DEEPMARK BENCHMARK: REDEFINING AUDIO WATERMARKING ROBUSTNESS

Slavko Kovačević & Murilo Z. Silvestre
DeepMark
{slavko, murilo}@deepmark.me

Kosta Pavlović
Computer Science Center
Faculty of Science and Mathematics
University of Montenegro
kosta@ucg.ac.me

Petar Nedić
Montenegrin Academy of Science and Arts
petarnedic@ucg.ac.me

Igor Djurović
Montenegrin Academy of Science and Arts
Faculty of Electrical Engineering
University of Montenegro
igordj@ucg.ac.me

ABSTRACT

This paper introduces DeepMark Benchmark, a novel and comprehensive framework for evaluating the robustness of audio watermarking algorithms. Designed with modularity and scalability in mind, the benchmark enables systematic testing of watermarking methods against a diverse set of attacks. These include basic audio editing operations, advanced desynchronization techniques, and deep learning-based attacks that leverage generative models and neural processing methods. Additionally, we introduce a new class of attacks, termed Process Disruption Attacks, which target generative AI (GenAI) platforms. These attacks do not rely on prior knowledge of the system's architecture or signal processing methods and can arise inadvertently within the GenAI workflows. The code is available at https://github.com/deepmarkpy/deepmarkpy-benchmark.

1 INTRODUCTION

The exponential growth of digital data and the sophistication of AI-generated content have created significant security challenges. Deepfakes, in particular, undermine public trust and pose risks to personal security, democratic processes, and societal stability Chesney & Citron (2019). This is especially true in critical areas such as elections Agarwal et al. (2019) and information dissemination Farid (2022), where the misuse of GenAI platforms can cause widespread harm.

Rapid advances in AI leave gaps in the detection and prevention of malicious content spread, as existing solutions struggle to keep pace Mirsky & Lee (2021). Watermarking has emerged as a promising approach to mitigate these risks by enabling clear identification of the GenAI company and the user responsible for generating malicious content. Robust solutions have become increasingly critical with regulatory frameworks such as the EU AI Act European Union (2024), mandating watermarking on synthetic content.

This paper introduces DeepMark Benchmark, a modular and scalable platform designed to rigorously evaluate the robustness of audio watermarking systems. As AI-generated content evolves, conventional watermarking methods have become increasingly vulnerable, often failing against modern techniques such as AI-based audio editing and enhancement. The ability to withstand such sophisticated attacks is now critical for watermarking to remain viable in combating malicious content.

DeepMark Benchmark is designed to evaluate the robustness of watermarking against traditional and modern threats, including desynchronization attacks, common audio modifications, and adversarial methods specifically targeting GenAI workflows. The benchmark incorporates a novel class of attacks, termed Process Disruption Attacks, which are uniquely tailored to exploit vulnerabilities in GenAI platforms without requiring prior knowledge of the watermarking system. By focusing

on robustness as the cornerstone of watermarking evaluation, DeepMark Benchmark provides a comprehensive and standardized framework for testing and comparing algorithms.

The goal of DeepMark Benchmark is to increase transparency in watermarking evaluation, enabling GenAI companies to identify and test solutions tailored to their needs. It provides a trusted, standardized platform for verifying watermarking methods by third parties. With publicly available source code, DeepMark serves as a baseline for future expansions and can be extended with new attack types to address evolving challenges.

DeepMark is the successor to StirMark Steinebach et al. (2001), a foundational audio watermarking benchmark that posed a significant challenge to systems in the past. However, the emergence of deep learning-based watermarking techniques Pavlović et al. (2022); Chen et al. (2023); San Roman et al. (2024); Singh et al. (2024) has proven effective against most attacks within the StirMark, prompting the development of this new benchmark. Meanwhile, new technologies and attack strategies have been developed, providing additional methods to compromise watermarking systems, further motivating the need for a new benchmark.

The results of benchmarking novel approaches with DeepMark Benchmark highlight that the rapid evolution of generative models and adversarial attacks has exposed vulnerabilities in existing watermarking methods, demonstrating that incremental improvements are insufficient. To remain viable, watermarking systems should not merely adapt to current deep learning advances but instead drive a revolution in the field by inventing new architectural paradigms and training strategies.

2 ATTACKS

The attacks presented in this section adhere to well-defined conditions that ensure their practical relevance. First, an attack must preserve the integrity of the signal carrier, as destroying the signal renders the information useless and defeats the purpose of the attack. Second, attacks should introduce minimal audible artifacts to maintain the perceptual quality of the signal. Obvious distortions or tampering artifacts would compromise the effectiveness of the attack by signaling clear evidence of manipulation. These principles ensure that the evaluated watermarking systems are tested under both realistic and challenging conditions.

2.1 PROCESS DISRUPTION ATTACKS

Process Disruption Attacks refer to a class of attacks that exploit GenAI platforms to interfere with the core mechanisms of watermarking systems, disrupting their ability to embed or detect watermarks. These attacks do not require prior knowledge of AI, signal processing, or the underlying system and can even occur unintentionally, as they result from the misuse or normal usage of GenAI tools.

A representative scenario for a process disruption attack arises when a generative AI company produces synthetic audio content embedded with a watermark and provides it to downstream integrators via an API. If an integrator further applies its own watermarking, using the same model architecture, the original watermark may be entirely overwritten or erased.

- Same Model Watermarking (SMW): The signal is first embedded with one watermark and then re-embedded with a second watermark, effectively layering the watermarks. The goal is to evaluate whether the watermarking system is robust enough to detect and retrieve the first watermark after subsequent embeddings, which tests the system's resistance to overwriting and interference from repeated use of the same model.
- Cross Model Watermarking (CMW): The signal is first embedded with one watermark using a specific model and then re-embedded with another watermark using a different model. The primary objective is to assess whether the first watermark can still be detected after the second embedding. This tests the interoperability and robustness of watermarking systems against interference from different embedding techniques.
- Collusion Attack (CA): Involves independently embedding two different watermarks into the same original signal, resulting in two separately watermarked versions. These watermarked signals are then merged in a way that cancels out or disrupts both watermarks,

effectively removing them from the signal. This attack tests the resilience of watermarking systems against coordinated attempts to erase watermarks through signal combination and manipulation. The size of the collusion segment ranges from 5 to 100 samples.

2.2 AUDIO EDITING ATTACKS

These attacks refer to standard audio editing operations that are not inherently malicious but can unintentionally or subtly remove watermarks. StirMark introduced many such attacks, and numerous variations exist. However, we have selected a subset of these attacks based on empirical evidence from our experiments, focusing on those that remain consistently damaging to modern watermarking models. Although widely used in routine audio processing, these operations can disrupt or erase embedded watermarks, posing a significant challenge to the robustness of watermarking systems in the real-world usage scenarios.

- Clip Segments (CS): Randomly removes short subsequences from an audio signal. It uses several parameters, a maximum clip sequence length (default: 50 samples), a number of clip sequences (default: 20), clip duration (default: 0.5 seconds), and maximum clip value difference (default: 0.1).

  The attack begins by randomly selecting a segment of the audio signal based on the specified clip duration. Within this segment, it randomly picks multiple positions for potential clipping. As it processes the segment, it removes sequences of samples, up to the maximum allowed length, provided that the difference between the start and end sample values remains within the defined threshold.

- Wavelet Denoising (WD): Applying wavelet-based denoising techniques to the audio signal removes subtle variations that may carry the embedded watermark. It computes the universal threshold which is particularly effective for denoising signals corrupted by additive white Gaussian noise.

2.3 DESYNCHRONIZATION ATTACKS

Desynchronization Attacks involve altering the timing or structure of an audio signal to disrupt the embedded watermark. These attacks often change the length of the signal, which poses a significant challenge to many deep learning architectures that require fixed-length inputs. There have been attempts to develop DNN-based watermarking systems that are resistant to desynchronization attacks Pavlović et al. (2023). However, these approaches often come with significant trade-offs in other performance criteria, particularly in terms of capacity.

- Time Stretch (TS): Altering the playback speed of an audio signal by either slowing it down or speeding it up, without affecting its pitch. This transformation is applied with a time-stretch parameter value selected from a range that ensures perceptual quality is preserved (default: 1.4).

- Pitch Shift (PS): Altering the pitch of an audio signal by raising or lowering its frequency without changing its length. Pitch shifting is controlled by the pitch shift parameter in cents, with 1 cent equivalent to $1/100$ of a semitone (default: 5 cents). Although pitch shift in this specific implementation might not strictly qualify as a desynchronization attack, it is typically grouped with time stretching approaches, prompting us to include it in the same category.

- Inverted Time Stretch (ITS): This method first applies time stretching to slow down or speed up the audio, then performs second, inverse time stretch (e.g., a factor of 2 followed by 0.5) to restore the original duration. Both stretches are limited to a controlled range to minimize distortion. The intensity of the effect is determined by the time stretch parameter (default: 2.0).

- Zero Cross Inserts (ZCI): Introduces pauses into an audio signal by inserting short sequences of zeros at zero-crossing positions where the signal transitions between positive and negative values. The length of each pause is controlled by an assigned parameter (default: 20 samples), while the minimum distance between consecutive pauses is determined

by another (default: 1.0 seconds). With pauses spaced apart and occurring only at zero-crossing points, this attack minimizes perceptual distortions while effectively tampering with the embedded watermark.

- Flip Samples (FS): Disrupts the embedded watermark by randomly exchanging the positions of selected sample pairs within the audio signal. Two separate parameters guide this operation: the first (default: 100) determines how many pairs are flipped, while the second (default: 0.5 seconds) specifies the time window in which those flips occur. By ensuring that this attack introduces subtle rearrangements, the signal structure is effectively altered without causing noticeable distortions.
- Replacement Attack (RA): This attack exploits the observation that audio content is often highly repetitive. It was proposed in Kirovski et al. (2007) and works by replacing each signal block with another perceptually similar block sourced from the same signal.

  The attack is guided by several parameters, including the block size used for processing (default: 1024), the overlap factor between consecutive blocks (default: 0.75), and the similarity distance range that determines whether a block qualifies as a candidate for replacement (default: 0 to 10 dB). The replacement block is computed as a linear combination of up to K similar blocks (default: 30), with coefficients obtained through the least squares approximation.

  Previous studies Hua et al. (2016) have demonstrated that this attack is highly effective against watermarking systems based on conventional signal processing techniques. Since it was not originally included in the StirMark benchmark, we incorporate it here.

2.4 AI Attacks

AI-driven attacks exploit advanced machine learning or deep leaning techniques to compromise watermarking systems. Unlike conventional attacks, these methods leverage AI's adaptive capabilities to subtly disrupt the embedded watermark. These attacks present a significant challenge by simulating real-world scenarios where AI-driven tools, either deliberately or inadvertently, modify watermarked content, making robust detection increasingly complex.

- Speech Enhancement (SE): Recent advancements in speech enhancement (denoising) have been driven by deep neural network (DNN) techniques. These DNN-based models have demonstrated state-of-the-art performance in improving the perceptual quality and intelligibility of noise-contaminated speech signals. Their powerful denoising capabilities pose a challenge for watermarking systems, as they may alter or erase embedded watermarks. The attack implemented within this benchmark applies speech enhancement using a SepFormer architecture Subakan et al. (2020), trained on a dataset sampled at 16 kHz, with the SpeechBrain toolkit Ravanelli et al. (2021). Before enhancement, Gaussian noise is added to the signal to adjust the signal-to-noise ratio (SNR) to approximately 15 dB, after which the enhancement process is applied.
- Speech Tokenization (ST): This attack utilizes X-Codec Ye et al. (2024) speech tokenizer to process watermarked speech, transforming it into a compressed token representation and subsequently reconstructing it back into waveform format. Given that X-Codec is optimized for semantic preservation rather than strict waveform fidelity, this process can introduce distortions that may challenge watermark robustness. The model operates with a codebook of 65536 tokens at a tokenization rate of 50 tokens per second. It uses Wav2Vec2-BERT Barrault et al. (2023) semantic encoder which supports multilingual speech, making it a relevant attack scenario for evaluating watermark resilience across diverse speech data.
- Neural Vocoder (NV): This attack reconstructs raw waveforms from high-level acoustic representations, such as mel spectrograms, leveraging state-of-the-art vocoder networks. Within this benchmark, we include the BigVGAN vocoder Lee et al. (2022), which builds upon HiFi-GAN Kong et al. (2020) as its baseline generator architecture.

  Mel spectrograms are widely used in speech processing as they transform raw audio into a time-frequency representation, mimicking human auditory perception. However, this transformation inherently discards certain fine-grained details, particularly phase information and subtle signal variations. When a neural vocoder, such as BigVGAN, reconstructs

the waveform from a mel spectrogram, it must regenerate these missing details, often introducing artifacts or deviations from the original signal. For watermarking systems, this transformation poses a severe challenge. Converting a watermarked signal to a mel spectrogram and back can alter, degrade, or completely remove the watermark, depending on how the vocoder reconstructs the missing details. This makes neural vocoder attacks an effective method for testing the robustness of watermarking techniques, particularly against real-world scenarios where AI-based speech synthesis tools are increasingly deployed.

- Variational Autoencoder (VAE): This attack exploits RAVE Caillon & Esling (2021), a neural audio synthesizer to compress and reconstruct audio. During the attack, the watermarked audio is first encoded into a lower-dimensional latent space through the encoder network. The latent space representation is obtained through sampling, which introduces stochastic noise that can further alter the encoded signal. The sampled representation is then decoded back to the audio domain using the decoder, which is trained to preserve perceptual quality. This attack can be particularly effective as the encoding-decoding process itself may distort the watermark, with the added effect of RAVE's training objective prioritizing perceptual features over exact waveform reconstruction.

- Diffusion-based Audio Attack (DFA): Leverages diffusion models to reconstruct audio signals, while introducing subtle transformations that can challenge watermark robustness. Diffusion-based generative models have recently demonstrated state-of-the-art performance in high-fidelity audio synthesis and provide a powerful framework for denoising and generative tasks.

  For this benchmark, we include two versions of the attack: DDPM (Denoising Diffusion Probabilistic Models) Ho et al. (2020) and DDIM (Denoising Diffusion Implicit Models) Song et al. (2020), using models from the Audio Diffusion library Smith (2023). These models operate on mel spectrograms rather than raw waveforms. The reconstructed spectrograms are then converted back into waveforms using a neural vocoder. We opted for direct mel spectrogram diffusion rather than performing a diffusion process in the latent space, as our experiments showed that latent diffusion models excessively alter the signal, failing to meet the criteria for information preservation.

  To ensure that the generated signals remain within the prescribed limits of information preservation, we reduce the number of diffusion steps by 90% compared to typical values. Specifically, instead of the default 1000 steps for DDPM and 50 steps for DDIM, we use 100 steps for DDPM and 5 steps for DDIM. This adjustment prevents excessive alterations to the signal, which would render the transformation unrealistic for watermark robustness testing.

3 Benchmark Architecture

Designed with a modular and extensible architecture, the benchmark enables seamless integration of new models and attack implementations by a broad user base, thereby supporting standardized evaluation across different approaches.

Benchmark architecture is shown in Figure 1.

The core component of the framework is its plugin manager, which dynamically imports and manages attacks and models. This eliminates manual configuration, allowing users to test their methods with minimal effort. Due to the complexity of dependencies—particularly for AI-based attacks—the framework leverages Docker to ensure environment consistency and reproducibility. While certain models and attacks can be executed natively, others require isolated containerized environments to manage intricate dependencies effectively.

This dual execution strategy ensures that the benchmark remains scalable and supports a wide range of adversarial techniques while maintaining ease of integration. By abstracting model and attack interfaces, the framework provides a consistent and automated evaluation pipeline, enabling direct comparisons between different watermarking methods.

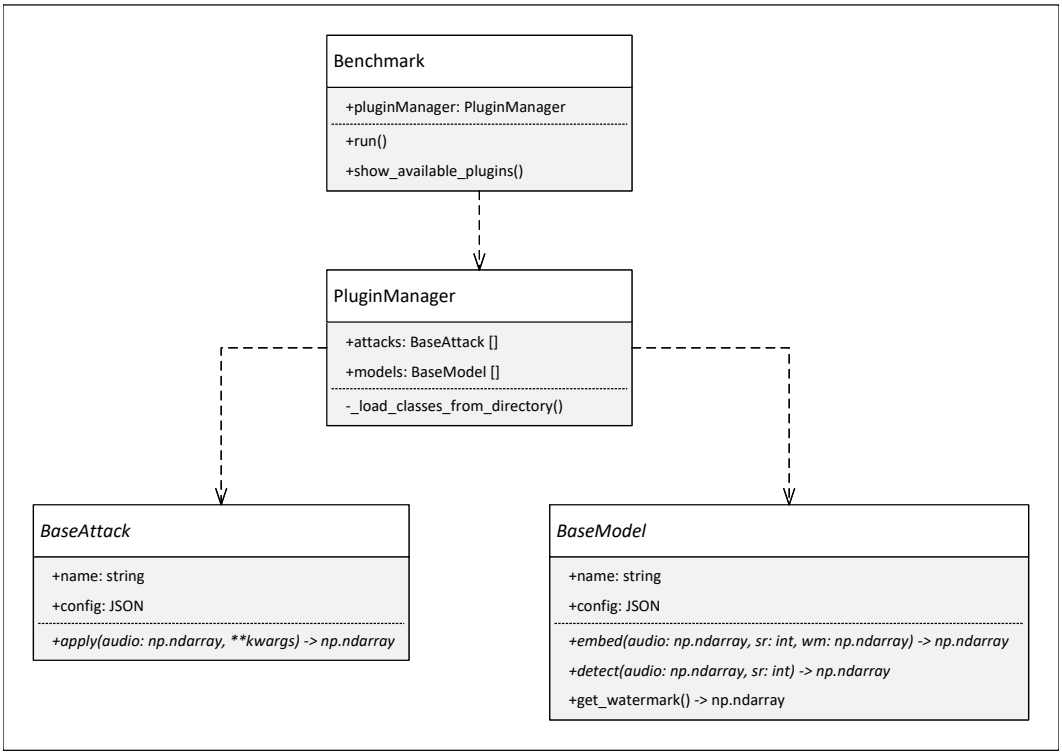

Figure 1: UML diagram of the benchmark architecture.

Table 1: Robustness evaluation using DeepMark benchmark. A '+' symbol indicates successful watermark detection, while a '-' symbol denotes failure.

Attack	Models		
	AudioSeal	WavMark	SilentCipher
SMW	-	-	-
CMW	+	+	+
CA	-	-	-
CS	+	+	+
WD	+	-	-
TS	+	-	-
PS	-	-	-
ITS	+	-	-
ZCI	-	+	+
FS	+	+	+
RA	+	+	+
SE	-	-	-
ST	-	-	-
NV	-	-	-
VAE	-	-	-
DFA	-	-	-

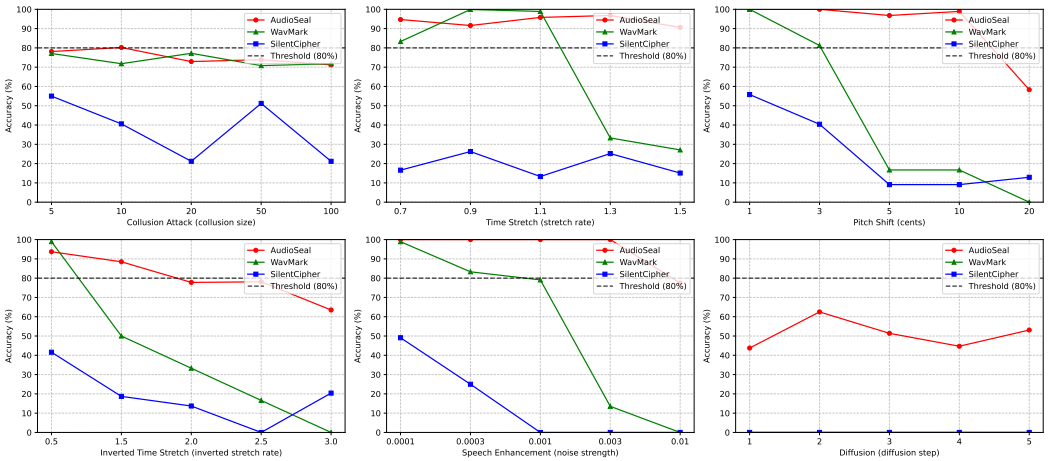

Figure 2: Comparison of AudioSeal, WavMark, and SilentCipher model accuracies when subjected to various parameterized audio attacks. The dashed line represents the threshold accuracy of 80%.

4 Benchmarking and Results

We evaluated several state-of-the-art watermarking models using our proposed benchmarking framework, including AudioSeal San Roman et al. (2024), WavMark Chen et al. (2023), and SilentCipher Singh et al. (2024). In our selection, we prioritized models with publicly available open-source implementations, as they allow for reproducible and transparent evaluation. Benchmarking results are presented in Table 1, with key findings summarized below. A model is considered non-robust against a specific attack if its Bit Error Rate (BER) exceeds 20%, indicating a substantial loss of watermark integrity beyond an acceptable threshold.

Our tests were performed on a VCTK Corpus Veaux et al. (2017).

Process disruption attacks pose a significant challenge for all tested models, except for the Cross Model Watermarking attack. This suggests that each model employs a distinct watermarking strategy, either embedding the watermark in different spectral regions or utilizing fundamentally different encoding mechanisms. Conversely, the Same Model Watermarking attack exhibits significantly greater disruption, as expected, likely due to interference between overlapping watermark patterns. However, despite this disruption, the second embedded watermark remains recoverable across all models, suggesting a degree of robustness in layered watermarking schemes.

Wavelet denoising demonstrates potential as a means of improving compression robustness. Given its impact on watermark integrity, we argue that future watermarking models should be explicitly designed to withstand wavelet-based transformations to ensure resilience against lossy compression techniques.

Even subtle pitch shifting leads to severe degradation of the embedded watermark, indicating that this attack is highly effective. Given its complexity, achieving robustness against pitch shifting would likely necessitate a fundamental shift in architectural design.

Zero-crossing insertion severely compromises AudioSeal. WavMark and SilentCipher exhibit greater resilience to this attack, likely due to their robust handling of silence and zero-value samples during watermark retrieval. This finding underscores the importance of careful silence management, as silent segments can implicitly encode information.

The idea was for the replacement attack to be more harmful, but due to quality criteria, a large number of similar blocks (K=30) were selected for least squares approximation. As a result, the replaced block is reconstructed with high fidelity, including the watermark pattern within it.

AI-based attacks prove to be completely destructive to the watermarking process across all evaluated models. This outcome was expected, as these attacks introduce significant transformations to the signal, effectively erasing the watermarking information in the process.

Because each model relies on a preferred sampling rate, resampling is a preprocessing step that applies to AI-based attacks. As a prerequisite, each watermarking model must be robust to this procedure; otherwise, evaluating it against AI-based attacks would be pointless. In our tests, all models proved capable of handling resampling without performance degradation, so we considered any disruption observed thereafter to stem from the attack itself rather than the resampling process.

Figure 2 illustrates the accuracy of the watermarking models under various parameterized attack scenarios. We selected attacks with meaningful parameter ranges, emphasizing those where at least one model exhibited partial robustness. Notably, the AudioSeal model shows promising results, nearing the robustness threshold (80%) in several attack configurations, indicating potential resilience upon moderate adjustments. Conversely, WavMark and SilentCipher display more rapid accuracy deterioration under similar conditions.

5 Conclusion and future work

The current state-of-the-art watermarking models demonstrate a certain level of robustness against various attacks but are far from being impervious. An optimal solution may involve combining multiple approaches, while the long-term goal lies in advancing deep learning technology. Certain attacks may be mitigated by refining the underlying operational logic employed by GenAI companies, but a philosophical question arises: How much robustness is truly necessary? Overediting, for instance, should not preserve the watermark.

We acknowledge the relevance of adversarial attacks in watermarking and plan to extend the DeepMark benchmark with a dedicated set of such attacks to evaluate model robustness under these emerging threats.

Additionally, we see the potential to improve usability by introducing a dedicated plugin manager for datasets, allowing for more flexible and streamlined integration of custom or standardized evaluation data, which is not supported in the present version.

Looking ahead, our benchmark aims to serve as a baseline for the community, remaining open to contributions such as new attacks and optimizations. At the same time, we will continue to refine and expand it to meet future challenges.

References

S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, J. Hoffman, M.-J. Hwang, H. Inaguma, C. Klaiber, I. Kulikov, P. Li, D. Licht, J. Maillard, R. Mavlyutov, A. Rakotoarison, K. R. Sadagopan, A. Ramakrishnan, T. Tran, G. Wenzek, Y. Yang, E. Ye, I. Evtimov, P. Fernandez, C. Gao, P. Hansanti, E. Kalbassi, A. Kallet, A. Kozhevnikov, G. M. Gonzalez, R. San Roman, C. Touret, C. Wong, C. Wood, B. Yu, P. Andrews, C. Balioglu, P.-J. Chen, M. R. Costa-jussà, M. Elbayad, H. Gong, F. Guzmán, K. Heffernan, S. Jain, J. Kao, A. Lee, X. Ma, A. Mourachko, B. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, A. Sun, P. Tomasello, C. Wang, J. Wang, S. Wang, and M. Williamson. Seamless: Multilingual expressive and streaming speech translation. *arXiv e-prints*, art. arXiv:2312.05187, 2023. doi: 10.48550/arXiv.2312.05187.

A. Caillon and P. Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv e-prints*, art. arXiv:2111.05011, 2021. doi: 10.48550/arXiv.2111.05011.

G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei. Wavmark: Watermarking for audio generation. *arXiv e-prints*, art. arXiv:2308.12770, 2023. doi: 10.48550/arXiv.2308.12770.

B Chesney and D. Citron. Deep fakes: A looming challenge for privacy. *California Law Review*, 2019. doi: 10.15779/Z38RV0D15J.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Official Journal of the European Union, L series, pp. 1–144.

H. Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1 (4), 2022. doi: 10.54501/jots.v1i4.56.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv e-prints*, art. arXiv:2006.11239, 2020. doi: 10.48550/arXiv.2006.11239.

G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. L. Thing. Twenty years of digital audio watermarking—a comprehensive review. *Signal Processing*, 128:222–242, 2016. doi: 10.1016/j.sigpro.2016.04.005.

D. Kirovski, F. A. P. Petitcolas, and Z. Landau. The replacement attack. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1922–1931, 2007. doi: 10.1109/TASL.2007.900088.

J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv e-prints*, art. arXiv:2010.05646, 2020. doi: 10.48550/arXiv.2010.05646.

S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon. BigVGAN: A universal neural vocoder with large-scale training. *arXiv e-prints*, art. arXiv:2206.04658, 2022. doi: 10.48550/arXiv.2206.04658.

Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1):1–41, 2021. doi: 10.1145/3425780.

K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski. Robust speech watermarking by a jointly trained embedder and detector using a dnn. *Digital Signal Processing*, 122:103381, 2022. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2021.103381.

K. Pavlović, S. Kovačević, I Djurović, and A. Wojciechowski. DNN-based speech watermarking resistant to desynchronization attacks. *International Journal of Wavelets, Multiresolution and Information Processing*, 21(5):2350009, 2023. doi: 10.1142/S0219691323500091.

M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio. SpeechBrain: A general-purpose speech toolkit. *arXiv e-prints*, art. arXiv:2106.04624, 2021. doi: 10.48550/arXiv.2106.04624.

R. San Roman, P. Fernandez, H. Elsahar, A. Defossez, T. Furon, and T. Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. doi: 10.5555/3692070.3693829.

M. K. Singh, N. Takahashi, W. Liao, and Y. Mitsufuji. SilentCipher: Deep audio watermarking. In *Proceedings Interspeech 2024*, 2024. doi: 10.21437/Interspeech.2024-174.

R. D. Smith. Audio Diffusion: Pytorch implementation of diffusion models for audio generation. https://github.com/teticio/audio-diffusion, 2023. Accessed: February 4, 2025.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv e-prints*, art. arXiv:2010.02502, 2020. doi: 10.48550/arXiv.2010.02502.

M. Steinebach, F. A. P. Petitcolas, F. Raynal, J. Dittmann, C. Fontaine, S. Seibel, N. Fates, and L. C. Ferri. StirMark benchmark: audio watermarking attacks. In *Proceedings International Conference on Information Technology: Coding and Computing*, pp. 49–54, 2001. doi: 10.1109/ITCC.2001.918764.

C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong. Attention is all you need in speech separation. *arXiv e-prints*, art. arXiv:2010.13154, 2020. doi: 10.48550/arXiv.2010.13154.

C. Veaux, J. Yamagishi, and K. MacDonald. SUPERSEDED - CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit, 2017.

Z. Ye, P. Sun, J. Lei, H. Lin, X. Tan, Z. Dai, Q. Kong, J. Chen, J. Pan, Q. Liu, Y. Guo, and W. Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. *arXiv e-prints*, art. arXiv:2408.17175, 2024. doi: 10.48550/arXiv.2408.17175.