* and can be deployed immediately within existing production workflows, provided that interoperable standards (e.g., C2PA) are adopted and integrity is ensured through cryptographic protection. **Audio watermarking** demonstrates strong potential for achieving *reliable and persistent marking*, yet its widespread adoption depends on the establishment of common standards and certification procedures to ensure interoperability and robustness across platforms. **Fingerprinting** remains a *complementary verification mechanism* rather than a marking method and is suitable for controlled domains with reference databases. In contrast, **forensic identification techniques** that detect AI-specific artefacts are still *experimental* and should currently be considered research tools rather than regulatory compliance mechanisms.

11. FUTURE PERSPECTIVES

Building on the shortcomings identified in the previous sections, several **technical, infrastructural, and strategic directions** can be pursued to advance the state of the art in marking, detecting, and identifying AI-generated audio.

- **Hybrid architectures:** Develop multi-layered provenance frameworks that combine complementary mechanisms—cryptographically signed metadata, robust watermarking, fingerprinting, and forensic model artefact detection—to maximize reliability. Such hybrid systems would allow verification even if one layer is removed, degraded, or compromised. For instance, metadata could provide rich provenance and authorship information, watermarking could ensure persistence through transformations, and fingerprinting could serve as a fallback when explicit marks are absent.
- **Robustness through adaptive threat modelling:** Move beyond generic robustness benchmarks by defining realistic, use-case–specific threat scenarios, such as those relevant to public media distribution, archival preservation, or forensic authentication. Tailored benchmarking would ensure that evaluation protocols reflect the most pertinent risks—e.g., recompression, transcoding, noise injection, or speech-to-speech revoicing—and support the development of adaptive defense mechanisms.

- **Interoperability by design:** Promote open and standardized protocols for embedding, extracting, and validating provenance information, ensuring compatibility across the content lifecycle—from creation tools and production workflows to streaming, archiving, and regulatory verification. Alignment with existing initiatives such as C2PA should be reinforced to avoid fragmentation and foster international coherence.
- **Centralized verification infrastructures:** Establish trusted registries or public databases to resolve and validate embedded watermarks, fingerprints, or metadata. Such infrastructures could provide: (1) API-based query services for platforms and regulators to verify provenance at scale, (2) Secure logging of verification requests and results to support audits, (3) Redundancy and mirroring across jurisdictions to ensure availability and resilience, and (4) These could be operated by public institutions, industry consortia, or public–private partnerships, ensuring neutrality and long-term governance.

Nevertheless, addressing these limitations requires more than technical innovation. The following complementary initiatives are also essential:

- **Standardization efforts:** Advance harmonized specifications for cryptographic metadata embedding, watermark encoding parameters, fingerprint extraction formats, and detection APIs. Coordinated standardization—through international bodies and EU initiatives—will reduce fragmentation, facilitate adoption, and ease compliance for smaller or emerging actors.
- **Public benchmarking resources:** Build and maintain open, representative datasets encompassing diverse real and synthetic audio samples annotated with provenance data. These should span multiple content types (music, speech, sound effects), languages, recording conditions, and generative models. Public leaderboards and challenge tasks could incentivize progress in robustness, accuracy, and efficiency while providing transparent performance comparisons.
- **Regulatory sandboxes:** Implement controlled pilot programs, jointly run by regulatory authorities, industry platforms, and research institutions. Sandboxes would allow technical solutions to be tested under real-world operational and legal conditions, helping to identify interoperability issues, compliance barriers, and potential unintended consequences before large-scale deployment.
- **Certification frameworks:** Develop reproducible testing suites and certification mechanisms aligned with Article 50 obligations. These frameworks should define measurable pass/fail criteria for robustness, interoperability, and transparency, and include provisions for periodic re-certification as technologies evolve.

In the longer term, sustained research and innovation efforts will be key to achieving substantial improvements:

- **Updateable watermarking:** Create watermark schemes that allow for controlled renewal, revocation, or replacement without degrading the underlying signal. This would enable the maintenance of provenance data across time without locking content to outdated or compromised marks.
- **Hybrid content provenance:** Create provenance models capable of representing partially AI-generated content combined with human editing. Such systems should clearly delineate which segments were machine-generated and

which were human-created, supporting accountability in hybrid creative workflows.

- **Symbolic and multimodal provenance:** Extend marking and detection methods beyond waveform audio to include symbolic representations (e.g., MIDI, MusicXML) and multimodal outputs (e.g., audiovisual works combining sound, video, and text). This will ensure comprehensive traceability across complex media ecosystems.
- **Explainability and trust:** Improve the interpretability of detection and attribution systems by providing not only binary outputs but also contextual explanations of the evidence (e.g., detected artefacts or signal features). Enhanced explainability will foster user confidence, support legal scrutiny, and strengthen ethical accountability.
- **Decentralized attribution:** Explore blockchain and distributed ledger technologies to verify metadata signatures without relying on a single central authority. Such approaches could improve transparency, resilience, and trust, particularly in cross-border or multi-stakeholder settings.
- **Privacy and data protection:** Develop privacy-preserving mechanisms that comply with fundamental rights while fulfilling transparency obligations. This includes data minimization, privacy-enhancing cryptography, and provenance frameworks that avoid embedding personally identifiable information. Some existing standards for marking media with metadata (such as C2PA and JPEG-Trust) already consider privacy provision to protect information when necessary.

Beyond the technical and regulatory dimensions, it is essential to recognize the **ethical and societal implications** of marking and detecting AI-generated audio. These technologies, if misused or applied without proper safeguards, could lead to **false attribution, reputational harm, or unjust liability** for creators and platforms. Detection errors—both false positives and false negatives—may have significant consequences for freedom of expression, artistic integrity, and media credibility. Furthermore, ensuring **equitable access** to trustworthy marking and detection tools is vital to prevent the concentration of control in a few large technology providers. Ethical deployment therefore requires **transparency, accountability, and inclusiveness** in both the design and governance of provenance technologies, ensuring they serve the broader public interest while supporting cultural diversity and creative freedom.

12. CONCLUSIONS

Marking and identifying AI-generated audio is fundamental to ensuring **transparency, trust, and accountability** in an era of pervasive generative media. Current technical solutions—**metadata**, **watermarking**, **fingerprinting**, and **generative model identification**—each provide valuable but incomplete capabilities. No single approach yet satisfies the full range of **regulatory, technical, and practical requirements** demanded by Article 50 of the AI Act and by real-world deployment scenarios.

Each method presents specific trade-offs:

- **Metadata** offers transparency, non-invasiveness, and straightforward integration into existing workflows, especially when cryptographically protected.

However, it remains fragile in open environments where it can be stripped or altered and depends on consistent adoption of interoperable standards.

- **Watermarking** enables the embedding of persistent and imperceptible identifiers directly in the audio signal, but robustness against deliberate tampering and interoperability across systems remain open challenges.
- **Fingerprinting** is effective for verifying known content within a trusted reference database, yet it cannot identify newly generated or substantially altered audio, limiting its applicability for novel or large-scale generative scenarios.
- **Generative model identification** allows the forensic detection of characteristic AI artefacts without prior marking, but its generalization across models and resilience to transformations are still limited.

Given these trade-offs, **a multilayered and risk-based approach** is the most viable path forward. We recommend combining **cryptographically protected metadata** (for transparency and authenticity) with **robust digital watermarks** (for persistence across transformations). When these identifiers are missing or unreliable, **forensic analysis** of model-specific artefacts should provide a fallback mechanism for detection. **Cross-modal integration**—linking audio with image, video, or text detection pipelines—can further enhance robustness against adversarial manipulation. **Fingerprinting systems** may complement these methods in controlled environments or for specific institutional use cases, despite known scalability constraints.

Ensuring **security against removal or obfuscation** of identifying marks remains one of the greatest technical and legal challenges. We therefore recommend that **tampering with or removing identifying marks** on AI-generated content be explicitly prohibited, in analogy to existing copyright protection frameworks such as the EU Copyright Directive 2001/29/EC. At the same time, it is essential to recognize that unintentional removal—e.g., through standard audio processing or neural codec re-encoding—can also occur, highlighting the need for continued research on **robustness and graceful degradation**.

The long-term reliability and societal acceptance of marking solutions will depend on a **coordinated ecosystem** of open standards, public benchmarking resources, and regulatory guidance. Sustained collaboration among **researchers, industry actors, and policymakers** is crucial to advance technical maturity while safeguarding privacy, artistic integrity, and interoperability.

Ultimately, **technological innovation and regulatory governance must evolve together**. Only through sustained, transparent, and inclusive efforts can we ensure that AI-generated audio remains **identifiable, verifiable, and accountable** throughout its lifecycle—supporting both the creative potential of AI and the public's trust in digital media authenticity.

13. REFERENCES

1. NIST AI 100-4. 2024. “Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency,” National Institute of Standards and Technology, U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-4
2. Hamon, R., Sanchez, I., Fernandez Llorca, D. and Gomez, E. 2024. “Generative AI Transparency: Identification of Machine-Generated Content,” European Commission, Ispra, 2024, JRC137136. https://publications.jrc.ec.europa.eu/repository/handle/JRC137136
3. Kriechbaum, W. 2009. “Audio Metadata,” In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. pp. 157–160. https://doi.org/10.1007/978-0-387-39940-9_1523
4. Riley, J. 2017. Understanding metadata. National Information Standards Organization (http://www.niso.org/publications/press/UnderstandingMetadata.pdf), 23, 7-10.
5. IFPI (International Federation of the Phonographic Industry). 2021. International Standard Recording Code (ISRC) Handbook, 4th Edition. https://www.ifpi.org/wp-content/uploads/2021/02/ISRC_Handbook.pdf, p. 11.
6. Nilsson, M. and Sundstrom, J. 1999. ID3 tag version 2.3.0. https://id3.org/
7. EBU (European Broadcasting Union). 2011. Specification of the Broadcast Wave Format (BWF): A format for audio data files in broadcasting (Version 2.0). https://tech.ebu.ch/files/live/sites/tech/files/shared/tech/tech3285.pdf
8. Qadir, A. M., and Varol, N. 2019. "A Review Paper on Cryptography," 7th International Symposium on Digital Forensics and Security (ISDFS), Barcelos, Portugal, 2019, pp. 1-6, https://doi.org/10.1109/ISDFS.2019.8757514.
9. Hardjono, T., Howard, G., Scace, E., Chowdury, M., Novak, L., Gaudet, M., ... & Vaughan, C. 2019. “Towards an open and scalable music metadata layer,” arXiv, https://doi.org/10.48550/arXiv.1911.08278.
10. Ali, Z., Imran, M., and Alsulaiman, M. 2017. "An Automatic Digital Audio Authentication/Forensics System," in IEEE Access, vol. 5, pp. 2994-3007, 2017, https://doi.org/10.1109/ACCESS.2017.2672681.
11. Yang, X., Wu, X., and Zhang, M. 2009. "Audio Digital Signature Algorithm with Tamper Detection," 2009 Fifth International Conference on Information Assurance and Security, Xi'an, China, 2009, pp. 15-18, https://doi.org/10.1109/IAS.2009.258.
12. C2PA (Coalition for Content Provenance and Authenticity). 2024. “C2PA Technical Specification” (Version 2.1). https://c2pa.org/specifications/specifications/2.1/specs/_attachments/C2PA_Specification.pdf
13. Krawetz, N. and Tucker, B. 2024. “Secure Evidence Attribution Label (SEAL)” (Version 1.01), GitHub. https://github.com/hackerfactor/SEAL
14. Bender, W. Gruhl, D. Morimoto, N. and Lu, A. 1996. "Techniques for data hiding," in IBM Systems Journal, vol. 35, no. 3.4, pp. 313-336, 1996. https://doi.org/10.1147/sj.353.0313
15. Gruhl, D. Lu, A. and Bender, W. 1996. “Echo hiding,” in Information Hiding: First International Workshop Cambridge, UK, May 30–June 1, 1996 Proceedings 1. Springer, 1996, pp. 295–315. https://doi.org/10.1007/3-540-61996-8_48

16. Cox, J. Kilian, J. Leighton, T. and Shamoon, T. 1996. “Secure spread spectrum watermarking for images, audio and video,” in Proc. ICIP, 1996, vol. 3, pp. 243–246. https://doi.org/10.1109/83.650120
17. Cox, I. J., Doerr, G. J., and Furon, T., “Watermarking is not cryptography,” in Digital Watermarking, 5th International Workshop, IWDW, 2006, pp. 1–15. https://doi.org/10.1007/11922841_1
18. Alattar, A., Sharma, R., and Scriven, J. 2020. "A System for Mitigating the Problem of Deepfake News Videos Using Watermarking" in Proc. IS&T Int’l. Symp. on Electronic Imaging: Media Watermarking, Security, and Forensics, 2020, pp. 117-1 - 117-10, https://doi.org/10.2352/ISSN.2470-1173.2020.4.MWSF-117
19. Chen, G., Wu, Y., Liu, S., Liu, T., Du, X., and Wei, F. 2023. “Wavmark: Watermarking for audio generation,” arXiv preprint. https://doi.org/10.48550/arXiv.2308.12770.
20. O’Reilly, Z. Jin, J. Su, and B. Pardo. 2024. “Maskmark: Robust neural watermarking for real and synthetic speech,” in Proc. ICASSP. IEEE 2024, pp. 4650–4654. https://doi.org/10.1109/ICASSP48485.2024.10447253
21. San Roman, R., Fernandez, P., Elsahar, H., Défossez, A., Furon, T., and Tran, T. 2024. “Proactive detection of voice cloning with localized watermarking,” Proc. ICML, 2024. https://dl.acm.org/doi/10.5555/3692070.3693829
22. Singh, M. K., Takahashi, N., Liao, W., and Mitsufuji, Y. 2024. “SilentCipher: Deep audio watermarking,” in Proc. Interspeech, 2024, pp. 2235–2239. https://doi.org/10.21437/Interspeech.2024-174
23. Dathathri, S., See, A., Ghaisas, S. et al. 2024. Scalable watermarking for identifying large language model outputs. Nature 634, 818–823 (2024). https://doi.org/10.1038/s41586-024-08025-4
24. Liu, M. Guo, Z. Jiang, L. Wang, and N. Z. Gong. 2024. “AudioMarkBench: Benchmarking robustness of audio watermarking,” in NeurIPS, Datasets and Benchmarks Track, 2024. https://doi.org/10.48550/arXiv.2406.06979
25. Wen, Yizhu et al. 2025. “SoK: How Robust is Audio Watermarking in Generative AI models?” ArXiv pre-print (2025) https://doi.org/10.48550/arXiv.2503.19176
26. O’Reilly, P., Jin Z., Su J., and Pardo, B. 2025. “Deep audio watermarks are shallow: Limitations of post hoc watermarking techniques for speech,” in ICLR 2025 Workshop on GenAI Watermarking, 2025. https://openreview.net/forum?id=44TCZ5XTuR
27. Özer, Y., Choi, W., Serrà, J., Singh, M.K., Liao, W.-H., Mitsufuji, Y. 2025. A Comprehensive Real-World Assessment of Audio Watermarking Algorithms: Will They Survive Neural Codecs? Proc. Interspeech 2025, 5113-5117, https://doi.org/10.21437/Interspeech.2025-1530
28. Westerfeld, S. 2025. “audiowmark: Robust audio watermarking library,” (2018-2020) https://github.com/swesterfeld/audiowmark, accessed: 2025-04-30.
29. ATSC. 2024. A/334: Audio Watermark Emission. The Broadcast Standards Association. https://www.atsc.org/atsc-documents/a3342016-audio-watermark-emission/
30. SMPTE. 2020. ST 2112-10:2018: Open Binding of Distribution Channel IDs and Timestamps (OBID-TLC). Society of Motion Picture and Television Engineers. https://pub.smpte.org/pub/st2112-20/st2112-20-2020.pdf

31. DCI. 2024. Digital Cinema Initiatives (DCI) Specification Version 1.4.4. Digital Cinema Initiatives, LLC. https://documents.dcimovies.com/DCSS/42cfeb997ae72dd484f7b027547e6e0bfad43ecf
32. Cano, P., Batlle, E., Kalker, T., and Haitsma, J. 2002. “A Review of Algorithms for Audio Fingerprinting,” in IEEE Workshop on Multimedia Signal Processing, 2002. https://doi.org/10.1109/MMSP.2002.1203274
33. Wang, A. 2006. “The Shazam music recognition service,” Commun. ACM, vol. 49, no. 8, pp. 44–48, 2006. https://doi.org/10.1145/1145287.1145312
34. Six, J., and Leman, M. 2014. “Panako - A Scalable Acoustic Fingerprinting System Handling Time-Scale and Pitch Modification,” in ISMIR, 2014. http://hdl.handle.net/1854/LU-5754913
35. Sonnleitner, R., and Widmer, G. 2016. “Robust Quad-Based Audio Fingerprinting,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, (Volume: 24, Issue: 3). https://doi.org/10.1109/TASLP.2015.2509248
36. Agüera y Arcas, B. et al. 2017. “Now Playing: Continuous low-power music recognition,” arXiv. http://arxiv.org/abs/1711.10958.
37. Báez-Suárez, A., Shah, N., Nolazco-Flores, J. A., Huang, S-H. S., Gnawali, O., and Shi, W. 2020. “SAMAF: Sequence-to-sequence Autoencoder Model for Audio Fingerprinting,” ACM Trans. Multimedia Comput. Commun. Appl. 16, 2, Article 43 (May 2020), 23 pages. https://doi.org/10.1145/3380828
38. Singh, A., Demuynck, K., and Arora, V. 2020. “Attention-Based Audio Embeddings For Query-By-Example,” in ISMIR, 2022. https://doi.org/10.48550/arXiv.2210.08624
39. Chang, S. et al. 2021. “Neural Audio Fingerprint for High-Specific Audio Retrieval Based on Contrastive Learning,” in ICASSP, 2021. https://doi.org/10.48550/arXiv.2010.11910
40. Yu, Z., Du, X., Zhu, B., and Ma, Z. 2020. “Contrastive Unsupervised Learning for Audio Fingerprinting,” arXiv. http://arxiv.org/abs/2010.13540
41. Wu, X., and Wang, H. 2022. “Asymmetric Contrastive Learning for Audio Fingerprinting,” IEEE Signal Process. Lett., vol. 29, pp. 1873–1877, 2022. https://doi.org/10.1109/LSP.2022.3201430
42. Wolff, D., Mignot, R., and Roebel, A. 2022. “Audio Defect Detection in Music with Deep Networks,” arXiv. https://doi.org/10.48550/arXiv.2202.05718
43. Li, Y., Sun, Q., Li, H., Specia, L., & Schuller, B. W. 2024. “Detecting machine-generated music with explainability – A challenge and early benchmarks,” arXiv. https://doi.org/10.48550/arXiv.2412.13421
44. Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. 2015. “Spoofing and countermeasures for speaker verification: A survey,” Speech Communication, vol. 66, pp. 130–153, 2015. https://doi.org/10.1016/j.specom.2014.10.005
45. Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., and Lee, K. A. 2023. “ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2507–2522, 2023. https://doi.org/10.1109/TASLP.2023.3285283
46. Yi, J., et al. 2022. “ADD 2022: The first Audio Deep Synthesis Detection Challenge,” in Proc. ICASSP, May 2022, pp. 9216–9220. https://doi.org/10.1109/ICASSP43922.2022.9746939

47. Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., and Larcher, A. 2020. “End-to-end anti-spoofing with RawNet2,” in Proc. ICASSP, 2020, pp. 6369–6373. https://doi.org/10.1109/ICASSP39728.2021.9414234
48. Jung, J. -w. et al. 2022. "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6367-6371, https://doi.org/10.1109/ICASSP43922.2022.9747766
49. Afchar, D., Meseguer-Brocal, G., and Hennequin, R. 2025. “AI-Generated Music Detection and its Challenges,” arXiv. https://doi.org/10.48550/arXiv.2501.10111
50. Afchar, D., Meseguer-Brocal, G., Akesbi, K., and Hennequin, R. 2025. “A Fourier Explanation of AI-music Artifacts,” arXiv. https://doi.org/10.48550/arXiv.2506.19108
51. Vila, L.C., Sturm, B.L.T., Casini, L. and Dalmazzo, D. (2025) “The AI Music Arms Race: On the Detection of AI-Generated Music,” Transactions of the International Society for Music Information Retrieval, 8(1), p. 179–194. https://doi.org/10.5334/tismir.254
52. Rahman, M. A., Hakim, Z. I. A., Sarker, N. H., Paul, B., and Fattah, S. A. “SONICS: Synthetic Or Not - Identifying Counterfeit Songs,” in International Conference on Learning Representations (ICLR), 2025. https://doi.org/10.48550/arXiv.2408.14080
53. Zang, Y., Zhang, Y., Heydari, M., & Duan, Z. 2024. “SingFake: Singing voice deepfake detection,” arXiv. https://doi.org/10.48550/arXiv.2309.07525
54. Sun, C., Jia, S., Hou, S., AlBadawy, E., & Lyu, S. 2023. “Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts,” *arXiv*. https://arxiv.org/abs/2302.09198
55. Kucherawy, M., Crocker, D., & Hansen, T. 2011. “DomainKeys Identified Mail (DKIM) Signatures (RFC 6376)”, RFC Editor. https://doi.org/10.17487/RFC6376
56. Krishna, K., Song, Y, Karpinska, M., Wieting, J., and Iyyer, M. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). https://dl.acm.org/doi/proceedings/10.5555/3666122
57. Rivest, R. L., Shamir, A., & Adleman, L. 1978. A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2), 120-126. https://hdl.handle.net/1721.1/148910

APPENDIX: QUESTIONNAIRE ON THE RELEVANCE OF VARIOUS TECHNICAL APPROACHES TO MARKING AI-GENERATED AUDIO

This questionnaire seeks expert input that can contribute to the writing of a technical report for the EU, which aims to review current technical solutions for marking/identifying AI-generated audio. The report will be publicly available; however, the responses to this questionnaire will remain confidential and will be used solely for drafting the report. Questionnaire participants will be acknowledged in the report unless they explicitly request otherwise (by answering the last item in this questionnaire).

The marking/identification approaches considered are: (1) Marking with Audio **Metadata**: textual information accompanying the audio, (2) Marking with Audio **Watermarking**: embedded information in the audio signal, (3) Identification with Audio **Fingerprinting**: compact representation of an audio signal that can be used to identify the content, (4) **Identification** of Generative Audio Models: audio analysis methodologies that can identify the AI model used in generating the audio signal.

The criteria used to evaluate them are grouped into four categories: (1) technical **effectiveness**, (2) **integration** and practicality, (3) information **security** and adaptability, and (4) **regulatory alignment**.

Please give your opinion on the adequacy of the approaches/technologies considered and answer some questions related to their shortcomings and about possible research directions that could help improve them.

Full name (optional)
Institution (optional)
Email (optional)

Your data will be processed according to the following privacy statement: LEGITIMACY: By consent. You may withdraw your consent at any time. DATA CONTROLLER: Universitat Pompeu Fabra, Barcelona. Email: mtg-info@upf.edu. In accordance with the General Data Protection Regulation, (EU) 2016/679. If you have any doubts, please see the contact details above.

State of the art technical solutions

We have identified four approaches, with example technical solutions within them, that can be used to mark or identify AI-generated audio content. We introduce them here and ask for other specific solutions that you might know about.

Marking with Audio Metadata:

One example of a state of the art technical solution based on metadata technologies is the standard of the Coalition for Content Provenance and Authenticity (C2PA, https://c2pa.org/) an open technical specification designed to embed cryptographic

signatures and detail data provenance into the metadata of audio files to verify the origin and integrity of audio content.

Do you know of other state of the art technical solutions based on metadata technologies for marking audio content? Please give their reference here.

Marking with Audio Watermarking:

One example of a state of the art technical solution based on audio watermarking technologies is the Standard A/334: Audio Watermark Emission of the Advanced Television Systems Committee (https://www.atsc.org/atsc-documents/a3342016-audio-watermark-emission/). It is based on digital signal processing techniques, modifying the least-significant bit in sub-band autocorrelation signals. Recently, there has been increased research activity in deep-learning-based audio watermarking techniques, such as using multiplicative spectrogram masks (O'Reilly, Z. Jin, J. Su, and B. Pardo. 2024. "Maskmark: Robust neural watermarking for real and synthetic speech," in Proc. ICASSP. IEEE 2024, pp. 4650–4654. https://doi.org/10.1109/ICASSP48485.2024.10447253).

Do you know of other state of the art technical solutions based on watermarking technologies for marking audio content? Please give their reference here.

Identification with Audio Fingerprinting:

One example of a state of the art technical solution based on audio fingerprinting for identifying audio content is landmark-based methods, such as Panako (Six, J., and Leman, M. 2014. "Panako - A Scalable Acoustic Fingerprinting System Handling Time-Scale and Pitch Modification," in 15th International Society for Music Information Retrieval Conference (ISMIR), 2014. http://hdl.handle.net/1854/LU-5754913) or QuadFP (Sonnleitner, R., and Widmer, G. 2016. "Robust Quad-Based Audio Fingerprinting," IEEE/ACM Transactions on Audio, Speech, and Language Processing, (Volume: 24, Issue: 3). https://doi.org/10.1109/TASLP.2015.2509248), which detect energy peaks within a signal's time-frequency representation.

Do you know of other state of the art technical solutions based on fingerprinting for identifying audio content? Please give their reference here.

Identification of Generative Audio Models:

One example of a state of the art technical solution for identifying generative audio models is based on detecting specific artifacts they introduce, such as the inherent characteristics of neural decoders (Sun, C., Jia, S., Hou, S., AlBadawy, E., & Lyu, S. 2023. "Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts," arXiv. https://arxiv.org/abs/2302.09198; Afchar, D., Meseguer-Brocal, G., and Hennequin, R. 2025. "AI-Generated Music Detection and its Challenges," arXiv. https://doi.org/10.48550/arXiv.2501.10111).

Do you know of other state of the art technical solutions based on identification of audio models technologies for identifying audio content? Please give their reference here.

Technical effectiveness

This refers to the ability to reliably embed, detect, and preserve marks without significantly altering the original audio.

1. What is your opinion on the technical effectiveness of using **Metadata** approaches/technologies, together with **Cryptographic** techniques, to mark audio files:

Accurate (effective in identifying AI-generated content):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Robust (able to withstand various content transformations):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Preserves fidelity (does not introduce perceptible content degradation):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Efficient (efficient in terms of computation and storage):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

2. What is your opinion on the technical effectiveness of using **Watermarking** approaches/technologies to mark audio files:

Accurate (effective in identifying AI-generated content):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Robust (able to withstand various content transformations):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Preserves fidelity (does not introduce perceptible content degradation):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Efficient (efficient in terms of computation and storage):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

3. What is your opinion on the technical effectiveness of using Audio **Fingerprinting** approaches/technologies to identify AI-generated audio files:

Accurate (effective in identifying AI-generated content):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Robust (able to withstand various content transformations):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Preserves fidelity (does not introduce perceptible content degradation):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Efficient (efficient in terms of computation and storage):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

4. What is your opinion on the technical effectiveness of using **Identification of Generative Audio** approaches/technologies to identify AI-generated audio files:

Accurate (effective in identifying AI-generated content):

Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Robust (able to withstand various content transformations):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Preserves fidelity (does not introduce perceptible content degradation):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Efficient (efficient in terms of computation and storage):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

Potential for integration and practicality

This refers to the potential for integration into existing ecosystems, workflows, and user needs.

1. What is your opinion on the integration potential and practicality of using **Metadata** approaches/technologies, together with **Cryptographic** techniques, to mark audio files:

Interoperable (compatible across platforms):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Non-disruptive (no impact on current production and distribution pipelines):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Accessible (stakeholders can easily implement and use it):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

2. What is your opinion on the integration potential and practicality of using **Watermarking** approaches/technologies to mark audio files:

Interoperable (compatible across platforms):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Non-disruptive (no impact on current production and distribution pipelines):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Accessible (stakeholders can easily implement and use it):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

3. What is your opinion on the integration potential and practicality of using Audio **Fingerprinting** approaches/technologies to identify AI-generated audio files:

Interoperable (compatible across platforms):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Non-disruptive (no impact on current production and distribution pipelines):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Accessible (stakeholders can easily implement and use it):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

4. What is your opinion on the integration potential and practicality of using

Identification of Generative Audio approaches/technologies to identify AI-generated audio files:

Interoperable (compatible across platforms):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Non-disruptive (no impact on current production and distribution pipelines):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Accessible (stakeholders can easily implement and use it):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

Information security and adaptability

This refers to the ability to effectively convey and manage relevant information while maintaining functionality under different conditions.

1. What is your opinion on the Information security and adaptability of using **Metadata** approaches/technologies, together with **Cryptographic** techniques, to mark audio files:

Sufficient information capacity (can convey enough amount and type of metadata):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Resilient to technical evolution (can adapt to evolving AI generation techniques):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Tamper-resistant (can detect instances where the marking has been removed or altered):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

2. What is your opinion on the Information security and adaptability of using **Watermarking** approaches/technologies to mark audio files:

Sufficient information capacity (can convey enough amount and type of metadata):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Resilient to technical evolution (can adapt to evolving AI generation techniques):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Tamper-resistant (can detect instances where the marking has been removed or altered):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

3. What is your opinion on the Information security and adaptability of using Audio **Fingerprinting** approaches/technologies to identify AI-generated audio files:

Sufficient information capacity (can convey enough amount and type of metadata):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Resilient to technical evolution (can adapt to evolving AI generation techniques):

Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Tamper-resistant (can detect instances where the marking has been removed or altered):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

4. What is your opinion on the Information security and adaptability of using **Identification of Generative Audio** approaches/technologies to identify AI-generated audio files:

Sufficient information capacity (can convey enough amount and type of metadata):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Resilient to technical evolution (can adapt to evolving AI generation techniques):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Tamper-resistant (can detect instances where the marking has been removed or altered):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

Regulatory alignment

This refers to the ability to operate within legal, ethical and regulatory boundaries.

1. What is your opinion on the regulatory alignment potential of using **Metadata** approaches/technologies, together with **Cryptographic** techniques, to mark audio files:

Transparent (can be understood and evaluated by stakeholders):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Can be standardized (can be standardized for widespread adoption):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

2. What is your opinion on the regulatory alignment potential of using **Watermarking** approaches/technologies to mark audio files:

Transparent (can be understood and evaluated by stakeholders):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Can be standardized (can be standardized for widespread adoption):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

3. What is your opinion on the regulatory alignment potential of using Audio **Fingerprinting** approaches/technologies to identify AI-generated audio files:

Transparent (can be understood and evaluated by stakeholders):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Can be standardized (can be standardized for widespread adoption):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree

Open comment and justification:

4. What is your opinion on the regulatory alignment potential of using **Identification of Generative Audio** approaches/technologies to identify AI-generated audio files:

Transparent (can be understood and evaluated by stakeholders):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Can be standardized (can be standardized for widespread adoption):
Move the slider or accept the initial position. Strongly disagree <-> Strongly agree
Open comment and justification:

Other questions

Do you see any major shortcomings in the use of the identified technologies for the task of marking or identifying AI-generated audio signals?

Can you identify any use cases in which the proposed technologies would have major problems or would not work?

Would you propose any other approach to mark or identify AI-generated audio signals?

Using the report being written, the EU will draft Codes of Practice for the transparency obligations specified in Article 50 (2), (4) and (5) of the AI Act. What do you think these Codes of Practice should, or should not, include?

What areas of research should be further promoted in order to improve the technological approaches to support the marking or identification of AI-generated content?

What role do you think decentralized technologies (e.g., blockchain, distributed ledgers) could play in verifying AI-generated audio?

Provide any final comment related to this questionnaire:

Do you want your name to be referenced in the report?


Publications Office
of the European Union