



NORTHWESTERN UNIVERSITY

Computer Science Department

Technical Report
Number: NU-CS-2026-39

September, 2026

Token Fixes: From Reactive to Proactive Algorithmic Interventions for Responsible Audio AI

Patrick O'Reilly

Abstract

While current AI systems represent significant technological achievements, their success in large-scale, low-cost automation of audio processing tasks comes with a number of documented and potential harms. As a result, researchers have proposed a variety of algorithmic interventions -- technical methods that disrupt, identify, constrain, or otherwise influence the operation of AI systems directly to mitigate harms -- with mixed results. This dissertation asks three questions: where and why have existing algorithmic interventions fallen short? Where have they shown promise? And how can they be made more effective going forward? To answer, I present original research on three classes of algorithmic intervention: adversarial interventions that interfere with irresponsibly deployed AI systems to protect the privacy and intellectual property of those interacting with them; cooperative interventions that operate alongside deployed AI systems with the support of developers or digital media platforms to identify or constrain AI outputs; and design interventions that incorporate governance and harm-mitigation objectives directly into the development of AI systems. I operationalize and analyze these intervention classes in the audio modality through (1) a novel adversarial intervention that perturbs voice communications in real-time to thwart surveilling speaker recognition systems; (2) a novel cooperative intervention for identifying AI-generated audio through an end-to-end neural network-based watermarking system with state-of-the-art robustness to removal attacks; (3) a study identifying failure modes of audio watermarking methods under challenging real-world conditions, revealing limitations of algorithmic interventions in audio more broadly; and (4) a novel neural audio codec training method that lays the groundwork for design-level interventions in the form of generative watermarks that operate within rather than alongside deployed AI systems. Through this work, I argue that algorithmic interventions can be made more effective by moving from reactive, extrinsic measures that sit outside of audio AI systems ("token fixes") to proactive, intrinsic measures that shape the design and behavior of audio AI systems.

Artificial Intelligence, Audio, Deep Learning

NORTHWESTERN UNIVERSITY

Token Fixes: From Reactive to Proactive Algorithmic Interventions for Responsible Audio AI

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Patrick O'Reilly

EVANSTON, ILLINOIS

September 2026

© Copyright by Patrick O'Reilly 2026
All Rights Reserved

Abstract

While current AI systems represent significant technological achievements, their success in large-scale, low-cost automation of audio processing tasks comes with a number of documented and potential harms. As a result, researchers have proposed a variety of algorithmic interventions – technical methods that disrupt, identify, constrain, or otherwise influence the operation of AI systems directly to mitigate harms – with mixed results. This dissertation asks three questions: where and why have existing algorithmic interventions fallen short? Where have they shown promise? And how can they be made more effective going forward? To answer, I present original research on three classes of algorithmic intervention: adversarial interventions that interfere with irresponsibly deployed AI systems to protect the privacy and intellectual property of those interacting with them; cooperative interventions that operate alongside deployed AI systems with the support of developers or digital media platforms to identify or constrain AI outputs; and design interventions that incorporate governance and harm-mitigation objectives directly into the development of AI systems. I operationalize and analyze these intervention classes in the audio modality through (1) a novel adversarial intervention that perturbs voice communications in real-time to thwart surveilling speaker recognition systems; (2) a novel cooperative intervention for identifying AI-generated audio through an end-to-end neural network-based watermarking system with state-of-the-art robustness to removal attacks; (3) a study identifying failure modes of audio watermarking methods under challenging real-world conditions, revealing limitations of algorithmic interventions in audio more broadly; and (4) a novel neural audio codec training method that lays the groundwork for design-level interventions in the form of generative watermarks that operate within rather than alongside deployed AI systems. Through this work, I argue that algorithmic interventions can be made more effective by moving from reactive, extrinsic measures that sit outside of audio AI systems (“token fixes”) to proactive, intrinsic measures that shape the design and behavior of audio AI systems.

Acknowledgment

I have many people to thank for what has been an enjoyable and rewarding PhD experience. First among these is my advisor, Bryan Pardo, who allowed me to chart a path while working to keep it clear at every turn: securing funding, building collaborations, nurturing promising ideas (and gracefully pointing out the not-so-promising ones), and finally nudging me out the door when the time came. Every research project of my PhD benefited from Bryan's steady guidance, technical expertise, and perspective, and I'm beyond grateful to have had him in my corner these past six years.

I would like to thank the members of my dissertation committee: Fabián Bustamante, who walked me through the basics of the PhD during my first term and helped ease my transition into the program against the backdrop of a global pandemic; Aravindan Vijayaraghavan, who helped guide me through my first PhD research project, graciously listening to my scattered ideas and pointing me in productive directions; and Adam Finkelstein, a constant source of encouragement, advice, and research inspiration in the late stages of my PhD.

I would like to thank all the mentors who have patiently shepherded me through research projects and crises of confidence (often related!) at Northwestern and beyond, including Prem Seetharaman, Ethan Manilow, Zeyu Jin, Jiaqi Su, and Pranjal Awasthi. I feel fortunate to have worked and learned alongside such creative and talented researchers.

I would like to thank my current and former labmates Hugo Flores García, Annie Chu, Julia Barnett, and Max Morrison, who made the PhD a fun and friendly experience above all. It's been a joy spending the last six years with such fundamentally cool people, and I can't imagine having done the PhD anywhere else. I would also like to thank my wonderful research collaborators Saumya Pailwan, Andreas Bugler, Adrian Hoffer, Keshav Bhandari, Anabella Isaro, and Nathan Pruyne, who helped turn manic whiteboard scribbles into prototypes and papers and contributed so many interesting ideas along the way.

I would like to thank the staff of Northwestern's Department of Computer Science for their oft-unseen but always appreciated work to ensure that my PhD went smoothly – from last-minute experiment expenses and IRB approvals to travel bookings and registration fixes.

I would like to thank my undergraduate advisor Justin London, who nurtured my interest in music research and inspired me to pursue it through graduate school, as well as my many excellent teachers and mentors at Carleton College and at the University of Illinois at Chicago for giving me the tools to navigate the PhD.

I would like to thank my mom, Catherine, and my dad, Gregory, who helped kick off this journey ten years ago by sending me a one-line email with the subject "NU" and a screenshot of the Interactive Audio Lab webpage. I owe everything to their tireless support in navigating the PhD and many much more important things. I am similarly indebted to my brothers, Sean and Declan, and my sister-in-law, Andrea, who have helped me break out of many a rut and kept me grounded even when my hours were at their least conventional. And I would like to thank my cousins, aunts, uncles, and grandparents for their advice, empathy, and company on this adventure. This dissertation exists because of you.

TABLE OF CONTENTS

Abstract	3
Acknowledgment	4
Table of Contents	6
List of Tables	10
List of Figures	12
1 Introduction	16
2 Background	22
2.1 Audio AI Fundamentals	22
2.1.1 Digital Audio Representations	23
2.1.2 Deep Learning	24
2.1.3 Relevant Deep Learning Paradigms	26
2.2 AI Interventions	33
2.2.1 Regulatory Interventions	35
2.2.2 Economic Interventions	37
2.2.3 Algorithmic Interventions	38
3 Adversarial Intervention	40
3.1 Preliminaries	41
3.1.1 Biometric Surveillance of Voice Communications	42
3.1.2 Speaker Recognition Systems	42
3.1.3 Prior Work	45
3.2 Method	46
3.2.1 Encoder	48
3.2.2 Bottleneck	50
3.2.3 Decoder	50

3.2.4	Training objective	51
3.2.5	Streaming Implementation	54
3.3	Experiments	54
3.3.1	Models	55
3.3.2	Data	55
3.3.3	Attack Algorithms	56
3.3.4	Objective evaluation	57
3.3.5	Subjective evaluation	60
3.4	Results and Discussion	62
3.4.1	Limitations and Future Work	63
3.5	Impact and Follow-On	65
4	Cooperative Intervention	67
4.1	Preliminaries	68
4.1.1	Speech Synthesis	68
4.1.2	Synthetic Audio Identification	69
4.2	Method	72
4.2.1	Architecture	73
4.2.2	Training	75
4.3	Experiments	76
4.3.1	Baselines and Comparison Methodology	77
4.3.2	Robustness to Audio Transformations	78
4.3.3	Perceptual Transparency	81
4.4	Results and Discussion	82
4.4.1	Limitations and Future Work	83
4.5	Impact and Follow-On	84
5	Limits of Algorithmic Intervention	86

- 5.1 Preliminaries 87
 - 5.1.1 (Post-Hoc) Digital Audio Watermarking 87
 - 5.1.2 Evaluations of Audio Watermark Robustness 90
- 5.2 Audio Transformations 91
 - 5.2.1 Signal-Processing Transformations 91
 - 5.2.2 Audio Codec Compression 91
 - 5.2.3 Neural Vocoder 92
 - 5.2.4 Neural Denoisers 92
 - 5.2.5 Watermark-Aware Transformations 93
- 5.3 Experiments 93
 - 5.3.1 Watermarking Methods 93
 - 5.3.2 Dataset 94
 - 5.3.3 Transformations 96
 - 5.3.4 Metrics 97
- 5.4 Results and Discussion 99
 - 5.4.1 Watermark Removal 99
 - 5.4.2 Implications for Post-Hoc Algorithmic Interventions 102
- 5.5 Impact and Follow-on 104
- 6 Towards Design Intervention 105
 - 6.1 Watermarking as Design Intervention 105
 - 6.2 Codec Idempotence: A Barrier to Generative Audio Watermarks 108
 - 6.3 Preliminaries 110
 - 6.3.1 Generative Watermarking for Audio 110
 - 6.3.2 Neural Audio Codecs 112
 - 6.3.3 Studies of Neural Codec Properties 113
 - 6.4 Measuring Idempotence 114

6.5	Encouraging Idempotence	117
6.6	Downstream Generative Modeling	120
6.7	Discussion	122
7	Conclusion	123
7.1	Summary of Research Contributions	124
7.2	Common Research Themes	127
7.3	Closing Remarks: Building Better Interventions	129
	References	132

LIST OF TABLES

3.1	Objective evaluation results. We perform attacks on the VoxCeleb1 dataset against seen (ResNetSe34v2) and unseen (Y-Vector) speaker recognition models, and compute both top-1 and top-10 recognition accuracies. Additionally, we compute objective speech quality metrics for each attack (PESQ, STOI) and measure the intelligibility of attacked speech over the LibriSpeech train-clean-360 dataset in terms of word- and character-error rates (WER, CER) obtained from a pretrained Wav2Vec 2.0 model.	59
3.2	Top-1 (T-1) and top-10 (T-10) recognition accuracy of the ResNetSE34v2 system on the de-identification attacks described in Section 3.3.3 when all query audio is passed through a Demucs [186] speech enhancement processing stage.	59
3.3	Effect of adversarial utterance enrollment on de-identification performance of VoiceBlock. For sets of 15 clean and adversarial query utterances, we compare the top-1 (T-1) and top-10 (T-10) accuracy of speaker recognition with the ResNetSE34v2 model under three conditions. Clean profile: 20 clean utterances are enrolled per speaker. Mixed profile: 10 clean and 10 adversarial (VoiceBlock) utterances are enrolled per speaker. Adversarial profile: 20 adversarial (VoiceBlock) utterances are enrolled per speaker.	60
3.4	Results of training VoiceBlock with different auxiliary loss functions. We consider a simple waveform regression loss (" L_1 ") and the mel-cepstral cosine-similarity loss used by O'Reilly et al. ("MFCC-cosine") [181] as simple drop-in replacements for the combined waveform/spectrogram loss of Defosséz et al. [186] (" $L_1 + \text{MRS}$ "). Additionally, we consider a feature-matching loss on the acoustic representations produced by a Wav2Vec 2.0 [241] automatic speech recognition model ("ASR"). This loss is computed by taking the L_1 distance between acoustic feature vectors produced by the Wav2Vec 2.0 encoder over clean and adversarial utterances, scaled by a factor of $1e^{-6}$	65
4.1	Objective speech quality metrics and human identification rates for the proposed MaskMark and baseline watermarking methods. DNN-A values are shaded to indicate it fails to detect watermarked speech at low false-positive rates ($< 4\%$). . .	82

5.1	Watermark detection under transformations. Under <i>Quality Preservation</i> , we evaluate the audio degradation incurred by each transformation in terms of intelligibility ("ASR-CER"), speaker identity similarity ("SIM"), and overall quality ("SQUIM-MOS"). Cells are shaded darker where degradation is minimized. Arrows in column headers indicate the optimal direction to minimize audio degradation. Under <i>Watermark Detection</i> , we evaluate the effect of each transformation on the true-positive detection rate of watermarks, given a fixed 1% false-negative rate ("TPR@1%FPR"). Lower values indicate stronger watermark removal. Cells are shaded darker as the true positive detection rate decreases. The transformations that most effectively and inconspicuously remove watermarks are therefore shaded in every column.	95
6.1	Comparison of Codecs	116
6.2	Generative Modeling on Codec Tokens	121

LIST OF FIGURES

2.1	Illustration of structure within a speech waveform at increasingly fine time-scales. The magnitude spectrogram, pictured above the waveform, represents the amplitudes of frequency components in short time segments of the waveform; brighter intensities correspond to larger amplitudes.	24
2.2	Illustration of common neural network-based latent representations used for generative modeling of audio. VAE latents are sequences of continuous vectors, VQ-VAE latents are sequences of integer indices into a codebook, and RVQ-VAE latents are stacked sequences of integer indices into n_q residual codebooks.	30
2.3	Illustrations of the three main generative modeling paradigms discussed in this dissertation: autoregressive language modeling of discrete token sequences; masked language modeling of discrete token sequences; and diffusion modeling of continuous token sequences. Each pictured method features a generator network G operating on a latent or token representation (shown as colored squares or rectangles), and samples instances iteratively over a number of generation steps. "Mask" tokens are shown in dark grey, and the diffusion timestep is represented through shading. The neural network architecture used for the generator is arbitrary, although current models typically use transformer-based [71] generators.	33
3.1	Illustration of biometric profiling via speaker recognition systems and the effect of the proposed VoiceBlock method. 1. Speaker recognition systems use neural networks to map speech utterances to an embedding space in which proximity corresponds to perceptual speaker similarity. 2. Given enrolled <i>profiles</i> of speakers consisting of labeled embedded utterances, the speakers of new <i>query</i> utterances can be identified by embedding and checking the stored label of nearby utterances. 3. VoiceBlock intervenes by perturbing a user's audio stream such that any captured utterances will be mapped to a distant region of a speaker recognition system's embedding space.	44
3.2	Left: The proposed VoiceBlock architecture. Acoustic features extracted by the encoder are fed to the recurrent bottleneck to predict filtering controls, which are regularized and applied to the input by the decoder to obtain adversarial audio. Right: VoiceBlock adversarially perturbs the user's audio stream such that any extracted queries are scored by a speaker recognition system as dissimilar to the user's enrolled utterances, hampering identification or retrieval.	47

3.3	Time-varying magnitude responses of 128-band filter controls predicted by VoiceBlock (top) to de-identify a given audio recording (bottom). Filter bands are arranged vertically by frequency, and the frequency magnitude response of each band over time is represented by the corresponding curve. The thin grey line beneath each curve represents a "neutral" frequency magnitude response of 1, with deviations above and below this line representing boosts and cuts, respectively. Note that the controls vary smoothly, and often exhibit sparseness in quiet regions of the waveform.	52
3.4	Spectrogram plots of each evaluated attack method operating on source audio (Clean, at right). Note that while Noise, Spectral Gating, and Universal attacks introduce perceptible degradations to the source audio (visible in the corresponding spectrograms), the proposed VoiceBlock attack modifies the source audio more subtly.	57
3.5	Distributions of quality ratings for VoiceBlock and baseline methods from crowd-sourced MUSHRA-style listening test. Higher numbers are better. Black dots are means and white dots are medians. Wilcoxon signed-ranked tests between all pairs of conditions show statistical significance at $p < 0.05$	61
4.1	Top row: spectrograms of a two-second speech utterance watermarked with the proposed MaskMark and baseline methods. Bottom row: waveform difference between original and watermarked utterances for each method, highlighted against the original waveform.	71
4.2	The proposed MaskMark architecture. Given a binary key vector and carrier audio to be watermarked, the embedder network modifies the carrier audio to encode the key. During training, embedded audio is then passed through a variety of acoustic transformations to simulate realistic conditions. The detector network is trained to predict the key vector embedded in (transformed) watermarked audio; because these keys are randomly distributed at training time, the detector's predictions over unwatermarked audio at inference time are likewise random. Detection is performed by computing the proportion of matching bits between the detector's prediction and a known watermark key.	73
4.3	TPR@1%FPR for watermarks under selected transformations, indicating the robustness of each method when configured to operate with a fixed 1% false-positive rate. Scores in the top row are for the DAPS dataset, scores in the bottom row for our synthesized VITS dataset. The proposed MaskMark method is robust to more transformations than the baselines. DNN-A has a true positive rate of 0% for false-positive rates below 4%, and thus its bars are not visible.	79
4.4	Receiver Operating Characteristic (ROC) curves and detector score histograms for the proposed MaskMark method and the DNN-A method of Pavlović et al. over the DAPS-clean subset. The dotted red line marks the average score of unwatermarked segments, while the solid red line marks the maximum observed score of unwatermarked segments. The method of Pavlović et al. produces markedly higher detection scores for unwatermarked audio, making operation at low false-positive rates impossible.	80

5.1	Watermark signatures. Given a short audio excerpt ("Original"), we apply five watermarking methods described in Section 5.3.1 ("AudioSeal," "WavMark," "MaskMark," "Timbre," and "Collaborative") and visualize the difference between original and watermarked audio at the spectrogram (top row) and waveform (bottom row). For each watermarking method, the original waveform is shown in grey and the difference in color.	89
5.2	Robustness to low-bitrate codecs. We plot the achievable true-positive detection rate at a fixed 1% false-positive rate ("TPR@1%FPR") for each watermarking method under transformation via low-bitrate neural codecs. Watermark detection declines at low bitrates while audio quality is generally preserved, as indicated by SQUIM-MOS scores.	98
5.3	Denoiser attack. We illustrate the application of the denoiser attack to Timbre-Watermark with spectrogram plots. Artifacts introduced by the watermark (blue rectangle) can be seen more clearly when compared to the original spectrogram (green rectangle); these are removed by the application of noise followed by the denoiser (yellow rectangle).	100
6.1	Generative watermarking for audio: whereas post-hoc watermarking methods embed signatures in generated audio through imperceptible perturbations (top row), generative watermarks intervene within the generative process to embed statistical signatures (bottom row). Pictured here, a generative watermark intervenes in the sampling of tokens by an autoregressive audio language model; the watermark can then in theory be detected by mapping generated audio back to tokens and performing a statistical test. <i>Note that if we cannot detect the presence of watermarked tokens after decoding and re-encoding, we cannot detect the generative watermark.</i>	106
6.2	Non-idempotence in neural codecs: decoding the tokens produced by a neural codec to audio and subsequently re-encoding may result in a substantially different set of tokens, hampering watermarking methods for audio generative models that rely on the sampling and recognition of specific tokens.	109
6.3	Audio and code stability over encodings: audio quality (in PESQ and SI-SDR) of selected neural codecs by encoding iteration, average codebook use by encoding iteration, and match rate between tokens encoded at the n^{th} and $(n + 1)^{\text{th}}$ iterations.	115
6.4	Phase sensitivity correlates with idempotence: For each selected codec, we compute the token match rate between encodings of audio inputs under time shifts between $\pm 2\text{ms}$. We then compute the correlation between average observed match rate and PESQ score after 25 encoding iterations. In general, codecs with higher phase sensitivity (as indicated by lower match rates) exhibit higher idempotence (as indicated by higher PESQ scores).	117

6.5	Effect of idempotent training: audio quality (in PESQ and SI-SDR), average codebook use, and match rate between tokens encoded at the n^{th} and $(n + 1)^{\text{th}}$ iterations as a function of encoding iterations for our DAC reproduction and fine-tuned variants with idempotence losses described in Section 6.5. Fine-tuning with $\mathcal{L}_{code}$ results in the highest overall idempotence through 25 encoding iterations, while preserving codebook use.	119
-----	---	-----

CHAPTER 1

INTRODUCTION

The past decade has seen an explosion in the capabilities – and correspondingly, adoption – of artificial intelligence (AI) systems for music and general audio. AI systems have powered the rise of massively popular commercial-grade song generation services such as Suno and Udio; driven growth in text-to-speech and voice cloning services such as ElevenLabs; and served as the backbone of popular audio editing and enhancement products from companies such as Descript and Adobe. Alongside such proprietary systems, open-source AI tools now widely facilitate transcription [1], symbolic and audio-domain music generation [2, 3], musical timbre modification [4], voice conversion [5], source separation [6], music mixing and mastering [7, 8], and other tasks relevant to creative workflows.

While current AI systems represent significant technological achievements, their success in large-scale, low-cost automation of audio processing tasks comes with a number of documented and potential harms. Transcription and speaker identification systems can be used to streamline mass surveillance [9–11]. Voice-cloning systems can be used to impersonate the voices of individuals, serving as tools for blackmail, fraud, and misinformation [12, 13]. AI music generation systems pose challenges to existing intellectual property frameworks due to their ability to produce realistic outputs that convincingly mimic artistic styles [14]. As the outputs of these systems flood digital platforms – streaming service Deezer claims that 44% of music recordings uploaded to its platform are now “AI-generated” [15] – musicians and listeners must reckon with established notions of creativity and artistry and navigate a radically reshaped media environment [16]. Worryingly, highly capable AI systems threaten to displace labor and siphon already-slim income streams from voice actors, musicians, audio engineers, and other creative professionals on whose works these

systems are often trained without license [17–19].

AI’s emerging social and economic impacts have led researchers, artist groups, and media platforms to propose a variety of interventions for mitigating the harms of irresponsibly deployed AI systems, ranging from regulatory measures (bans on AI use and development [16, 20, 21], content restrictions [22], and disclosure requirements [23–25]) to economic frameworks (data licensing [26, 27], attribution [28], and royalty distribution for AI outputs [29]). Within these broader mitigation efforts, a large number of AI researchers have turned to the tools of their discipline – namely, deep learning methods – to develop *algorithmic interventions* that disrupt, identify, constrain, or otherwise influence the operation of AI systems more directly [9, 30–33].

The results of these efforts have often been disappointing. Regulatory and economic interventions have run up against predictable roadblocks: institutional inertia, difficulty building political coalitions and securing industry buy-in, and the challenges of standardizing and scaling governance proposals to fit the rapidly evolving capabilities of AI technologies [34–38]. Troublingly, algorithmic interventions have struggled to effectively curtail AI harms despite leveraging the techniques that have proven so effective in driving AI capabilities forward [39–41]. This impasse has created a frustrating state of affairs for both the AI researchers pursuing these interventions and the creative professionals whose livelihoods and artistic practices stand to be affected by their success or failure.

This dissertation asks three questions: where and why have existing algorithmic interventions fallen short? Where have they shown promise? And how can they be made more effective going forward? To answer, I present original research on three classes of algorithmic intervention: *adversarial interventions* that aim to interfere with irresponsibly deployed AI systems in order to protect the privacy and intellectual property of those interacting with them; *cooperative interventions* that operate alongside deployed AI systems with the support of developers or platforms to identify or constrain AI outputs; and *design interventions* that incorporate governance and harm-mitigation objectives directly into the development of AI systems. I analyze and operationalize these intervention classes in the audio modality through works spanning voice privacy, synthetic media

identification, and neural codec design. The dissertation is structured as follows:

- In **Chapter 2: Background**, I provide an overview of audio AI system design and review the broader context of audio AI harm-mitigation interventions.
- In **Chapter 3: Adversarial Intervention**, I present a novel adversarial intervention for de-identifying voice communications in real-time to thwart biometric surveillance systems.
- In **Chapter 4: Cooperative Intervention**, I present a novel cooperative intervention for identifying AI-generated audio through an end-to-end neural network-based watermarking system.
- In **Chapter 5: Limits of Algorithmic Intervention**, I present a study identifying failure modes of audio watermarking methods under challenging real-world conditions, revealing limitations of post-hoc algorithmic interventions in audio more broadly.
- In **Chapter 6: Towards Design Intervention**, I present a novel neural codec training method that lays the groundwork for design-level interventions in the form of generative watermarks that operate within rather than alongside deployed AI systems.
- I summarize my findings and provide closing thoughts in **Chapter 7: Conclusion**.

Across these chapters, I argue that many algorithmic interventions are merely *token fixes*¹: technical methods that attempt to constrain AI systems from the outside, post-hoc, after capabilities and system behaviors have already been determined. As a result, such interventions risk serving as symbolic gestures toward responsibility while remaining detached from the design choices and circumstances that shape the use of AI systems in practice.

The published works presented in this dissertation outline a corrective shift from reactive, extrinsic measures that sit outside of AI systems to proactive, intrinsic measures that shape the

¹In AI research, a “token” is a unit of data consumed or produced by an AI model. More generally, “token” refers to perfunctory or symbolic efforts.

design of AI systems directly. While these works center on biometric privacy and watermarking, they provide useful tools and lessons for the development of more effective algorithmic interventions in audio AI systems more broadly.

Research Contributions

This dissertation has been undertaken as a series of published research projects. In Chapters 3–6, I present expanded and reformatted versions of these works in addition to analysis and discussion. These works span the following areas:

Adversarial intervention: I present a novel method for perturbing voice communications in real-time to thwart surveilling speaker recognition systems, published at NeurIPS 2022 as VoiceBlock: Privacy through Real-Time Adversarial Attacks with Audio-to-Audio Models [10].

Cooperative intervention: I present a method for identifying AI-generated audio through an end-to-end neural network watermarking system with state-of-the-art robustness under channel distortions and watermark removal attacks, published at ICASSP 2024 as MaskMark: Robust Neural Watermarking for Real and Synthetic Speech [42].

Limits of algorithmic intervention: I diagnose failure modes of audio watermarking methods under challenging real-world conditions and demonstrate neural network-based attacks capable of removing state-of-the-art watermarks in work presented at ICLR WMARK 2025 as Deep Audio Watermarks are Shallow: Limitations of Post-Hoc Watermarking Techniques for Speech [40].

Design intervention: I identify a technical limitation of neural network-based audio codecs that hinders the implementation of “generative” watermarking methods in the audio modality and present a codec fine-tuning method to overcome this limitation in work published at ICASSP 2025 as Code Drift: Towards Idempotent Neural Audio Codecs [43].

Broader Impact and Intellectual Interest

This dissertation encompasses a number of algorithmic interventions that address AI harms spanning biometric surveillance, fraud, and misinformation. Through the development and analysis of these algorithmic interventions, I provide technical tools to aid in the detection of AI-generated media and prevent misuse of AI systems; draw connections between intervention strategies across research areas (e.g., between audio-modality adversarial attacks and watermarking methods) and situate these strategies within the broader context of algorithmic interventions; identify limitations in existing intervention methods; and show how proactive interventions within the design space of AI systems can offer advantages over reactive, external interventions that seek to constrain these systems only after deployment.

The techniques developed in this dissertation have direct relevance to research in a number of areas, including adversarial data protection; AI-generated audio identification and watermarking; AI-generated media detection beyond the audio modality (e.g. image and video); neural network-based steganography and information-hiding; deep generative modeling for audio; and AI ethics.

The development and analysis of these techniques also holds relevance for a number of communities, including the providers of AI systems for music and audio generation; the users of digital media platforms on which AI-generated media may appear; artists and creative laborers whose livelihoods are impacted by the spread of AI-generated media; and the legal bodies regulating the use, identification, and attribution of AI-generated media. For example, improved tools for the identification of AI-generated audio can allow model providers to give stronger assurances that harmful content produced by their systems can be detected in accordance with regulations, as well as to track general use of their products; can inform users of digital media platforms as to the nature and provenance of AI-generated content to which they are exposed; can let artists and other digital content creators trace the use of their own content, and to make informed decisions as to their use of generated content by exposing provenance information; and can help legal bodies to take

effective action on AI-generated media identification by providing practical solutions and realistic characterizations of their limitations.

CHAPTER 2

BACKGROUND

This chapter covers the fundamentals of modern audio AI system design and reviews the broader context of interventions intended to address the social and economic impacts of increasingly capable AI systems. In doing so, I establish necessary information to ground the research contributions presented in Chapters 3 – 6. Discussions of my proposed algorithmic interventions, the specific AI systems targeted by these interventions, and more narrowly relevant prior works are deferred to the corresponding chapters.

2.1 Audio AI Fundamentals

Modern AI systems perform a variety of information processing tasks by leveraging *deep representation learning*, the training of neural networks to extract structured information from data [44]; *deep generative modeling*, the training of neural networks to sample from probability distributions [45, 46]; and *deep reinforcement learning*, the training of neural networks to perform sequential tasks through the promotion of preferred outputs [47]. Research advances within these paradigms have combined to produce remarkable improvements in the capabilities of AI systems. A decade ago, AI proponents could point to LSTM systems [48] (2016) that performed competent translation between language pairs [49] or early Variational Autoencoder [50] (2013) and Generative Adversarial Network [51] (2014) systems that produced blurry, low-resolution images. Today, AI systems can create realistic high-resolution images [52] and videos [53], converse fluently in natural language [54], prove mathematical theorems [55], and discover critical software vulnerabilities [56].

These advances have been similarly sharp for AI systems operating in the audio modality,

spanning tasks as diverse as text-to-speech [57], voice conversion [5], audio enhancement and editing [58], sound event detection [59] and music generation [3]. In the following sections, I provide an overview of concepts in audio processing and AI system design relevant to the discussion of these capabilities, the algorithmic interventions that grapple with them, and my own research contributions.

2.1.1 Digital Audio Representations

Audio is often represented digitally as a *waveform* – a sequence of values measuring changes in air pressure. These values are sampled at a regular interval, typically between 16,000 and 48,000 times per second (Hz), resulting in long sequences, even for short audio recordings [60]. For example, a 44.1kHz recording of a two-minute song may contain over 5 million samples (or over 10 million in stereo). Speech, music, and other audio signals of interest often exhibit complex structures spanning both nearby and distant sequence values (i.e., at both fine and coarse time scales), making statistical modeling of waveforms difficult [61].

For purposes of analysis and manipulation, audio can also be represented as a spectrogram, a time-frequency matrix holding the Fourier transform of short (typically overlapping) segments of the waveform. The resulting complex-valued representation contains both magnitude and phase information for each time-frequency index, and is thus invertible (i.e., the waveform can be recovered from the spectrogram via an inverse Fourier transform). Spectrograms benefit from a reduction in sequence length relative to the waveform (due to the windowed spectrum computation), as well as the disentanglement of frequency bands. For analysis, phase information is often discarded, resulting in a non-invertible magnitude spectrogram. Figure 2.1 illustrates waveform and spectrogram representations of a speech recording.

In addition to waveform and spectrogram representations, deep learning methods can be used to produce more compact and task-specialized representations of audio recordings, at the expense of exact invertibility (see Section 2.1.3) [62–64]. The audio AI systems discussed in this dissertation

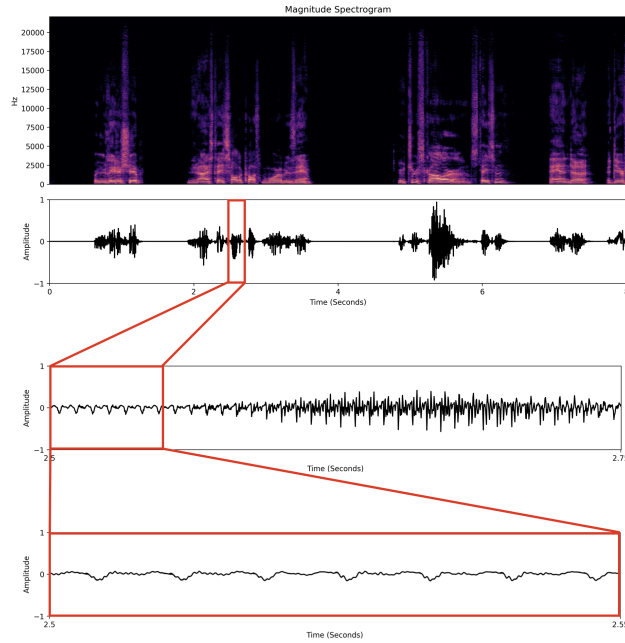


Figure 2.1. Illustration of structure within a speech waveform at increasingly fine time-scales. The magnitude spectrogram, pictured above the waveform, represents the amplitudes of frequency components in short time segments of the waveform; brighter intensities correspond to larger amplitudes.

operate on one or more of waveform, spectrogram, and deep learning-based audio representations.

2.1.2 Deep Learning

Modern audio AI systems primarily process information via deep neural networks – function approximators composed of multiple operations or "layers" whose parameters are learned via gradient descent on a task-defined objective or "loss" function [65, 66]. Often, individual layers represent simple vector-valued functions such as linear projections or element-wise scaling; in general, these functions are differentiable to facilitate the flow of gradient information through the network. The vectors or vector sequences ingested and produced by these layers are often referred to as “features,” and layers can be combined in a modular fashion to construct a variety of specialized neural network architectures.

Deep neural networks have become popular tools for audio analysis and generation, often com-

plementing or even supplanting traditional signal-processing methods [61, 67, 68]. The neural network models discussed in this dissertation build upon standard architectural components such as the aforementioned linear layers and nonlinear scaling functions, convolutional layers (which process input features via the application of learned convolutional kernels) [69, 70], transformer attention layers (which process feature sequences in terms of pairwise information flow between sequence elements) [71], and pooling layers (which aggregate long feature sequences into a smaller number of feature vectors). The research contributions discussed in this dissertation do not center around novel architectures, but rather the use of standard architectures within novel task formulations, training schemes, and evaluations.

The audio AI systems on which the algorithmic interventions proposed in this dissertation operate leverage two main applications of deep learning: *deep representation learning* and *deep generative modeling*. (While the third aforementioned application – *deep reinforcement learning* – is an increasingly important ingredient in the design of audio AI systems, it is not relevant to the specific systems discussed in this dissertation and is therefore omitted here).

Deep representation learning encompasses a variety of algorithms for training neural networks to map data to informative representations [44]. These mappings are often learned through reconstruction objectives (in which the network must map a representation back to the original instance from which it was derived) [72] and metric objectives (in which the network must preserve known data-space relationships in its representation space) [73]. The learned representations – typically real-valued vectors or vector sequences – may be further regularized to be sparse [74] or low-dimensional [75], or constrained to follow a specific probability distribution [50]. Deep representation learning enables AI systems to perform audio analysis tasks such as speaker and speech recognition [76, 77] by automating the extraction of useful structure from raw data, and allows for the mapping of audio data to compact representations that can be processed efficiently [62].

A *generative model* is a statistical model of a probability distribution $p_{\text{data}}(x)$ that allows for

sampling instances x from the distribution $x \sim p_{\text{data}}(x)$. When generative models are parameterized by deep neural networks, they are commonly referred to as *deep generative models* [45, 46]. These models are often *conditional*, meaning they fit a distribution $p_{\text{data}}(x|c)$ where c is a variable on which the instance x depends – for example, a text description c may condition the generation of a music recording x in a text-to-music model [3]. Current deep generative modeling techniques allow audio AI systems to produce highly realistic outputs across varied distributions such as speech, music, and sound effects [57, 78, 79].

Deep representation learning and deep generative modeling are closely related, and often overlap within the systems discussed in this dissertation. Models trained for representation-learning tasks may leverage generative techniques (e.g., neural codecs that compress audio data for efficient processing often incorporate generative adversarial networks [51, 62] or flow-matching [80, 81]). Correspondingly, models trained for generative tasks often leverage pre-trained representations (e.g., voice conversion systems often extract phonetic content using external speech representation models [82]).

2.1.3 Relevant Deep Learning Paradigms

I next survey the main deep learning techniques relevant to this dissertation, including deep representation learning and deep generative modeling methods.

Metric Representation Learning methods train neural networks to map data to a compressed representation space that captures important structural relationships in the data [44]. These representations are typically vectors or *embeddings* whose proximity under some embedding-space metric (e.g. Euclidean distance) corresponds to some abstract notion of data-space proximity, such as facial similarity in face recognition systems [83]. Embedding models are often trained with contrastive objectives that penalize embedding-space proximity between instances assumed to differ meaningfully in the data space [73, 84], triplet objectives that draw an anchor instance’s embedding close to a positive embedding of a known-similar instance and away from a negative embedding

of a known-dissimilar instance [83], or prototypical objectives that draw a query instance close to the centroid of an embedded support set of similar instances and away from the centroids of embedded support sets of dissimilar instances [85]. The speaker recognition models discussed in Chapter 3 are trained with metric learning objectives to construct speaker embedding spaces in which utterances of the same speaker are clustered together, enabling identification through simple nearest-neighbors lookup [76].

Variational Autoencoders (VAEs) [50] are deep neural networks trained to map data to and from regularized *latent* representations probabilistically. VAEs represent the data distribution $p_{\text{data}}(x)$ via a two-step process in which a latent z is first sampled from a (typically Gaussian) prior distribution $p(z)$, following which an instance x is sampled from the conditional distribution $p_{\theta}(x|z)$ parameterized by θ . Estimating θ directly is not feasible due to the intractability of the data likelihood $p(x) = \int p(z)p_{\theta}(x|z)dz$ and posterior density $p_{\phi}(z|x) = p_{\theta}(x|z)p(z)/p(x)$; instead $p_{\theta}(x|z)$ and the posterior distribution $p_{\phi}(z|x)$ are approximated by neural networks θ and ϕ trained jointly to optimize a lower bound on the data likelihood. This approach results in a noise-regularized auto-encoding formulation in which an input x is probabilistically mapped to a latent representation z via the *encoder* $q_{\phi}(z|x)$, the latent representation z is mapped back to a reconstruction of x via the *decoder* $p_{\theta}(x|z)$, and the latent distribution $p_{\phi}(z|x)$ is encouraged to follow the predetermined prior $p(z)$. Figure 2.2 illustrates a VAE operating on audio.

VAEs sit at the intersection of deep representation learning and generative modeling. Because their prior-constrained latent representations are typically well-structured, low-dimensional, and can be approximately mapped to the data space via the decoder, it is common to train *latent generative models* to sample in a VAE latent space rather than the data space [52], and then pass samples through the decoder to obtain final generations; this configuration is discussed in more depth below. Beyond furnishing latent representations for generative modeling, VAEs can themselves be used to draw samples from a learned data distribution by sampling from the prior $p(z)$ and decoding to obtain a sample from the conditional $p_{\theta}(x|z)$. VAEs serve as components of several audio AI

systems discussed in this dissertation.

Vector-Quantized Variational Autoencoders (VQ-VAEs) [72] represent data through discrete rather than continuous latents. Certain data modalities exhibit naturally discrete structure (e.g. phonemes in speech), and as such may be well suited for generative modeling over discrete rather than continuous representations. The use of discrete latents also allows for the application of language modeling techniques (see below) to continuous modalities such as audio. VQ-VAEs modify the VAE architecture via the replacement of a defined prior distribution with a discrete "codebook" of learnable latent vectors. During training, encoded instances are quantized to the nearest codebook vector, which is passed to the decoder for reconstruction. In a Residual Vector-Quantized Variational Autoencoder (RVQ-VAE), a hierarchy of multiple codebooks can be used to iteratively store residuals between the quantized and un-quantized vectors, allowing for improved reconstruction quality at the expense of an enlarged latent representation [62]. While the lack of a well-defined prior distribution over codebooks hampers unconditional sampling from the latent space, this discrete representation facilitates the learning of categorical priors over codebook vectors [86, 87]. Latent generative modeling can thus be performed on quantized representations, and just as with VAEs sampled instances can be deterministically mapped to the data space via the decoder $p_{\theta}(x|z)$. Many of the neural network-based audio "codec" models discussed in this dissertation utilize a residual VQ-VAE architecture [62–64]. Figure 2.2 illustrates both VQ-VAE and RVQ-VAE networks operating on audio.

Generative Adversarial Networks (GANs) represent the data distribution using a similar latent-variable formulation as VAEs, in which a latent z is sampled from a prior distribution $p(z)$ and an instance x is sampled from the conditional distribution $p_G(x|z)$ [51]. This mapping of latents to the data space is parameterized with a *generator* network G , which can be thought of as analogous to the decoder network in a VAE. However, while VAEs consist of cooperative encoder and decoder networks trained in an autoencoding formulation, GANs pair the generator network with a *discriminator* network D tasked to distinguish between generated samples $\hat{x} \sim p_G(x|z)$

and real instances x drawn from the data distribution $p_{\text{data}}(x)$. The generator network learns to "fool" the discriminator by producing samples $G(z)$ indistinguishable from real data, while the discriminator learns to recognize (and thus discourage) characteristic artifacts in the generator's outputs by mapping any instance x to an estimated probability $D(x)$ that it is real. The generator and discriminator networks are optimized jointly according to the objective

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]$$

In practice, this objective is optimized via alternating updates of the generator and discriminator network. The overall framework is flexible, and GAN architectures in the audio domain are often conditional (i.e. take additional information as input to the generator) [88], deterministic (rely solely on conditioning information rather than sampled latents z) [89], and use alternate parameterizations of the adversarial losses (e.g. least-squares [90] and feature-matching [91]). This allows for the introduction of adversarial losses into the training of virtually any generative model via a GAN discriminator, and in this sense many of the audio codec models discussed in this dissertation can be considered "GANs."

Adversarial Machine Learning leverages a core training mechanic of GANs – shaping the inputs of a neural network to optimize an objective defined on its corresponding outputs – to steer the behavior of neural network-based AI systems outside of generative modeling contexts. Just as the generator network in a GAN is trained to produce outputs that elicit a desired response from the discriminator network, data instances can in general be optimized via gradient descent or gradient-free methods to produce specific outputs from a (possibly unseen) neural network [92, 93]. These optimized inputs are referred to as *adversarial examples*, and are often studied in the context of malicious attacks on deployed AI systems. For instance, an attacker may optimize an image to fool a vision neural network performing AI-generated image detection [94]; optimize an audio waveform to conceal hidden commands for a speech recognition network operating within a

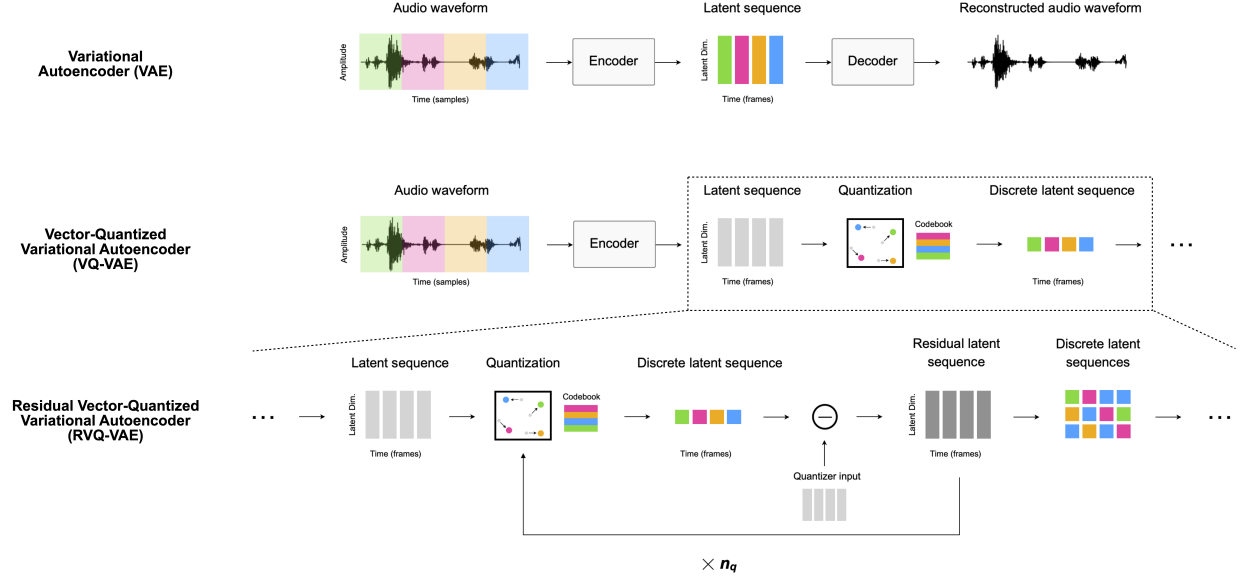


Figure 2.2. Illustration of common neural network-based latent representations used for generative modeling of audio. VAE latents are sequences of continuous vectors, VQ-VAE latents are sequences of integer indices into a codebook, and RVQ-VAE latents are stacked sequences of integer indices into n_q residual codebooks.

voice assistant [95, 96]; or realize an adversarial example as an inconspicuous physical sticker that hides a stop sign from the view of a video-processing network within an autonomous self-driving system [97]. Conversely, the adversarial interventions discussed in this dissertation aim to steer the behavior of malicious or irresponsibly deployed neural network systems to beneficial ends – for instance, concealing the identity of a speaker on a phone call from a surveilling speaker recognition system [10].

Language Models are a class of generative model that represent data as a sequence of *tokens*. These tokens may be discrete, in which case each token is an integer index drawn from a learned categorical distribution over the token vocabulary [98], or continuous, in which case each token is a vector drawn from a learned continuous distribution over a token space [99, 100]; however, the overwhelming majority of recent language modeling work has focused on discrete tokens. Given a trained language model, an instance can be sampled by repeatedly drawing tokens from the model to populate a complete sequence. For variable-length sequences, instances are usually generated

via "left-to-right" autoregression over tokens, while for fixed-length sequences, tokens can be predicted "out-of-order" or in parallel. Language models have been most widely used to generate text, which is naturally discrete. Symbolic music formats such as MIDI [101] are likewise discrete and are thus easily compatible with popular language modeling approaches [102, 103]. However, audio is naturally continuous, and much research has focused on the development of *tokenizers* or *codecs* that map audio to highly compressed, discrete token representations amenable to language modeling. The majority of existing audio codecs follow the (R)VQ-VAE or FSQ autoencoding frameworks discussed above, and often incorporate GAN losses to improve the perceptual quality of decoded audio [62–64, 104]. Both autoregressive [3, 72, 86, 87, 105] and parallel [78, 106] language modeling approaches have been explored for tokenized audio representations. The most popular parallel approaches are inspired by MaskGIT [78, 106, 107], an image-domain *masked language modeling* method in which special "mask" tokens are iteratively replaced with predicted tokens to "fill in" a sequence. Because multiple tokens are predicted at each iteration, the total number of inference steps is typically far fewer than would be required for pure autoregression. The generative modeling systems discussed in this dissertation include both autoregressive and MaskGIT-style language models operating on tokens obtained from audio codecs.

Diffusion Models are a class of generative model that iteratively transform samples from a prior distribution (typically an isotropic Gaussian) into samples from a data distribution through a series of intermediate mappings [108, 109]. Samples from the prior are generally interpreted as pure noise, samples from the data distribution as pure signal, and the multi-step mapping as a denoising or *reverse* process. A diffusion model is trained to undo a corresponding predefined *forward* process that maps from data to the prior (e.g., by adding noise). The forward and reverse processes can be thought of as functions of a discrete or continuous time index t , where the step $t = 0$ corresponds to fully-denoised data and the noise level increases with t . The reverse process is parameterized by a neural network $G(x_t, t)$ that, given step t and correspondingly noised instance x_t , predicts a quantity used to transition to some partially denoised instance x_{t-1} . Instances can then be sampled by using G

to iteratively refine a pure noise instance $z \sim p(z) \approx p(x_T)$ where T is the maximum noise index and $p(z)$ is the prior/noise distribution. Researchers have proposed numerous variants of this denoising process: intermediate predictions in the reverse process can be interpreted as "scores" or gradients of the intermediate noised distributions $\nabla_{x_t} \log p(x_t)$, allowing for the use of differential equation solvers for sampling at inference time [109]; the reverse prediction task can be reparameterized to target the noise, the denoised data, or some linear combination of the two [110]; the magnitude of noise at each time step can be adjusted via "noise schedules" to improve the training and sampling processes; sampling can be performed efficiently without the iterative prediction of all noise levels [111]; etc. Closely related are *Flow Matching Models*, which similarly refine samples from a prior distribution into samples from a data distribution through mappings between intermediate representations [80]. Flow matching models learn to predict a velocity field $v(x_t, t) = \frac{dx_t}{dt}$ that defines a path from the prior distribution to the data distribution conditioned on the timestep t , and perform inference by integrating the vector field over t . Theoretical distinctions aside, flow matching models closely resemble diffusion models in practice (isotropic Gaussian priors, training to predict refinement-like perturbations of interpolated noise-data instances, iterative inference processes that admit differential equation solvers). Therefore, except where otherwise noted, in this dissertation "diffusion models" refers more broadly to generative models that sample by iteratively refining continuous representations, including both diffusion and flow matching.

Just as audio language models are typically trained on discrete representations obtained from codecs rather than "raw" waveforms, many audio diffusion models are trained on continuous representations obtained from VAEs or other autoencoder variants. As in language modeling, these input representations are often called "tokens," and the title of this dissertation refers in part to this fundamental component of latent generative modeling.

The autoregressive language modeling, masked language modeling, and diffusion modeling paradigms discussed above are illustrated in Figure 2.3. In each paradigm, a *generator* neural network iteratively predicts sequences of tokens over the course of multiple "generation steps" to

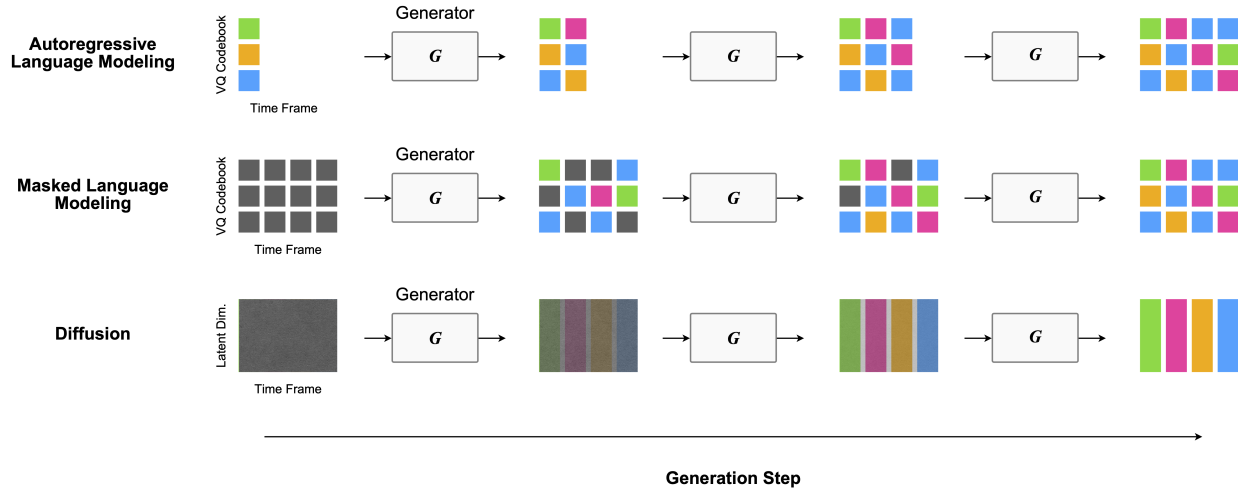


Figure 2.3. Illustrations of the three main generative modeling paradigms discussed in this dissertation: autoregressive language modeling of discrete token sequences; masked language modeling of discrete token sequences; and diffusion modeling of continuous token sequences. Each pictured method features a generator network G operating on a latent or token representation (shown as colored squares or rectangles), and samples instances iteratively over a number of generation steps. "Mask" tokens are shown in dark grey, and the diffusion timestep is represented through shading. The neural network architecture used for the generator is arbitrary, although current models typically use transformer-based [71] generators.

sample instances from the data distribution. Autoregressive language models iteratively predict tokens along the sequence order, masked language models iteratively predict tokens to fill "masked" positions within the sequence, and diffusion models iteratively denoise the full sequence of tokens. In all three cases, generated tokens can be mapped to the data space (i.e., actual audio waveforms that can be played out of a speaker) via a codec or VAE decoder to obtain the final sampled instances.

2.2 AI Interventions

The architectures, modeling frameworks, and training techniques discussed above have enabled the development of highly capable AI systems in audio and other modalities such as text, images, and video. Many of these systems serve broadly beneficial purposes, such as improving accessibility via automatic transcription and captioning of video content [1, 112], enabling low-latency speech translation [113, 114], performing noise reduction and audio enhancement within wearable devices

such as hearing aids [115, 116], and enabling novel artworks that draw on generative models to expressively transform creative materials [117]. However, these proficiencies in analyzing, generating, and manipulating digital media can also facilitate large-scale harms when AI systems are deployed irresponsibly or maliciously.

This intertwining of capabilities and harms has been a prominent topic in AI research for decades. Since the emergence of AI as a named field of study in the 1950s [118, 119], researchers have grappled with the potential impacts of highly capable machine intelligences. Alan Turing famously predicted that machines would produce text that could not be reliably distinguished from human compositions [118] and considered hypothetical social disruptions resulting from machines surpassing human intelligence [120]; Norbert Wiener warned of the risks of uninterpretable AI systems, emergent AI behaviors under misspecified learning objectives, and loss of human control [121]; and I. J. Good raised the possibility of recursive self-improvement or “intelligence explosion” in AI systems [122]. While these writings may have been shaped more by the discussions of machine consciousness in fictional works like Samuel Butler’s *Erewhon* than any observed AI harms – at the time of Wiener’s article, IBM’s checkers-playing programs represented the pinnacle of machine intelligence – they proved prescient in many ways.

In the following decades, rule-based systems demonstrated the risks of automated decision-making software long before the emergence of sophisticated artificial intelligence capabilities. Joseph Weizenbaum’s conversational system ELIZA (1966) surfaced human tendencies to anthropomorphize and attribute intent to even rudimentary pattern-matching programs [123, 124], while Dr. Geoffrey Franglen’s admissions screening system for St. George’s Hospital Medical School in London (1979) was found to have explicitly codified racial and gender biases implicit in the decisions of human admissions committee on which it was modeled [125, 126]. As automated systems have become more complex and been deployed in a broader range of contexts, the scope of realized harms has expanded substantially to include racial bias in recidivism risk prediction for criminal sentencing [127] and within more general-purpose technologies such as facial recognition

[128]; the use of realistic “deepfake” or AI-generated media to impersonate, harass, and defraud individuals and spread misinformation through digital media platforms [129]; the use of AI technologies by governments of authoritarian countries and liberal democracies alike to surveil citizens, including tracking protesters [130–133]; and the displacement of labor as AI capabilities provide rationale for cost-cutting and capital reallocation [134–136], allow for the large-scale production of goods and performance of services that compete directly with human outputs in the marketplace [137–141], and in some cases appear to support direct initiatives by employers to train on and automate employee labor [142].

Harms like these have motivated a number of *interventions* intended to curb the risks of irresponsibly or maliciously deployed AI systems. These interventions can be broadly grouped into three categories: regulatory, economic, and algorithmic¹.

2.2.1 Regulatory Interventions

Regulatory interventions aim to mitigate AI harms by providing external governance frameworks, including enforceable legal mechanisms, to shape the conditions under which AI systems are developed, deployed, and used. These interventions include the use of existing laws to combat misuses of AI technologies such as unauthorized facial recognition [143–145]. Governments have also adopted a range of legal measures specifically targeting harmful uses of AI systems. For example, Article 22 of the EU’s General Data Protection Regulation (2016) grants “data subjects” protections from profiling via automated decision-making systems [146], and the city of San Francisco has imposed a ban on the use of facial recognition technology by law enforcement (2019) [147]. In addition to prohibitions on categories of AI misuse, regulatory interventions can also provide frameworks for responsible use. For example, Canada’s Directive on Automated Decision-Making (2019) lays out guidelines for “transparency, accountability, legality, and procedural fairness” in the

¹This tripartite framing is in many ways analogous to the Legal, Market, and Technical response categories discussed by Chesney & Citron [129], while encompassing regulatory interventions beyond legal action and economic interventions beyond services anticipated to fill emergent market niches.

use of AI systems by the federal government [148]. Similarly, the United States National Institute of Standards and Technology’s Artificial Intelligence Risk Management Framework (2023) provides a voluntary framework for mitigating AI harms within public and private organizations [149]. And at smaller scales, artist groups have investigated governance structures for the use of their data within AI systems [150].

Of particular relevance to this dissertation are regulatory interventions concerning AI-generated media. The emergence of realistic image, video, and audio deepfakes and their use for unambiguously malicious purposes such as nonconsensual pornography has sparked public outcry [151, 152], instigated lawsuits against the providers of AI systems [22], and prompted the introduction of numerous state laws [153]. Partly in response to the rising prevalence of AI-generated media, the Biden administration issued Executive Order 14110, which set policy goals including provenance tracing for AI content and the protection of civil liberties in the face of AI capabilities for profiling and surveillance [154, 155] (and which has since been rescinded by the Trump administration [38]). While the legal apparatus continues to adapt to changing AI capabilities and harms [156], the diffusion of AI technologies across society and the wide ranges of uses they serve makes broadly prohibitive regulation difficult to enforce and justify [129], and has led legal scholars to favor more cautious approaches leveraging civil liability frameworks, carefully scoped criminal enforcement via existing statutes, and the delegation of AI-generated content moderation to digital media platforms [129]. Reflecting this reality, private entities and communities have enacted a range of regulatory interventions of their own: major digital media platforms have adopted disclosure requirements for AI-generated content [24, 25], and in a growing patchwork of commercial prohibitions, digital and physical art marketplaces from BandCamp to comics conventions have instituted bans on the display and sale of AI-generated works [16, 20].

Existing regulatory interventions have run up against significant obstacles in mitigating the harms posed by AI-generated media. The lack of a stable political consensus on AI regulation has led to conflicting policies, with the Trump administration rescinding Biden-era measures and

seeking to challenge state-level AI laws [38, 157]. Without robust technical solutions for identifying AI-generated content, regulatory interventions centered around prohibition and disclosure can be rendered toothless in the face of noncompliant actors who take even rudimentary steps to conceal usage [158, 159]. Regulatory approaches that intervene post-hoc after the generation of AI media are also necessarily limited in their ability to curtail misuse before it spreads widely [129]. Additionally, many harms stemming from AI-generated media lie beyond the reach of prohibition, restriction, or guidance, and instead concern the capture and distribution of economic value [160, 161]. To address these limitations, researchers have proposed a number of *economic interventions* and *algorithmic interventions*, discussed in the following sections.

2.2.2 Economic Interventions

Economic interventions aim to measure, attribute, and distribute the value captured (and costs incurred) by AI systems fairly. The attribution of value is most commonly approached through the lens of data, as modern AI systems derive their capabilities from the ingestion of large volumes of data often scraped from the public internet without consent of or payment to the authors of this data [19, 162]. Researchers have sought to highlight the harms of this appropriation and proposed interventions such as “opt-out” mechanisms that allow authors to withdraw their data from AI training [163], licensing agreements to provide fixed payments to authors of data in return for its use [27], and levies or payments extracted from AI developers for unlicensed use of data [163].

Proposed economic interventions face a number of unresolved issues. One-time payments from licensing agreements may be disproportionate to the value of data and insufficient to support data authors [29]. Training prohibitions may inadvertently serve to exclude data authors from evolving cultural contexts as AI media grows more prevalent, cutting them off not just from revenue but from exposure to new audiences [29]. Moreover, opt-outs may be legally complicated where AI systems are trained on licensed catalogs, e.g. for song-generation systems trained by major record labels [164, 165]. One method for addressing these issues is a royalty-based payment framework

in which revenue generated by an AI system is distributed “pro-rata” (or proportionally) to data authors based on the usage of that data to generate the outputs from which revenue is derived [29]. While preliminary research has explored training data attribution for AI music generation systems, further work is needed to build production-grade attribution systems capable of operating within commercial AI services [28, 29, 166, 167].

In addition to accompanying regulatory interventions that ensure adoption and provide enforcement frameworks, economic interventions may require technical methods to facilitate accurate credit assignment. For example, royalty-based approaches rely on attribution mechanisms to determine which data affect any given system output, including machine learning techniques for estimating the influence of training data [168, 169] and inference procedures that explicitly ground generation in the retrieval and mimicry of examples from a licensed corpus [166, 167]. In the following section, I discuss the development of such technical methods for intervening within AI systems and enabling external regulatory and economic interventions.

2.2.3 Algorithmic Interventions

Whereas regulatory and economic interventions are concerned with the construction, maintenance, and evaluation of frameworks surrounding AI systems, algorithmic interventions aim to shape the internals of AI systems themselves or otherwise leverage machine learning techniques to disrupt, identify, constrain, or influence the operation of AI systems [9, 30–33].

As researchers have uncovered failure modes and risk factors within AI system design including fragility, reward-hacking, and misalignment, they have sought to develop technical methods for addressing these factors at the algorithmic level [170–172]. These methods include training schemes to improve the robustness of neural network-based AI systems to maliciously perturbed inputs [173, 174], training schemes that protect the privacy of data subjects [175, 176], bias mitigation measures that promote fair outcomes in the operation of AI systems [177, 178], and techniques for embedding desirable values and behaviors within AI systems [172, 179, 180].

As mounting harms of irresponsible and malicious AI uses have pushed researchers to move beyond neutral “accident-prevention” framings [170] of this work, three categories of algorithmic intervention have emerged: *adversarial interventions* (discussed in Chapter 3) that impede the function of harmful AI systems [9, 30, 31], *cooperative interventions* (discussed in Chapter 4) that augment the function of preexisting AI systems with safeguards, and *design interventions* (discussed in Chapter 6) that operate within AI systems to steer them to more beneficial ends. Crucially, these interventions can function not just as standalone harm-mitigation measures, but as tools for regulatory and economic interventions (e.g., adversarial mechanisms to support use prohibitions, data attribution algorithms to support compensation schemes, and watermarking algorithms to support provenance tracing requirements).

This dissertation presents original research into the development and evaluation of algorithmic interventions for AI systems operating in the audio modality. In the following chapters, I examine interventions spanning the aforementioned three categories and chart a path towards proactive, intrinsic algorithmic interventions that serve as more than “token fixes” for real-world AI harms.

CHAPTER 3

ADVERSARIAL INTERVENTION

In this chapter, I present VoiceBlock, a novel adversarial intervention that modifies voice communications in real-time to circumvent AI systems performing biometric identification. This work was published at NeurIPS 2022 as VoiceBlock: Privacy through Real-Time Adversarial Attacks with Audio-to-Audio Models [10], and is presented in edited form here with additional discussion and analysis. While I proposed and led the development of VoiceBlock, it is important to acknowledge that this work was conducted in collaboration with Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo. Similarly, I led earlier work establishing the adversarial filtering method by which VoiceBlock modifies audio data, published at ICASSP 2022 as Effective and Inconspicuous Over-the-Air Adversarial Examples with Adaptive Filtering (ICASSP 2022) [181], in collaboration with Pranjal Awasthi, Aravindan Vijayaraghavan, and Bryan Pardo. I therefore use the pronoun “we” when discussing research contributions in this chapter.

VoiceBlock circumvents surveilling AI systems by interfering with the underlying deep neural networks used by these systems to process audio data. As discussed in Section 2.1.3, deep neural networks are known to be vulnerable to adversarial examples – data instances (e.g. a recording of “Bob” speaking) modified to cause a model to make an incorrect prediction (the recording’s speaker is labeled as “Maria”) [182]. This presents opportunities for privacy-minded individuals to mislead or evade neural network-based biometric surveillance systems with adversarially-crafted inputs.

Researchers have proposed adversarial attacks against a variety of audio systems relevant to biometric surveillance [183, 184]. To fool a given neural network-based system, many audio attacks modify a recording by adding a perturbation signal directly at the waveform representation [185].

This perturbation is typically crafted using gradient information obtained from the system’s neural network—or from a similarly-constructed surrogate—and requires iterative optimization. Once the optimization is complete, the modified recording is played to the system in an effort to induce arbitrary incorrect predictions (an *untargeted* attack) or a specific incorrect prediction chosen by the attacker (a *targeted* attack).

In theory, such attacks might allow individuals to evade systems that surveil and analyze voice communications, thereby protecting privacy. However, many existing attack algorithms require a costly optimization for each audio recording to which they are applied. Continuous, real-time voice communication precludes the adoption of such approaches through online optimization (which is typically too slow) or through the use of a set of precomputed adversarial examples (which would necessarily limit the user’s interaction). This suggests a need for algorithms capable of modifying speech on-the-fly. Additionally, standard attack methods introduce audible degradations to audio signals, potentially harming the utility of voice communications.

Recent years have seen the development of models for audio-to-audio tasks such as denoising, voice conversion, and musical timbre transfer [186–188]. Such models have sufficient expressive power to modify audio in coherent and task-specific ways, and are often designed to run in real-time on device. Inspired by these models, VoiceBlock leverages a deep neural network that learns to apply a time-varying finite impulse response (FIR) filter to outgoing audio, producing highly effective adversarial perturbations on a small fixed delay suitable for streaming. VoiceBlock advances beyond prior work by (1) obfuscating a user’s identity from voice-based biometric surveillance in real-time and (2) preserving the perceptual quality of de-identified voice communications, offering a more practical approach to combating biometric surveillance.

3.1 Preliminaries

In this section, I provide background information necessary to ground the research contributions embodied in the VoiceBlock system.

3.1.1 Biometric Surveillance of Voice Communications

VoiceBlock targets AI systems designed to facilitate mass biometric surveillance. Mass surveillance of voice communications is an ongoing and pervasive issue. While section 702 of the United States Foreign Intelligence Surveillance Act allows the government to perform targeted monitoring of foreign communications, bulk collection practices have resulted in the warrantless surveillance of large numbers of "incidental" foreign and domestic individuals [189]. Despite the fact that millions of these communications are obtained without warrants, they have been used in ordinary criminal investigations [190], undermining a core purpose of the Fourth Amendment of the US constitution [191]: to protect the people against searches without probable cause. Many corporations also possess the capability for large-scale collection of voice data [192], which may be leveraged to profile users for advertising or accessed by government entities through upstream and downstream surveillance [193].

Collected voice communications can be used to build biometric profiles of individuals, enabling automated identification of speakers from voice communications such as phone calls. The use of such biometric profiling technology has spread beyond authentication systems to surveillance applications [194]. For instance, some prison systems have mandated the enrollment of biometric profiles for both inmates and those participating in phone calls with them [195, 196]. As individuals are faced with these growing surveillance apparatuses, it is worth remembering that routine surveillance of private voice communications is also corrosive to free speech and association and tends to disproportionately affect marginalized groups [197, 198].

3.1.2 Speaker Recognition Systems

VoiceBlock intervenes against a specific category of audio AI system: *speaker recognition*. In the absence of identifying metadata, speaker recognition systems can facilitate mass surveillance by allowing an operator to search a database of recorded voice data for utterances from a chosen

speaker [199] or to diarise (assign utterances to individuals) transcripts of recordings [200]. Prior to the advent of automatic speaker recognition, these tasks required human analysts, forming a natural check on surveillance overreach. Through VoiceBlock, we seek to restore this check by degrading the efficacy of speaker recognition models while maintaining the original perceptual quality of the voice communication, a step that could grant users a measure of privacy from mass surveillance.

As discussed in Section 2.1.3, modern speaker recognition systems rely on deep neural networks to construct and cross-reference biometric profiles from speech utterances. These networks are trained to map speech utterances to a low-dimensional embedding space in which distance corresponds to speaker similarity [201]. In this work, we consider models f for which the speaker distance D_f between utterances u and v can be measured via a cosine distance between embeddings $f(u)$ and $f(v)$:

$$D_f(u, v) = 1 - \frac{f(u) \cdot f(v)}{\|f(u)\|_2 \|f(v)\|_2} \quad (3.1)$$

At inference time, *query* audio from the user is embedded and scored for similarity against *enrolled* utterances – pre-computed embeddings stored in the database. Scores may be evaluated for all enrolled embeddings, or for only a representative of each unique speaker (e.g. speaker centroids). The highest-ranking result or results (e.g. the closest speaker identity) are then returned.

A variety of methods have been proposed to efficiently compute and search over low-dimensional speaker representations. Generally, a hashing algorithm is applied to speaker embedding vectors to reduce storage and search costs [201–203]. Because the resulting hash representation merely serves as an efficient point of access for embedding-space distances, we omit hashing algorithms from consideration and construct our adversarial intervention to target the speaker embedding space directly.

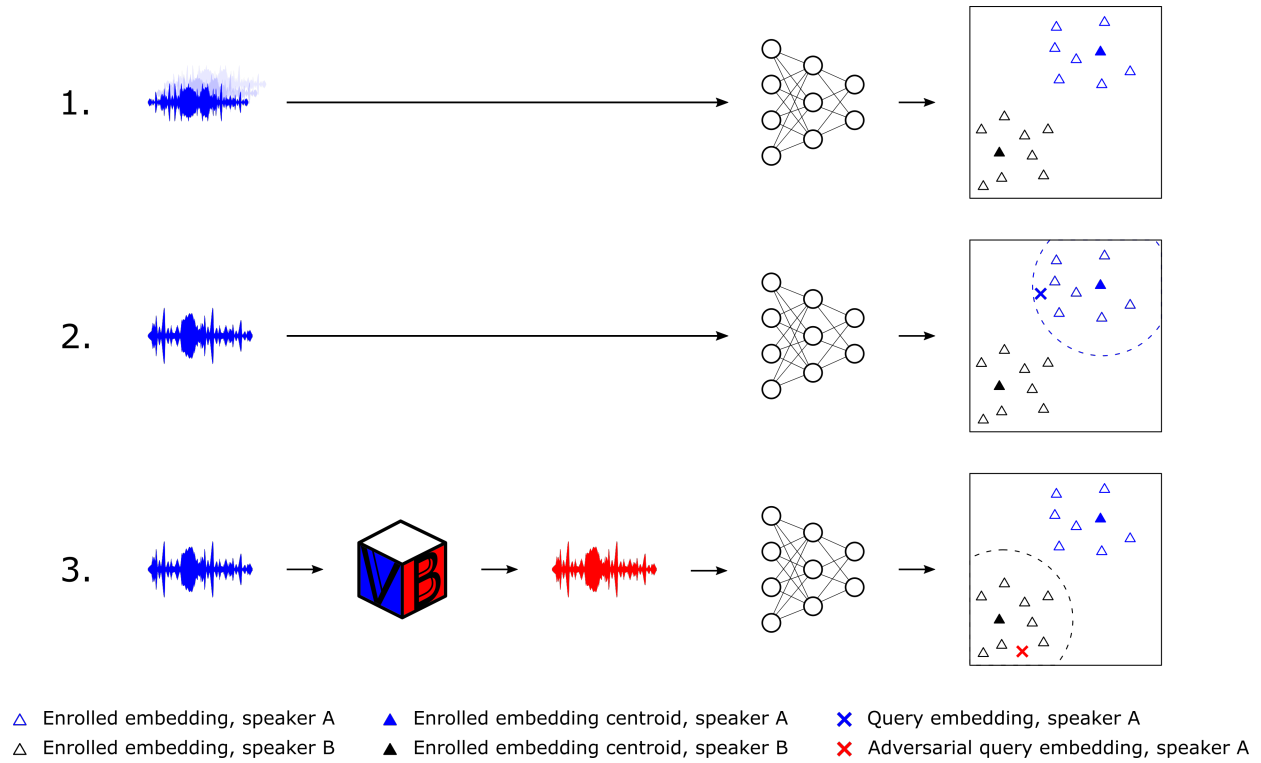


Figure 3.1. Illustration of biometric profiling via speaker recognition systems and the effect of the proposed VoiceBlock method. **1.** Speaker recognition systems use neural networks to map speech utterances to an embedding space in which proximity corresponds to perceptual speaker similarity. **2.** Given enrolled *profiles* of speakers consisting of labeled embedded utterances, the speakers of new *query* utterances can be identified by embedding and checking the stored label of nearby utterances. **3.** VoiceBlock intervenes by perturbing a user's audio stream such that any captured utterances will be mapped to a distant region of a speaker recognition system's embedding space.

3.1.3 Prior Work

Previous works in speech de-identification have proposed methods for obfuscating speech to evade surveilling speaker-recognition systems. One such method is voice conversion, which modifies the speech of a source speaker to sound like that of another target speaker, both to humans and machines. Jin et al. [204] and Alegre et al. [205] analyze the vulnerability of speaker recognition systems against de-identification attacks performed with voice conversion models. However, the conversion models considered are incapable of real-time operation. While the recent availability of low-latency voice conversion models [187, 206, 207] may make such approaches more practical, our aim is to modify speech in a way that is inconspicuous and minimally invasive to the user experience, and that has the potential to generalize to tasks beyond speaker recognition. This also rules out recent anonymization approaches cataloged through the VoicePrivacy Challenge, as all evaluated submissions to date noticeably alter speaker characteristics [208]. Similarly, we do not consider obvious transformations like pitch-shifting that significantly alter or degrade recorded audio.

Separately, real-time audio attacks have recently been explored. Chiquier et al. [209] propose an attack on automatic speech recognition systems that is capable of inducing significant transcription errors in an *over-the-air* environment (where attack audio is played to the system in a physical space), but which introduces conspicuous noise due to the use of additive perturbations at the audio waveform. This approach also necessitates both a large model (requiring approximately 16 GPU-days to train) and lengthy receptive field, resulting in an initial “idle” period of 2.5s during which the attack does not affect speech. By contrast, we focus on an *over-the-line* setting, where attacks are passed to the victim model over a digital channel, and propose a lightweight network capable of producing inconspicuous perturbations through the use of filtering, with an initial delay of milliseconds rather than seconds.

Universal adversarial attacks optimize a short perturbation that can be played alongside arbitrary

speech, optionally in real-time, to evade a victim model [183, 210, 211]. However, these attacks tend to introduce conspicuous noise due to the use of additive perturbations at the audio waveform. The same holds for the attack of Xie et al. [212], which uses a Wave-U-Net [213] model to efficiently generate universal perturbations. Rather than generate additive perturbations, we perform multiplicative attacks in the frequency domain through the use of filtering, and in doing so avoid the bias towards noise-like artifacts typical of additive attacks.

Other works have proposed crafting adversarial attacks in the frequency domain. The “Kensville” attack of Abdullah et al. [214] performs spectral gating, removing low-energy frequency components in an effort to hinder classification. The authors demonstrate the effectiveness of their approach against seen and unseen speaker- and speech-recognition systems, and the proposed method is fast and gradient-free. However, the spectral gating must be constrained to a small number of key audio frames to avoid conspicuous artifacts, which requires offline optimization using word- or phoneme-level alignment information. Ahmed et al. [215] optimize bandpass filters to perform targeted impersonation attacks on speaker-recognition models; these filters can then be physically realized as resonant tubes through which an attacker can speak to the victim system, allowing real-time operation. However, once optimized and realized, the bandpass filters are fixed in time and cannot adapt to the attacker’s speech. The attack is significantly less effective than traditional approaches, even in controlled acoustic environments, and requires an active intervention on the part of the user to both realize and apply.

In summary, VoiceBlock aims to improve over prior approaches that lack the capability to simultaneously perform voice de-identification and remain inconspicuous to human listeners while operating in real-time on-device.

3.2 Method

VoiceBlock adversarially modifies recorded speech by specifying the frequency magnitude response of a finite impulse response (FIR) filter. By carefully varying the filter’s response many times per

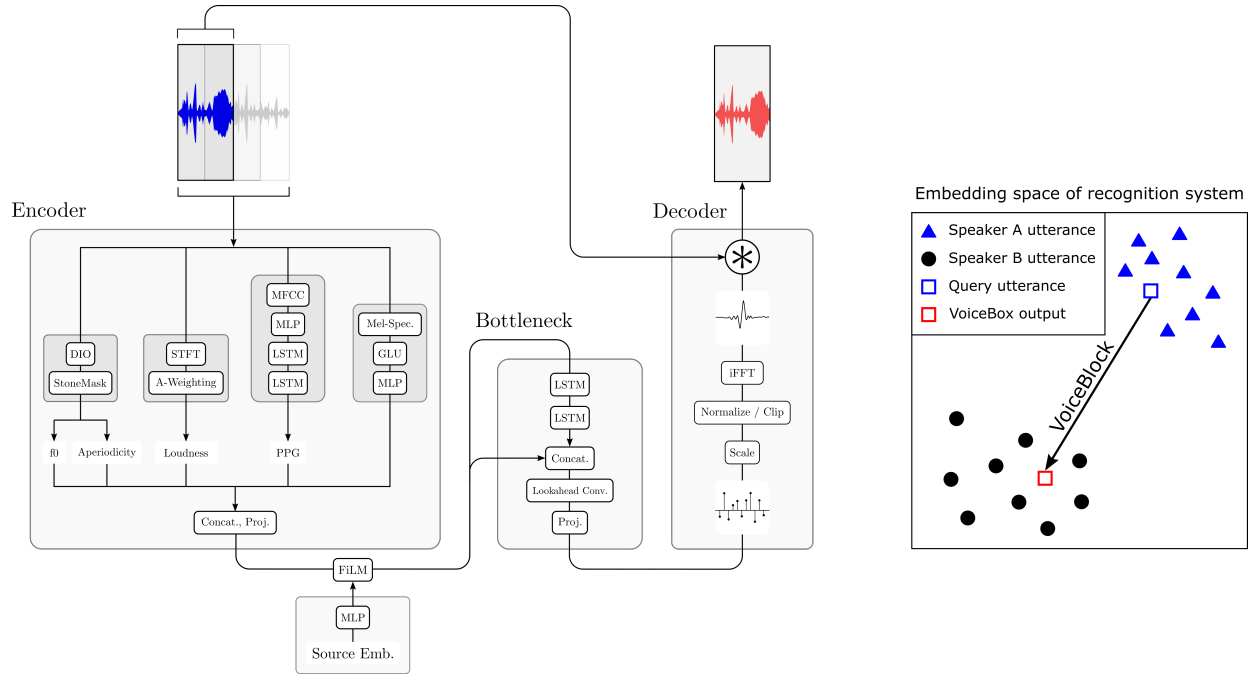


Figure 3.2. **Left:** The proposed VoiceBlock architecture. Acoustic features extracted by the encoder are fed to the recurrent bottleneck to predict filtering controls, which are regularized and applied to the input by the decoder to obtain adversarial audio. **Right:** VoiceBlock adversarially perturbs the user’s audio stream such that any extracted queries are scored by a speaker recognition system as dissimilar to the user’s enrolled utterances, hampering identification or retrieval.

second, the speech is de-identified to an automated system while speaker identity is preserved for human listeners. This is accomplished without introducing noisy artifacts at the waveform, as filtering can only amplify or attenuate frequency energy present in the original signal through multiplication in the Fourier domain (see Section 3.2.3 for details of the filtering implementation used in our experiments).

VoiceBlock consists of three main modules, shown in Figure 3.2: (1) an *encoder* module extracts acoustic features from input audio frames, (2) a recurrent *bottleneck* module incorporates context from past frames to predict a set of adversarial filter controls for each frame, and (3) a *decoder* module regularizes the predicted filter controls and applies them to the corresponding frame of input audio. We provide an overview of each component in the following sections.

3.2.1 Encoder

VoiceBlock’s encoder module extracts acoustic features from 16kHz audio (double the sample rate of telephone speech) segmented into frames of 256 samples with 50% overlap. Though VoiceBlock can operate in fully causal fashion, we find we achieve stronger attacks using a lookahead of five frames (see Section 3.2.2), resulting in a theoretical latency of 56ms. For reference, Verizon reports 25ms or less is typical latency for voice-over-IP communication in North America and Europe [216], while experimental data [217] show that 130ms of latency is acceptable for high-quality voice communications. Therefore, adding VoiceBlock to the signal chain leaves latency well within this limit.

Audio frames entering the encoder are passed to four sub-modules in parallel to obtain the following features:

- *Pitch features*: Given the known sensitivity of speaker-recognition models to pitch [215] and the importance of spectral structure in delineating linguistic content, we extract fundamental frequency and aperiodicity estimates for each frame using the DIO algorithm [218]. Pitch estimates are refined by the StoneMask algorithm [219]. For both stages, we use the `pyworld` [220] implementation of the WORLD vocoder [219].
- *Loudness*: We take an A-weighted average of each frame’s log-magnitude spectrum to obtain a loudness estimate [221].
- *Phonetic posteriorgrams*: Following the method of Ronssin & Cernak [187], we use a trained phoneme classifier with frozen weights to encode linguistic content. The classifier consists of a multi-layer perceptron followed by two LSTM [222] layers and a linear classification layer, and takes as input 13 mel-frequency cepstral coefficients (MFCC) with first- and second-order deltas for each frame. We train our phoneme classifier on the "train-clean-100" subset of the LibriSpeech dataset [223] with frame-aligned phoneme labels [224, 225] for 25 epochs using

the Adam optimizer [226] and learning rate $1e-3$. Rather than directly use the classifier’s predicted distributions over phoneme labels – known as phonetic posteriorgrams (PPGs) [227] – we discard the classification layer after training and pass along the output of the final LSTM layer, which Ronssin & Cernak also refer to as PPGs in their architecture.

- *Spectrogram features*: Finally, we pass 64-bin mel-spectrograms through a 1×1 convolutional layer and gated linear unit (GLU) operating along the frequency axis, followed by two linear layers with leaky ReLU activations and hidden size 512. The convolution layer upsamples the channel (frequency) dimension to 1024, from which the GLU obtains a 512-dimensional representation to pass to the linear layers. This small network captures residual information not present in the coarse-grained pitch, loudness, and phonetic features. We note that Engel et al. [188] also make use of a spectrogram encoder, alongside pitch and loudness features, to control differentiable signal-processing components.

For each frame of audio, we concatenate the 256-dimensional output of the phoneme encoder, the 512-dimensional output of the spectrogram encoder, and the one-dimensional pitch / periodicity / loudness features and linearly project to obtain a 512-dimensional vector.

We then introduce additional speaker information by passing a pre-computed 512-dimensional embedding of the source speaker through a FiLM layer [228] to modulate the encoder output. This embedding helps to guide the de-identification task, allowing us to train a single model capable of de-identifying arbitrary users. This imposes a requirement that users must record a small amount of speech to obtain an embedding before first using VoiceBlock. We find that less than a minute of speech is sufficient, and use a pre-trained ResNetSE34V2 model [229] to compute embeddings (see sections 3.3.1, 3.3.2).

3.2.2 Bottleneck

Encodings entering VoiceBlock’s bottleneck are passed through two LSTM layers with hidden size 512, and the outputs are concatenated with a skip connection from the encoder. To enable streaming, we use unidirectional LSTM layers and pass concatenated outputs through a small lookahead convolutional network [230] to incorporate information from future frames at the expense of a small fixed delay. We find a lookahead of 5 frames (48 ms) is sufficient to craft strong de-identification attacks. Finally, a linear projection layer maps the concatenated representations to a vector of 128 filter controls, with each element representing the unnormalized frequency magnitude response of a filter band for the current frame.

3.2.3 Decoder

VoiceBlock’s decoder applies time-varying filtering to the input audio based on frame-wise controls obtained from the bottleneck. We find that a filtering-based decoder affords a number of advantages over the synthesis-based (e.g., transposed-convolutional) decoder architectures present in many audio-to-audio models [186]:

- The filtering module is not prone to the periodic upsampling artifacts that transposed-convolutional architectures often introduce [231].
- Our decoder is a simple deterministic module with no trainable parameters, keeping VoiceBlock lightweight.
- The capacity of the decoder and conspicuousness of perturbations can easily be constrained in terms of the number of filter bands or their allowed range of motion, providing interpretable control to the user.

To regularize and apply filter controls to each frame of audio, we use the method proposed in my previous work Effective and Inconspicuous Over-the-Air Adversarial Examples with Adaptive

Filtering [181]. We bound filter controls to the range $[-0.5, 0.5]$ to constrain deviations from a "neutral" control configuration of $\vec{1}$. Each set of scaled filter controls is transformed into a time-domain impulse response via the inverse Fourier transform. We shift the impulse response to zero-phase (symmetric) form, apply a Hann window, and finally convolve with the corresponding input audio frame by taking the Fourier transform and performing element-wise multiplication. After compensating for shift, we overlap-add the resulting frames with a Hann window to obtain the final filtered audio. Notably, while the filtering attack developed in my prior work operates offline via an optimization loop, VoiceBlock adapts this method for real-time, online adversarial modification of audio. Figure 3.3 shows an example of VoiceBlock’s predicted filter controls applied to speech.

VoiceBlock is implemented in PyTorch [232] and contains 6.3m trainable parameters, and 7.5m in total, counting the frozen phoneme encoder.

3.2.4 Training objective

To demonstrate the ability of VoiceBlock to perform inconspicuous privacy-preserving audio transformations, we train our model to attack speaker recognition systems. Given a speaker embedding model (as described in Section 3.1.2), we aim to modify query audio drawn from a user’s stream such that its embedding-space distance from any enrolled utterances of the user is large, and its evaluated similarity is small. We quantify this de-identification in terms of distance thresholds in the embedding space, set according to percentiles of the estimated distribution of all inter-speaker embedding distances. Let P_r represent the distance corresponding to the r^{th} percentile of this distribution; then for query audio u and enrolled utterance v , $D_f(u, v) > P_r$ implies that roughly r percent of database entries should be scored as more similar to u than v . Thus, we can set distance thresholds that correspond directly to the strength of de-identification applied to user audio, and construct a loss function that penalizes our model when the embedding-space cosine distance between query and enrolled utterances falls below the threshold. To do so, we use a variant

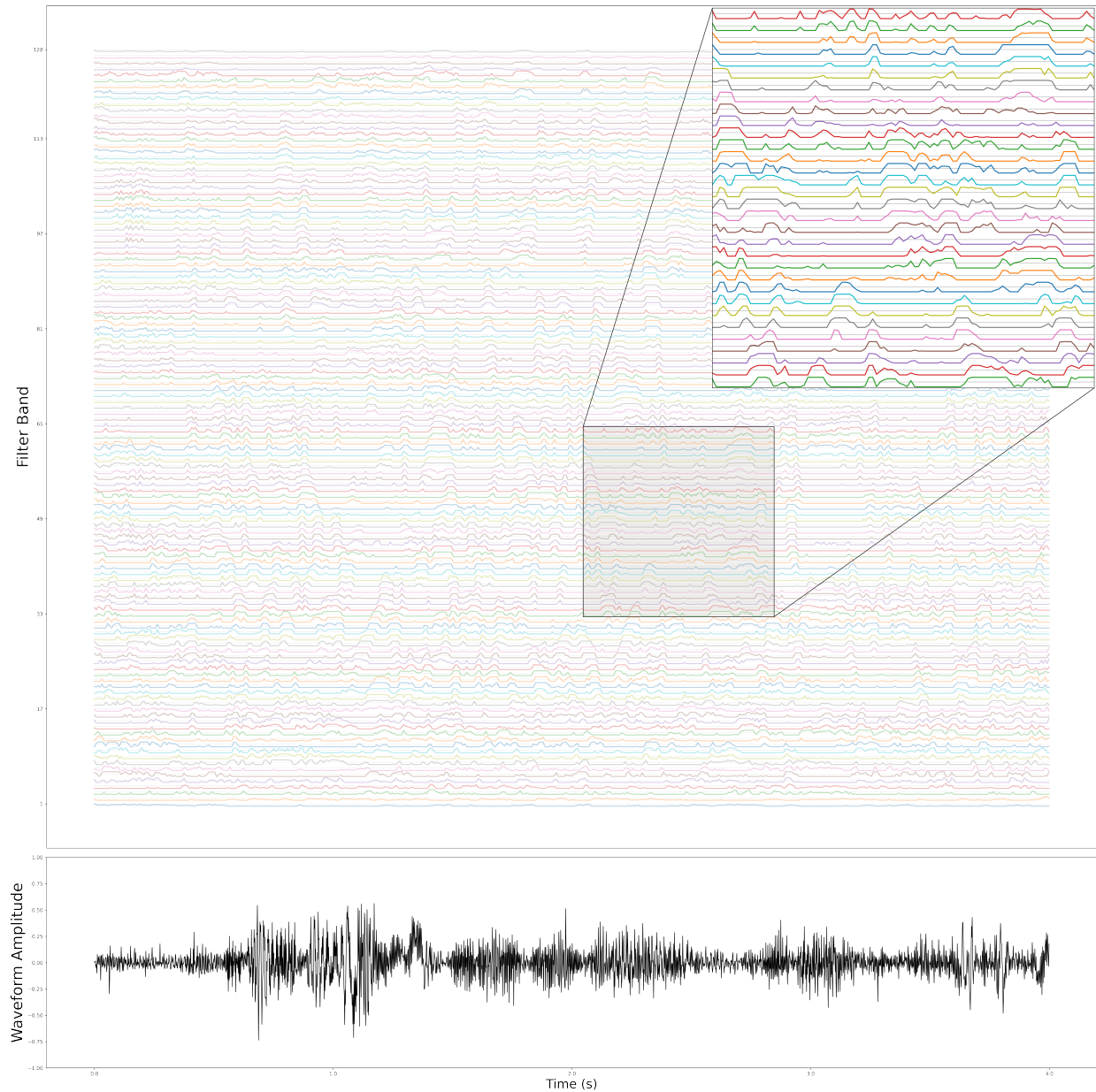


Figure 3.3. Time-varying magnitude responses of 128-band filter controls predicted by VoiceBlock (top) to de-identify a given audio recording (bottom). Filter bands are arranged vertically by frequency, and the frequency magnitude response of each band over time is represented by the corresponding curve. The thin grey line beneath each curve represents a "neutral" frequency magnitude response of 1, with deviations above and below this line representing boosts and cuts, respectively. Note that the controls vary smoothly, and often exhibit sparseness in quiet regions of the waveform.

of the adversarial loss proposed by Zhang et al. [183]. Let f represent the victim model, g our VoiceBlock network, u an utterance from our user’s audio stream, and P_r our de-identification threshold; then

$$\mathcal{L}_{adv}(f, g, u) = (P_r - D_f(g(u), u) + \kappa)^+ \quad (3.2)$$

where $(\cdot)^+ = \max(\cdot, 0)$ and κ is a confidence parameter encouraging the attack to fully cross the threshold. To ensure that our VoiceBlock model learns to perturb user audio inconspicuously, we incorporate an additional loss function to penalize perceptible filtering artifacts. We compute the combined waveform L_1 and multi-resolution spectrogram losses proposed by Defosséz et al. [186] on the clean and adversarial audio:

$$\mathcal{L}_{aux}(g, u) = \mathcal{L}_{stft}(u, g(u)) + \|u - g(u)\|_1 \quad (3.3)$$

where $\mathcal{L}_{stft}$ is given by a sum of magnitude and spectral convergence losses computed over several spectrogram resolutions:

$$\mathcal{L}_{stft}(a, b) = \sum_{i=1}^M \left[\mathcal{L}_{sc}^{(i)}(a, b) + \mathcal{L}_{mag}^{(i)}(a, b) \right] \quad (3.4)$$

$$\mathcal{L}_{sc}(a, b) = \frac{\| |STFT(a)| - |STFT(b)| \|_F}{\| |STFT(a)| \|_F} \quad (3.5)$$

$$\mathcal{L}_{mag}(a, b) = \| \log |STFT(a)| - \log |STFT(b)| \|_1 \quad (3.6)$$

Here $\| \cdot \|_F$ and $\| \cdot \|_1$ refer to the Frobenius and L_1 norms, respectively. Following Defosséz et al., we use $M = 3$ spectrogram resolutions with hop sizes of 50, 120, and 240 samples, FFT lengths of

512, 1024, and 2048, and window lengths of 240, 600, and 1200, respectively.

Combining the above adversarial and auxiliary losses, we obtain our final training objective:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{aux} \quad (3.7)$$

3.2.5 Streaming Implementation

To demonstrate the ability of our proposed system to modify voice communications in real-time, we implement a streaming configuration of VoiceBlock. All components are modified to support the computation of perturbations on overlapping windows of audio while the input and output streams process non-overlapping chunks. Where possible, components are complied to TorchScript¹ to improve performance. Three buffers are added to process a stream of audio into overlapping chunks:

- Last Frame Input Buffer: Contains the last 128 samples of the last input audio to be appended to the oncoming chunk, before computing the overlapping windows.
- Lookahead Buffer: Contains overlapping frames to be used as lookahead.
- Output Buffer: Contains the last 128 samples output by VoiceBlock, to be overlap-added with the next processed chunk.

We provide code for the proposed VoiceBlock system – including training, inference, and the streaming implementation – as well as listening examples on a supplementary webpage².

3.3 Experiments

We next describe experiments used to validate the claimed advances of our work, namely that VoiceBlock can de-identify speech in real-time while remaining significantly less conspicuous

¹<https://docs.pytorch.org/docs/main/jit.html>

²<https://interactiveaudiolab.github.io/project/voiceblock.html>

than existing methods of similar effectiveness. We first introduce the models, datasets, and attack algorithms considered in our experiments. Following this, we detail our experiment configurations and present the results of both objective (de-identification performance) and subjective (human listening) evaluations.

3.3.1 Models

We train VoiceBlock and baseline de-identification attack methods against the ResNetSE34v2 model [229] provided in the VoxCeleb Trainer repository [233]. The model takes mel-spectrogram inputs and uses 2D convolutions with residual connections, squeeze-and-excitation, and attentive statistics pooling to generate frame-level features and aggregate them into 512-dimensional speaker embeddings. To examine the transferability of our approach against unseen systems, we additionally evaluate trained attacks against the Y-Vector model proposed by Zhu et al. [234]. The model uses a multiscale 1D-convolutional waveform encoder to extract acoustic features, followed by squeeze-and-excitation blocks, feature aggregation, and a time-delayed neural network to map variable-length utterances to 128-dimensional speaker embeddings.

Both the ResNetSE34v2 and Y-Vector models were trained on the development set of the VoxCeleb2 dataset [235]. Note that while our attack is causal—modulo a short, five-frame lookahead—we do not impose the same restriction on a hypothetical surveillance system; instead, we evaluate our attack against strong, non-causal systems capable of aggregating speaker characteristics from across full utterances before rendering predictions.

3.3.2 Data

We use both the `train-clean-100` and `test-clean` subsets of the LibriSpeech dataset [223] for training VoiceBlock. The former comprises 28,539 utterances from 251 speakers while the latter comprises 2,620 utterances from 40 speakers. To simulate large-scale surveilling speaker recognition, we evaluate attacks on the VoxCeleb1 dataset [236], comprising 153,516 utterances

from 1,251 speakers. This also ensures that no evaluation speakers are seen during training. As with the LibriSpeech dataset, we trim or pad all utterances to 4 seconds.

For all experiments, we carefully divide the data to imitate a realistic attack setting. During training, we select fifteen utterances (one minute total) from each source speaker in the training set and compute embeddings using the ResNetSE34v2 model. The centroid of these embeddings is then used as an enrolled target for the computation of the adversarial loss with all utterances of that speaker. We find that this produces stronger attacks than using individual utterance embeddings as targets, possibly by ensuring more consistent gradient information across the optimization. For our VoiceBlock attack (see Section 3.3.3), we select a further ten utterances (40s total) from each source speaker in the training set and again compute embeddings using the ResNetSE34v2 model. The centroid of these embeddings is then fed as a fixed conditioning vector to the VoiceBlock model alongside all utterances from that speaker (see Section 3.2.1). Similar to training, during evaluation we select fifteen utterances per speaker as a query set. We again select a further ten utterances to serve as conditioning for the VoiceBlock attack. Finally, twenty utterances of each speaker are enrolled in the speaker recognition system, serving as the database against which query utterances are matched.

3.3.3 Attack Algorithms

We perform untargeted de-identification attacks on speaker recognition systems using the following algorithms. *VoiceBlock*: We implement the proposed VoiceBlock attack as described in Section 3.2 and train for 10 epochs on 3 NVidia RTX 2080 Ti GPUs, using the Adam optimizer with learning rate $1e-4$; this takes approximately 40 minutes. *Universal*: We optimize a short (2s) universal additive perturbation for 10 epochs using the established penalty method [183, 210, 211] and the same adversarial objective and optimization configuration as VoiceBlock. On 3 NVidia RTX 2080 Ti GPUs, this takes approximately 16 minutes. To limit the perceptibility of the attack, we scale the perturbation to have L_∞ norm 0.08 times that of the unperturbed speech. During training

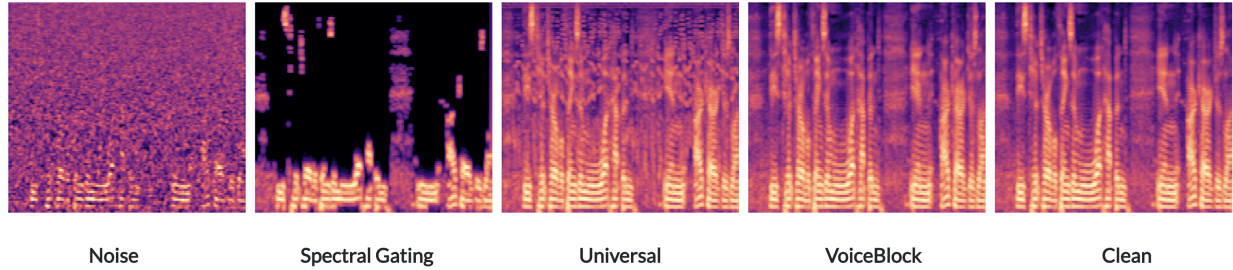


Figure 3.4. Spectrogram plots of each evaluated attack method operating on source audio (Clean, at right). Note that while Noise, Spectral Gating, and Universal attacks introduce perceptible degradations to the source audio (visible in the corresponding spectrograms), the proposed VoiceBlock attack modifies the source audio more subtly.

and evaluation, the perturbation is aligned arbitrarily with query utterances and looped to match durations, serving as a constant adversarial "background" signal. *White noise:* We add Gaussian noise to utterances at the waveform representation at a signal-to-noise ratio of -10dB . *Spectral gating:* We modify the "Kenansville" attack of Abdullah et al. [214] to allow for streaming use by performing spectral gating at all frames, using a threshold of 4dB relative to the maximum-energy spectral bin of each frame. In Figure 3.4, we provide spectrogram plots of each evaluated attack to demonstrate how each method affects the source audio.

3.3.4 Objective evaluation

We evaluate VoiceBlock and baseline de-identification attacks in a large-scale closed-set speaker recognition task and provide objective measures of de-identification strength, de-identified audio quality, and streaming performance. Additionally, we examine the performance of VoiceBlock in two challenging scenarios: neural network-based audio processing intended to "wash out" adversarial modifications to voice communications, and enrollment of adversarially-modified utterances in a speaker recognition database to re-identify VoiceBlock users.

First, we use the `test-clean` LibriSpeech subset to obtain a rough estimate of the distribution of distances between an individual utterance and the centroids of each distinct speaker in the embedding space of the ResNetSE34V2 model. To encourage strong de-identification attacks, we

take the distance corresponding to the 25th percentile of this distribution as the target threshold P_{25} for our training loss (see Section 3.2.4). We train each attack as discussed above.

We evaluate closed-set speaker recognition performance under each attack on the VoxCeleb1 dataset. We compute the distance between each query embedding and the centroid of all embeddings of each speaker; recognition is then performed by returning the speaker identity of the nearest speaker centroid. We find this is slightly more accurate and robust to attack than using the identity of the nearest embedded utterance. We perform exact nearest-neighbors search over the embedding space, and report the top-1 (T-1) and top-10 (T-10) accuracy of the relevant speaker recognition model given both clean and adversarial queries. Additionally, we compute the following objective speech quality metrics over the clean and adversarial query audio as a proxy measure of the imperceptibility of attacks: Perceptual Evaluation of Speech Quality (PESQ) [237], operating in the wide-band configuration and using the `python-pesq` implementation [238]; and Short-Time Objective Intelligibility (STOI) [239], using the `pystoi` implementation [240]. Finally, to measure the intelligibility of voice communications processed by the proposed and baseline methods, we evaluate the word-error-rate (WER) and character-error rate (CER) of audio processed with each attack on the LibriSpeech `train-clean-360` subset using a Wav2Vec 2.0 automatic speech recognition model [241]. The `train-clean-360` dataset contains utterances and accompanying transcriptions from 921 speakers not seen during training. The results of our evaluation are presented in table 3.1.

In real-world settings, the operator of a biometric surveillance system may apply various audio processing stages to enhance audio quality and remove adversarial modifications before computing embeddings with a speaker recognition model. To simulate the presence of such stages, we perform attacks using the experiment configuration discussed above but pass query utterances through a pretrained Demucs [186] speech enhancement model en route to the ResNetSE34v2 recognition system; results are reported in 3.2. The speaker recognition system maintains its top-1 and top-10 accuracy on clean queries. In contrast to Chiquier et al. [209], we do not incorporate the enhancement model into our adversarial optimization, meaning that Demucs functions as an unseen

Table 3.1. Objective evaluation results. We perform attacks on the VoxCeleb1 dataset against seen (ResNetSe34v2) and unseen (Y-Vector) speaker recognition models, and compute both top-1 and top-10 recognition accuracies. Additionally, we compute objective speech quality metrics for each attack (PESQ, STOI) and measure the intelligibility of attacked speech over the LibriSpeech train-clean-360 dataset in terms of word- and character-error rates (WER, CER) obtained from a pretrained Wav2Vec 2.0 model.

Approach	Speech Quality		Speech Intelligibility		ResNetSe34V2		Y-Vector	
	PESQ $\uparrow$	STOI $\uparrow$	WER $\downarrow$	CER $\downarrow$	T-1$\downarrow$	T-10$\downarrow$	T-1$\downarrow$	T-10$\downarrow$
White noise	1.03	0.41	1.00	1.00	0.16	0.47	0.00	0.01
Spectral gating	1.11	0.63	0.50	0.29	0.02	0.12	0.03	0.13
Universal	1.35	0.78	0.16	0.08	0.09	0.20	0.48	0.73
VoiceBlock	3.74	0.92	0.02	0.01	0.03	0.10	0.35	0.66
No attack	4.64	0.99	0.01	0.01	0.99	0.99	0.96	0.99

Table 3.2. Top-1 (T-1) and top-10 (T-10) recognition accuracy of the ResNetSE34v2 system on the de-identification attacks described in Section 3.3.3 when all query audio is passed through a Demucs [186] speech enhancement processing stage.

Approach	ResNetSe34V2		+Demucs	
	T-1 $\downarrow$	T-10 $\downarrow$	T-1 $\downarrow$	T-10 $\downarrow$
White noise	0.13	0.40	0.02	0.09
Spectral gating	0.02	0.11	0.02	0.11
Universal	0.14	0.22	0.87	0.97
VoiceBlock	0.02	0.10	0.04	0.15
No attack	0.97	0.99	0.96	0.99

adversarial defense.

Beyond protective processing, we consider an additional scenario that may thwart adversarial attacks on biometric surveillance systems. It is possible that adversarially-perturbed query utterances extracted from a VoiceBlock user’s audio stream may themselves be enrolled by a surveilling speaker recognition system. In such cases, there are two possibilities:

1. VoiceBlock successfully de-identifies user speech, and adversarial queries are enrolled as a separate speaker profile from any existing clean utterances of the user
2. VoiceBlock fails to de-identify user speech or adversarial queries are otherwise manually

Table 3.3. Effect of adversarial utterance enrollment on de-identification performance of VoiceBlock. For sets of 15 clean and adversarial query utterances, we compare the top-1 (T-1) and top-10 (T-10) accuracy of speaker recognition with the ResNetSE34v2 model under three conditions. **Clean profile:** 20 clean utterances are enrolled per speaker. **Mixed profile:** 10 clean and 10 adversarial (VoiceBlock) utterances are enrolled per speaker. **Adversarial profile:** 20 adversarial (VoiceBlock) utterances are enrolled per speaker.

	Clean profile		Mixed profile		Adversarial profile	
Query processing	T-1 ↓	T-10 ↓	T-1 ↓	T-10 ↓	T-1 ↓	T-10 ↓
VoiceBlock	0.02	0.10	0.51	0.80	0.78	0.93
None	0.97	0.99	0.83	0.95	0.09	0.28

incorporated into an existing profile of the user containing clean utterances.

We examine both cases, again using the VoiceBlock attack discussed in Section 3.3.3. Results are presented in table 3.3.

Finally, we evaluate the real-time performance of the streaming implementation of VoiceBlock described in Section 3.2.5 by measuring its average real-time factor (RTF). We measure performance on a single thread on two different CPUs, an Intel i7-5600U @ 3.2 GHz and an Apple M1 Chip. In the streaming configuration, VoiceBlock processes a chunk of 4 overlapping frames at a time, equivalent to 640 samples / 40 ms. We compute the average RTF for processing a chunk over all chunks in a 4 second audio clip, and find that VoiceBlock has an average RTF of .255 on the Intel i7-5600U and .200 on the Apple M1.

3.3.5 Subjective evaluation

We further evaluate the subjective quality of de-identified audio under processing with VoiceBlock and baseline attack algorithms via a human listening study. For our subjective evaluation, we sampled 100 clean speech recordings from the VoxCeleb1 dataset. To standardize comparisons, we trimmed or padded all utterances to 4 seconds. Each of the four attacks described in Section 3.3.3 was applied to all 100 recordings, resulting in 100 comparison sets with five recordings per set: the original clean speech and the speech modified by each of the four attacks. These sets were

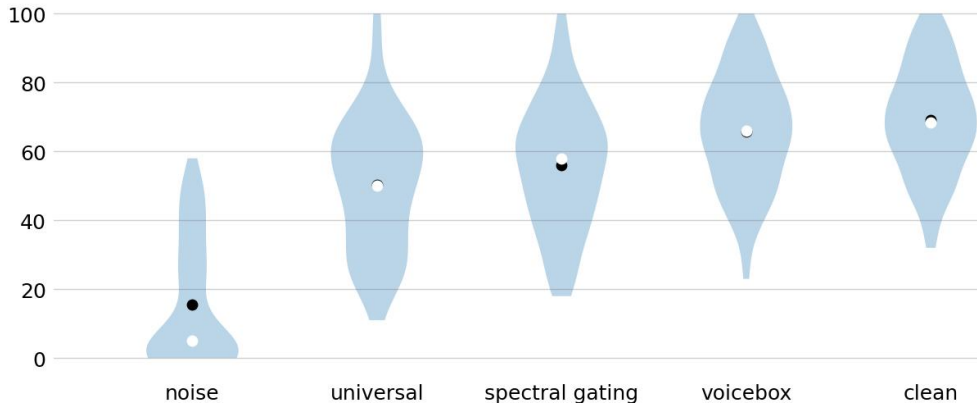


Figure 3.5. Distributions of quality ratings for VoiceBlock and baseline methods from crowd-sourced MUSHRA-style listening test. Higher numbers are better. Black dots are means and white dots are medians. Wilcoxon signed-ranked tests between all pairs of conditions show statistical significance at $p < 0.05$.

then evaluated in a MUSHRA-style listening study [242] deployed on Amazon Mechanical Turk using the open-source Reproducible Subjective Evaluation (ReSEval) [243] system. IRB approval was obtained prior to conducting this study, and there were no known risks to the participants in this study.

We recruited 20 participants. Participants were screened with a listening test prior to beginning the study; those that passed rated 20 comparison sets. For each comparison set, the participant was asked to listen to and rate the relative audio quality of each of the five audio files on a scale from 0 to 100. The instructions were as follows:

Listen to all recordings of a person speaking. Then, move the sliders to **rate the quality of each audio file from 0 (worst) to 100 (best)**. The higher-quality audio files are the ones that are more natural sounding, or have fewer audio artifacts (e.g., clicks, pops, noise, or otherwise sound ‘unnatural’). **Note**—Each slider cannot be moved until its corresponding audio file has been listened to in its entirety.

It took listeners approximately 15 minutes to complete all evaluation tasks. If a participant succeeded in the listening test and completed the evaluation tasks, they were paid 3.00 USD, or equivalently, 12.00 USD / hour. If the participant failed the listening test, they were paid 0.50 USD. We omitted responses by four participants who failed our prescreening listening test and nine who

rated the white noise attack (our low anchor) as superior in quality to ground-truth clean audio (our high anchor), giving us a total of 140 five-way comparisons.

The results of our crowdsourced subjective evaluation can be found in Figure 3.5. We find VoiceBlock is preferred to all other methods, and exhibits similar perceptual quality to ground-truth speech recordings. The difference between VoiceBlock and ground-truth audio is significant ($p < 0.05$) using a Wilcoxon signed-rank test ($p = 0.024$) but not significant using Welch’s T-test ($p = 0.071$), which assumes that human perceptual scores are normally distributed but with potentially different variances. All other pairs of conditions are significantly different using either test.

3.4 Results and Discussion

Our experimental results indicate VoiceBlock can de-identify speech from arbitrary unseen users in real-time on a standard M1 CPU. As shown in Table 3.1, VoiceBlock lowers the top-1 speaker identification performance of a ResNetSe34V2 system from 99% to 3%, outperforming both a universal additive attack (9%) and high-magnitude white noise (16%) while remaining competitive with spectral gating (2%). Crucially, VoiceBlock de-identifies speech while maintaining significantly higher audio quality than competing methods, as measured by PESQ (3.74 vs. 1.35 for the next-highest method), STOI (0.92 vs. 0.78 for the next-highest method), ASR WER (0.02 vs. 0.16 for the next-lowest method), and ASR CER (0.01 vs. 0.08 for the next-lowest method, and on par with clean speech). These results are supported by our subjective evaluation, in which human listeners rated the VoiceBlock highest of all evaluated methods in audio quality.

VoiceBlock also achieves nontrivial de-identification results against a Y-Vector speaker recognition system it was not trained on, outperforming a far more perceptible universal attack. This is notable given that we take no explicit steps to improve the transferability of our method, which leverages a lightweight model trained on a small corpus of clean speech without augmentation. While pure signal-processing approaches such as spectral gating and white noise transfer between

systems more successfully, they severely degrade audio quality as noted above. For instance, spectral gating and white noise produce ASR word error rates $20\times$ and $50\times$ that of VoiceBlock, respectively. This limits their practicality in real-world voice communications, where intelligibility is crucial. By contrast, our method produces adversarial audio that is virtually indistinguishable from the clean source, as indicated by our subjective evaluation.

We find that VoiceBlock loses almost no effectiveness when adversarial utterances are processed with an unseen Demucs speech enhancement model; by comparison, the universal attack loses most of its effectiveness, as Demucs is able to separate the low-magnitude but noisy perturbation from clean speech. Both the white noise and spectral gating attacks retain their effectiveness, presumably because they degrade audio beyond Demucs’s capability to provide coherent reconstructions of the original speech. We hypothesize that VoiceBlock’s filtering-based processing produces perturbations closely aligned with the source signal – as filtering only amplifies or attenuates frequency content present in the source – making them more difficult for enhancement models such as Demucs to remove.

3.4.1 Limitations and Future Work

While the proposed VoiceBlock system improves substantially on existing algorithms for intervening adversarially against AI-assisted biometric surveillance of voice communications, it also exhibits a number of limitations.

The first major limitation is transferability: while VoiceBlock demonstrates nontrivial de-identification performance against the unseen Y-Vector speaker recognition system, its effectiveness drops significantly when compared to the ResNetSe24V2 system seen during training. As shown in Table 3.1, VoiceBlock reduces the top-1 recognition accuracy of ResNetSe24V2 from 99% to 3%, but the reduces top-1 recognition accuracy of Y-Vector from 96% only to 35%. Transferability is a major topic of research within the broader area of adversarial machine learning, and a number of techniques have been proposed for improving the transferability of standard attacks [244–247].

Future work may therefore explore the application of such methods to VoiceBlock. I later discuss this limitation as it applies to adversarial interventions more broadly in Chapter 5.

The second major limitation exhibited by VoiceBlock is its degraded de-identification performance under the enrollment of adversarial utterances. We find that while VoiceBlock is highly effective at de-identifying users against a profile constructed from clean (unperturbed) utterances, its de-identification performance suffers significantly when the operator of a speech recognition system is able to build a dedicated profile of VoiceBlock-processed utterances or incorporate such utterances into a pre-existing profile of the user’s clean speech. Table 3.3 illustrates this degradation in de-identification performance. This suggests that future work should focus on ensuring that a VoiceBlock user remains a "moving target" to surveilling speaker recognition systems, preventing the construction of accurate speaker profiles. One possible solution is to leverage targeted rather than untargeted attacks: by proactively "spoofing" specific (random) locations in the speaker embedding space, VoiceBlock may hamper the creation of a single matching enrolled profile. In theory, this requires only minor modifications to the adversarial training loss (Eq. 2.2) and assignment of training targets. Subsequent work has demonstrated the feasibility of this general approach [248].

Finally, VoiceBlock stands to benefit from improved auxiliary training objectives. In an ablation experiment, we replace the combined waveform and multi-resolution spectrogram loss of Defosséz et al. [186] with a waveform-only L_1 loss, the mel-frequency cepstral coefficient cosine-similarity loss used in my prior work on filtering-based adversarial attacks [181], and an ASR feature-matching loss that penalizes the divergence between activations within a Wav2Vec 2.0 speech recognition model given clean speech and speech processed by VoiceBlock. Results are reported in Table 3.4 alongside de-identification performance. Training with ASR feature-matching loss results in a model that outperforms the “main” VoiceBlock configuration in all objective metrics, suggesting that future work could explore further improvements to our training objectives.

Table 3.4. Results of training VoiceBlock with different auxiliary loss functions. We consider a simple waveform regression loss (" L_1 ") and the mel-cepstral cosine-similarity loss used by O'Reilly et al. ("MFCC-cosine") [181] as simple drop-in replacements for the combined waveform/spectrogram loss of Defosséz et al. [186] (" L_1 + MRS"). Additionally, we consider a feature-matching loss on the acoustic representations produced by a Wav2Vec 2.0 [241] automatic speech recognition model ("ASR"). This loss is computed by taking the L_1 distance between acoustic feature vectors produced by the Wav2Vec 2.0 encoder over clean and adversarial utterances, scaled by a factor of $1e^{-6}$.

Auxiliary Loss	Speech Quality		ResNetSe34V2	
	PESQ $\uparrow$	STOI $\uparrow$	T-1 $\downarrow$	T-10 $\downarrow$
L_1	3.49	0.89	0.02	0.07
MFCC-cosine	3.34	0.89	0.02	0.07
ASR	4.08	0.94	0.02	0.09
L_1 + MRS	3.74	0.92	0.03	0.10

3.5 Impact and Follow-On

My work on VoiceBlock shows that large-scale, indiscriminate, automated surveillance can be hindered by introducing inconspicuous perturbations to audio in real-time. Because VoiceBlock does not conceal speaker identity from human listeners, it still allows high-effort targeted surveillance (e.g. authorized human-attended wiretaps of criminal enterprises). In this way, the adversarial intervention embodied in VoiceBlock is an attempt to return the status quo of the 20th and early 21st centuries – in which the need for human listeners provided an important check on mass surveillance.

VoiceBlock's core design element – a neural network parameterizing a time-varying frequency-domain perturbation of an audio signal – also inspired the design of the audio watermarking method presented in Chapter 4. The use of frequency-domain perturbations to conceal information inconspicuously in audio signals is a recurring motif in this dissertation, and a good example of the ways in which domain-specific knowledge (e.g., audio filtering mechanisms) can be used to improve on more general techniques (e.g., adversarial waveform optimization inspired by attacks in the vision domain).

I have followed up VoiceBlock with additional work developing adversarial interventions. In

2024, I collaborated with researchers to explore methods for protecting music recordings from unauthorized AI mimicry, and presented a demo of our preliminary work Audio Data Defenses: Protecting Music and Speech Data from Targeted Attacks at ISMIR 2024³ [249]. This work examined the effectiveness of optimization-based adversarial attacks on audio used to condition an autoregressive music generation system, targeting an underlying neural audio codec (see Section 2.1.3) rather than an embedding model as in VoiceBlock. While this work demonstrated the feasibility of such attacks, it also identified significant limitations in the application of adversarial attacks for audio data protection and adversarial interventions more broadly (discussed further in Chapter 5).

My work on VoiceBlock and adversarial data protection exists within a broader context of adversarial interventions intended to interfere with maliciously or irresponsibly deployed AI systems in audio and other modalities. This paradigm presents numerous challenges – for instance, protections must maintain their effectiveness against previously unencountered AI systems, outside the lab, in unpredictable real-world scenarios. However, many developers of AI systems – particularly *generative AI* systems that produce digital media such as images, text, voice recordings, and music – are strongly motivated to deploy their technology safely. Given these factors, significant research effort has shifted from purely adversarial interventions to *collaborative interventions* that leverage related techniques to identify and mitigate potential AI harms. In the next chapter, I turn my focus from de-identification to identification, and discuss my research on tools to aid safe AI system deployment through the labeling of AI-generated audio.

³https://oreillyp.github.io/assets/manuscript/494_lbd.pdf

CHAPTER 4

COOPERATIVE INTERVENTION

In this chapter, I present MaskMark, a novel neural network-based watermarking method for the identification of AI-generated audio. This work was published at ICASSP 2024 as MaskMark: Robust Neural Watermarking for Real and Synthetic Speech [42], and is presented in edited form here with additional discussion and analysis. I use the pronoun “we” in this chapter to acknowledge the contributions of my co-authors Zeyu Jin, Jiaqi Su, and Bryan Pardo.

Whereas the VoiceBlock system presented in Chapter 3 is an adversarial intervention intended to sabotage a harmful AI system, MaskMark is a *cooperative intervention* intended to operate alongside a deployed AI system with the support of its developers to mitigate potential harms. More specifically, MaskMark *watermarks* the output of a speech synthesis system by concealing within it an inaudible signal that can be detected algorithmically – allowing for the automated identification of synthetic audio, and giving digital media platforms and policymakers a tool to aid in the tracing and moderation of harmful AI-generated media. This work was performed in collaboration with Adobe Inc., a firm with a growing array of AI-powered audio products, offering insights into the challenges of production-grade audio watermarking under difficult real-world constraints on perceptual quality and performance. In the next section, I discuss the limitations of prior synthetic audio identification methods under these constraints and provide technical background. I then describe the design and evaluation of the proposed MaskMark system, and close the chapter by situating MaskMark within the growing body of research on cooperative audio watermarking interventions.

4.1 Preliminaries

In this section, I provide background information necessary to ground the research contributions embodied in the MaskMark system.

4.1.1 Speech Synthesis

Speech synthesis systems – encompassing capabilities such as text-to-speech (the mapping of graphemes to audio recordings), voice cloning (the imitation of a specific speaker’s identity in an audio recording, typically within the context of text-to-speech), and voice conversion (the mapping of recorded speech from one speaker’s voice to another) – have numerous beneficial applications, such as aiding individuals with speech difficulties and fostering creative expression [250, 251]. However, these systems can also be used for malicious purposes, including voice impersonation and disseminating false information. Recent speech synthesis systems allow users to “deepfake” the voices of public figures to make statements ranging from incongruous and comedic to offensive and hateful [12, 252]. Voice conversion has been used to replicate the voices of famous singers [14], igniting debates on copyright and intellectual property. Moreover, the prevalence of realistic synthetic media has been used in attempts to cast doubt on the authenticity of digital evidence in court [253]. Developers and providers of speech synthesis systems therefore need methods to identify speech produced by their models, while preserving the sonic quality of the generated speech, thereby ensuring that beneficial applications are not hindered.

Modern speech synthesis systems leverage a variety of neural network architectures to perform text-to-speech, voice cloning, and voice conversion. In this work, we evaluate the effectiveness of the proposed MaskMark watermark (Section 4.3) using the VITS [254] speech synthesis system. VITS combines a number of deep learning components discussed in Section 2.1.3; at its core, a variational autoencoder (VAE) framework is used to map text representations to an acoustically-aligned latent space, from which audio is decoded using a generative adversarial network (GAN).

Subjective listening evaluations show that VITS approaches the perceptual quality of human speech in certain settings, with some listeners rating outputs on par with non-synthetic recordings. As a cooperative intervention, MaskMark is intended to sit outside a deployed speech synthesis system like VITS, treating it as a black box; we therefore design our watermarking method in a manner agnostic to the architectural details of any specific speech synthesis system, and develop a temporally redundant frequency-domain information-hiding mechanism robust to variations in the content and timbre of audio to be watermarked (see Section 4.2).

4.1.2 Synthetic Audio Identification

Researchers have developed a number of methods for identifying synthetic audio and media more broadly. I next survey these methods – grouped into four main categories – and discuss their limitations in reliably identifying synthetic speech recordings under real-world conditions.

Metadata: the simplest approach to marking synthetic audio recordings is to attach labels as file metadata. Metadata allows for the storage of detailed provenance information, enabling digital media platforms to label and trace metadata-bearing AI-generated content [255–258]. However, metadata can be scrubbed intentionally via simple editing of the marked files or unintentionally via re-capture (e.g., screenshots of metadata-bearing images) [259–261], making it a brittle and partial solution at best. Orthogonally, this brittleness can be leveraged to instead mark genuine or human-generated content with cryptographically signed provenance information. This approach faces its own challenges, including the widespread adoption of a unified protocol across media capture devices and editing software [255, 262], and can be thought of as complementary to the synthetic media identification methods discussed in this chapter.

Classification: a straightforward approach to synthetic media identification is to train models to classify this media, removing producer-side requirements for marking (e.g., metadata). Researchers have developed numerous classification-based systems for synthetic audio identification, with the majority of published work focusing on synthetic speech due to its clear potential harms.

These systems typically use neural network models trained to perform binary classification on standard corpora of natural and synthetic speech [263–266]. The development of these classifiers presupposes that speech synthesis systems exhibit characteristic artifacts not present in natural speech, and that identification performance on a small set of existing speech synthesis systems will generalize to future, unseen systems. In practice, this has not been the case. The ASVSpooF challenge [263] has been organized to evaluate approaches for discriminating between authentic and synthetic speech utterances; updated evaluations from the most recent completed edition of the challenge show that detecting arbitrary synthetic speech remains far from solved, with the best-scoring submission achieving an equal error rate (EER) of 15% on unseen synthesis systems. Existing classifiers have been found to be vulnerable not only to optimization-based adversarial attacks [39], but to simple low-effort transformations of synthetic audio such as reverberation (which not only occurs naturally when audio is re-recorded in a physical environment, but is often applied to speech and music for artistic effect) [267]. Thus, existing classifiers can be circumvented easily and even unintentionally. Moreover, closing this robustness gap will not be sufficient for producing generalizable systems, as the distribution of synthetic speech is not stationary: speech synthesis systems continue to approach human-indistinguishable quality [57, 268] as architectures and training paradigms evolve to mitigate characteristic artifacts like aliasing and pitch inconsistency [269, 270]. It therefore appears likely that as speech synthesis technology improves, synthetic speech classification systems will face an increasingly difficult task. The same trends hold for the detection of synthetic music [271], indicating that *passive* classification methods may face hard limitations, and that researchers should explore *active* identification methods that intervene cooperatively to make synthetic audio more recognizable.

Retrieval: one such active identification method is retrieval or “fingerprinting,” in which an AI system provider maintains a centralized database of all outputs from a particular generative model; given any content of unknown provenance, the database can be searched for a closest match according to a similarity metric, and if the similarity exceeds a fixed threshold the content can

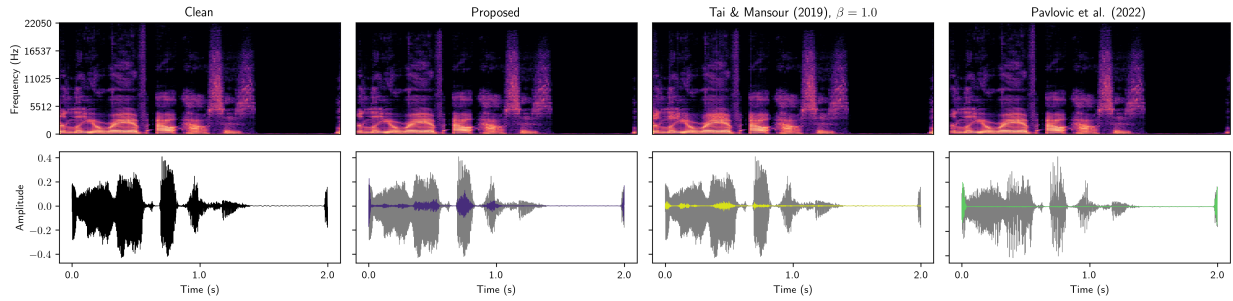


Figure 4.1. Top row: spectrograms of a two-second speech utterance watermarked with the proposed MaskMark and baseline methods. Bottom row: waveform difference between original and watermarked utterances for each method, highlighted against the original waveform.

be assumed synthetic. Similarly to classification, this does not require active modification of the generated content; unlike classification, this determines provenance via content matching rather than characteristic artifacts, potentially allowing for more precise and interpretable detection [272]. However, retrieval methods require storing large amounts of potentially overlapping user-generated media and thus raise questions of scalability and privacy. Perhaps as a result, retrieval has seen limited adoption in the audio modality [273].

Watermarking: a promising alternative approach for synthetic audio identification is watermarking, in which an embedding algorithm conceals a traceable signature (*key*) in audio signals and a corresponding detection algorithm determines whether the key is present in an audio signal. This allows the watermarked output of a speech synthesis system to be distinguished from unwatermarked natural speech even in the absence of other identifying artifacts [274]. Unlike retrieval, watermarking methods do not require the storage of generated media because the information required for identification is embedded in the media itself.

To enable practical synthetic speech detection, a watermark should satisfy two criteria. First, to avoid compromising the quality of synthetic speech, the watermark should be *perceptually transparent* – imperceptible to human listeners. Second, the watermark should be *robust* – capable of operating with strong detection accuracy even when the watermarked audio is significantly altered. Such alterations may include passive transformations of the watermarked audio from transmission

over digital and physical channels (e.g. equalization, reverberation, codec compression); active transformations that occur through benign user interactions (e.g. processing synthetic speech clips to remove pauses); and transformations specifically intended to thwart watermark detection (e.g. pitch/time-scale modification).

Speech watermarks published prior to MaskMark struggle to provide robust discrimination between watermarked and unwatermarked speech while maintaining the requisite perceptual transparency for widespread use. Tai & Mansour propose an autocorrelation-based spread-spectrum watermark for suppressing wake-word detection on Amazon Alexa [275]. While their proposed “Eigen” watermark achieves strong performance in discriminating between short watermarked and unwatermarked speech segments in realistic acoustic environments, the watermark is vulnerable to certain transformations such as pitch- and time-scale modification. Others have sought to leverage the strengths of deep neural networks for watermarking audio in a data-driven manner. Liu et al. propose training paired embedder and detector networks to perform music watermarking in the wavelet domain [276]. However, their method requires long audio segments for embedding and produces clearly audible artifacts when applied to speech. Pavlović et al. also propose training paired networks to watermark short speech segments [32]. While their “DNN-A” system demonstrates strong perceptual transparency, it is not robust to common editing distortions. Moreover, in Section 4.3.2 we show that the low-capacity training scheme employed by Pavlović et al. results in the detector “overfitting” to a small set of key vectors, producing unacceptably high false-positive rates when discriminating between watermarked and unwatermarked speech.

4.2 Method

Motivated by the limitations of prior audio watermarks, we propose MaskMark, a robust neural network-based watermarking method for speech audio. MaskMark leverages a novel network architecture to embed watermarks in short (≤ 2 s) speech segments via a multiplicative spectrogram mask, and builds on the training method of Pavlović et al. [32] to significantly improve robustness

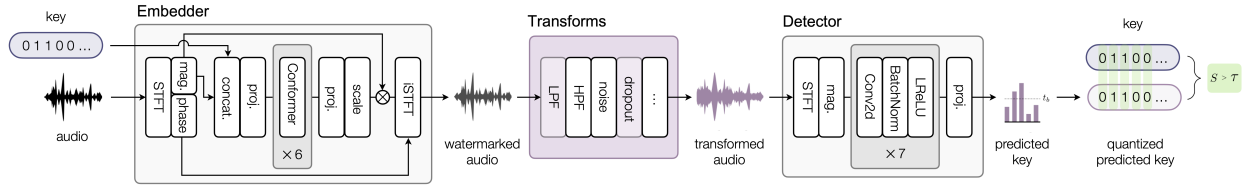


Figure 4.2. The proposed MaskMark architecture. Given a binary key vector and carrier audio to be watermarked, the embedder network modifies the carrier audio to encode the key. During training, embedded audio is then passed through a variety of acoustic transformations to simulate realistic conditions. The detector network is trained to predict the key vector embedded in (transformed) watermarked audio; because these keys are randomly distributed at training time, the detector’s predictions over unwatermarked audio at inference time are likewise random. Detection is performed by computing the proportion of matching bits between the detector’s prediction and a known watermark key.

and discriminative performance. We show that compared to Eigen and DNN-A, the strongest prior signal-processing and neural network-based speech watermarks, MaskMark achieves superior robustness under a comprehensive set of audio transformations encompassing editing manipulations, channel effects, and attacks designed to remove watermarks. Additionally, objective and subjective evaluations show that MaskMark maintains perceptual transparency competitive with or better than Eigen, the only baseline method to demonstrate practical performance. These strengths allow MaskMark to intervene cooperatively alongside deployed speech synthesis systems to render their outputs identifiable in real-world settings.

The following sections describe the design and training of the MaskMark system in more detail.

4.2.1 Architecture

We propose a neural network-based method for watermarking speech audio. Building upon the approach of Pavlović et al. [32], we use a two-network system to embed and detect watermarks, where each watermark is represented as a binary key vector of length 512, drawn from a fixed pool of n keys. An *embedder* network conceals a watermark key within a given *carrier* audio recording, while a *detector* network recovers the watermark by predicting the key vector contained in the carrier. The networks are trained such that the vector predicted by the detector closely matches the

key embedded within a watermarked carrier but is essentially random for unwatermarked audio. Thus, given knowledge of the embedded key, the operator of the detector can distinguish between watermarked and unwatermarked audio by comparing any predicted vectors to the known key via a calibrated similarity threshold. The proposed architecture is illustrated in Figure 4.2 and discussed component-by-component below.

Embedder: Our novel embedder network takes as input a key vector and carrier signal to be watermarked. To allow for embedding in high sample-rate audio, the carrier is resampled to 48kHz and a magnitude spectrogram is computed with a frame length of 1024 samples, hop length of 512 samples, and Hann window. The key vector is repeated and concatenated with each frame, and the result projected to a hidden dimension of 576 channels. This embedded representation then passes through six Conformer [277] layers with feed-forward dimension 768, six attention heads, and kernel size 5. The output is projected to match the frequency dimension of the carrier spectrogram and bounded to $[1e^{-5}, \infty)$ to obtain a mask. After multiplying the mask and carrier magnitude spectrogram, we take the inverse Fourier transform using the original carrier phase to obtain the watermarked waveform. In total, the embedder network contains 14.3m parameters. Whereas DNN-A directly generates the complex spectrogram of the watermarked audio, we find that the use of a real-valued spectrogram mask to embed the watermark allows our approach to easily generalize across input sample rates and prevents the embedder from introducing noise in silent regions of the carrier, as the masking operation is multiplicative. Additionally, the redundant frame-level embedding and Conformer module allow our embedder to robustly encode a larger set of key vectors and thereby avoid the “overfitting” behavior described in Section 4.3.2.

Detector: We adopt the fully-convolutional detector of Pavlović et al. with minor modifications. We prepend an additional strided convolutional layer, resulting in a receptive field of approximately one second at 48kHz (versus the original two seconds at 16kHz), and operate the detector on the same magnitude spectrogram representation as the embedder rather than a complex spectrogram. The detector predicts a 512-dimensional vector of unnormalized probability scores v , where v_i

represents the probability that the i^{th} bit in the watermark key is 1. At inference time, we quantize detector outputs to $\{0, 1\}^{512}$ using a fixed threshold of $t_b = 0.5$. In total, our detector network contains 5.1m parameters.

Scoring: As in previous works, we assume the detector operator has access to a watermark key [32, 275, 278]. Given candidate audio, we compute the proportion of matching bits between the detector’s quantized output and the known key to obtain a score S that is then compared to a calibrated threshold τ . If $S \geq \tau$, we conclude the candidate audio contains the watermark. We find this outperforms other scoring methods such as cosine distance, mean squared error, and binary cross-entropy.

4.2.2 Training

Through MaskMark, we seek to develop a cooperative intervention that does not rely on the specifics of any one speech synthesis architecture, and which will generalize to future synthesis systems that continue to close the perceptual gap to natural human speech. We therefore train MaskMark exclusively on non-synthetic speech recordings and evaluate on a mixture of synthetic and non-synthetic recordings.

While Pavlović et al. use a private speech dataset, we train on the VCTK dataset [279], totalling 44 hours of audio across 108 speakers sampled at 48kHz. We train on 2-second random excerpts for 70,000 iterations with batch size 64 and mixed precision. We use a fixed set of $n=64$ random watermark keys rather than the $n=6$ used by DNN-A, as we find this prevents the detector from “overfitting” to the key set and encourages more uniform random predictions over unwatermarked audio (see Section 4.3.2).

At each training iteration, we normalize carrier audio to -24dBFS , apply a random phase shift, and pass the carrier to the embedder network alongside a randomly selected key vector to obtain watermarked audio. Watermarked audio is then passed through a differentiable channel simulation. While past works have focused on additive noise, pass filters, and reverberation [32, 276], we

also consider pitch- and time-scale modification, as such transformations often prove especially destructive to watermarks. Our simulation encompasses low- and high-pass filtering at 4000Hz and 500Hz, respectively; recorded environmental noise [280–282] at 20-35 dB SNR; sample dropout with probability 0.001; reverberation applied via recorded and simulated impulse responses [281, 283]; phase-vocoder pitch shift of up to ± 1 semitone; speed change of $\pm 5\%$; and spectral gating applied via a threshold relative to the highest-magnitude bin of each spectrogram frame [10]. We find that training on this relatively small set of transformations allows MaskMark to generalize to a much broader set of unseen conditions (see Section 4.3.2).

At each training step, we randomly select three of the aforementioned transformations and apply each with an independent probability of 25%. Transformed audio is then passed to the detector network, which predicts an unnormalized probability distribution over the key vector. We compute two losses: binary cross-entropy loss on the detector predictions and embedded watermark keys; and a multi-resolution spectrogram loss on the embedder output and input [284] to encourage perceptual transparency. Gradients of the cross-entropy loss are propagated to both the detector and embedder (through the differentiable channel simulation), while gradients of the spectrogram loss are propagated only to the embedder. We balance the respective norms of these gradients within the embedder in a ratio of $\frac{100}{1}$ using the method of Défossez et al. [63] before performing an update. To ensure stability, we use AutoClip [285] to perform adaptive gradient clipping in both the embedder and detector networks at the 10th percentile. We use the NAdam optimizer [286] with learning rate $2e-4$, $\beta_1=0.9$, $\beta_2=0.999$, and weight decay $1e-6$.

4.3 Experiments

We perform experiments (1) validating MaskMark’s ability to robustly discriminate between watermarked and unwatermarked audio and (2) measuring its perceptual transparency through a subjective evaluation and objective quality metrics. We evaluate on two datasets unseen in training. To represent current synthesis systems, we use the VITS text-to-speech model [254] to resynthesize

utterances from 10 unseen VCTK speakers at 24kHz; and to represent high-quality future synthesis systems, for which generated speech may be “indistinguishable” from natural speech, we use the natural speech DAPS “clean” subset [287], containing studio recordings of 20 speakers sampled at 44.1kHz.

4.3.1 Baselines and Comparison Methodology

We next describe the baseline watermarking methods against which we compare MaskMark’s performance.

Eigen: We implement the Eigen watermark of Tai & Mansour [275] using an embedding band of 3-4kHz, non-overlapping windows of length 100ms, segment length $N_s = 2$, and repetition $N_r = 50$ to obtain a watermark length of one second as in the original work. Because the watermark strength parameter β is not specified in the original work, we evaluate two configurations with $\beta = 1.0$ and $\beta = 0.5$ to balance perceptual transparency and robustness.

DNN-A: We implement the DNN-A watermark of Pavlović et al. [32] using the hyperparameters detailed in the original work, including the differentiable transform set, and train on 2-second excerpts of the VCTK dataset for 140,000 iterations to ensure convergence. Our implementation achieves comparable robustness (as measured by key recovery under the dropout, filtering, and noise transforms considered in the original work) and perceptual transparency (as measured by PESQ [237] and SNR) to the authors’ implementation.

Conflicting sample rates: MaskMark operates at 48kHz, Eigen at 44.1kHz, and DNN-A at 16kHz. For MaskMark and Eigen, we account for sample rate differences by resampling audio before and after embedding. For DNN-A, we split carrier audio into two bands at 8kHz before embedding. The low band is resampled to 16kHz, embedded, and resampled to the original dataset rate (44.1kHz for DAPS, 24kHz for VITS). We then balance the energy of the two bands to match their pre-embedding ratio and sum. We find this does not impact robustness, and allows for fair listening comparisons between watermarking methods at the original dataset rates.

Watermark length: Following Pavlović et al., we perform robustness evaluations at a watermarked segment length of 2 seconds. The proposed MaskMark is capable of embedding at arbitrary lengths due to redundant frame-wise processing, and we simply repeat the Eigen watermark. For MaskMark and Eigen, which render detection scores at 1-second intervals, we take the maximum over all scores produced for a segment. Finally, for our perceptual evaluations we embed repeatedly to cover entire utterances.

4.3.2 Robustness to Audio Transformations

We evaluate the robustness of the proposed and baseline watermarks in a synthetic speech detection scenario. For each of the aforementioned watermarks and datasets, we sample 2000 two-second excerpts and embed with random keys to generate pairs of watermarked and unwatermarked utterances. We then apply an audio transformation (see below) and compute detector scores for the transformed utterances using the known keys. For Eigen, we compute the self-correlation score as defined by Tai & Mansour [275]. For both MaskMark and DNN-A, we compute the bitwise similarity score described in Section 4.2.1 to allow for fair comparisons. For a watermark to be of practical use, the distributions of synthetic (watermarked) and natural (unwatermarked) scores should be sufficiently separated to allow for reliable detection with a low false-positive rate; following previous work in synthetic media detection [288], we compute the achievable true positive rate when the threshold τ is empirically calibrated to fix the false positive rate at 1.0% (TPR@1%FPR).

We evaluate the robustness of the proposed and baseline methods over a set of 26 realistic audio transformations, including 18 unseen in training. **Signal-processing:** we modify the training transformations listed in Section 4.2.2 to use unseen impulse responses, noise recordings, and speed change ratios within the interval $\pm 5\%$; we also lower noise SNR to 10dB. For evaluation we implement the following additional transformations: Mu-Law quantization with 256 channels; dynamic range compression [289]; constant phase shift in $\pm\pi$; clipping applied at the 95th percentile of observed amplitudes; OGG-Vorbis compression; equalization across six bands, each with ran-

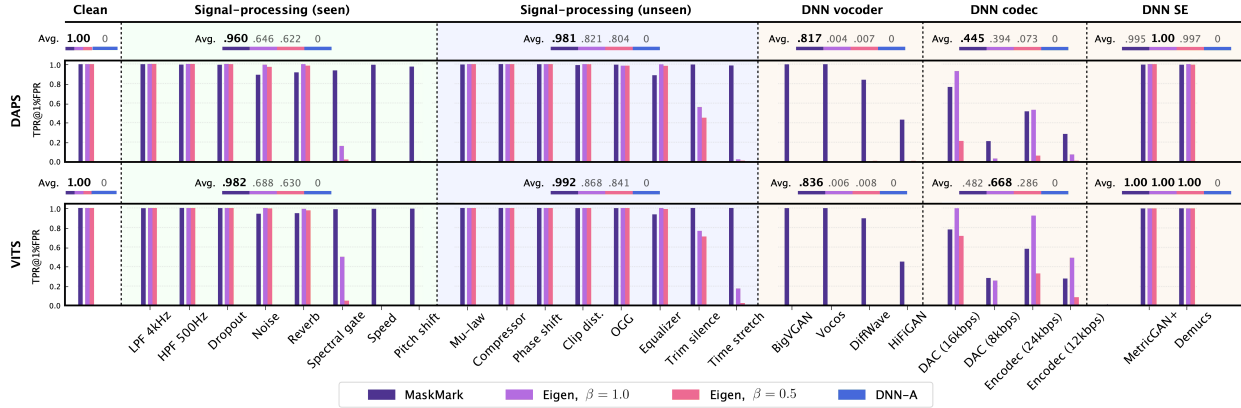


Figure 4.3. TPR@1%FPR for watermarks under selected transformations, indicating the robustness of each method when configured to operate with a fixed 1% false-positive rate. Scores in the top row are for the DAPS dataset, scores in the bottom row for our synthesized VITS dataset. The proposed MaskMark method is robust to more transformations than the baselines. DNN-A has a true positive rate of 0% for false-positive rates below 4%, and thus its bars are not visible.

domly sampled gain in ± 1 dB; silence removal; and phase-vocoder time-stretching in the interval $\pm 10\%$. **Neural vocoding:** we transform audio by resynthesizing from mel-spectrograms using the pre-trained neural vocoder models BigVGAN [269], Vocos [290], DiffWave [291], and HiFiGAN [292]. The mel-spectrogram representation discards a large amount of information from the audio, which is subsequently inferred by the vocoder. As a result, this transformation may substantially alter or destroy embedded watermarks while maintaining much of the audio’s perceptual quality. **Neural codec compression:** we transform watermarked audio by applying the state-of-the-art neural codec models Descript Audio Codec (DAC) [64] and Encodec [63]. Similar to vocoding, neural network-based compression reconstructs audio from a tight information bottleneck and may degrade the effectiveness of embedded watermarks. **Neural speech enhancement (DNN SE):** speech enhancement models remove or suppress non-speech sounds from a recording; thus, if an embedded watermark can be distinguished from the carrier speech (e.g. if it manifests audible artifacts), it may be partially or wholly removed. We use pretrained Demucs [284] and MetricGAN+ [293] models.

Experimental results are in Figure 4.3. We find that while DNN-A often recovers embedded

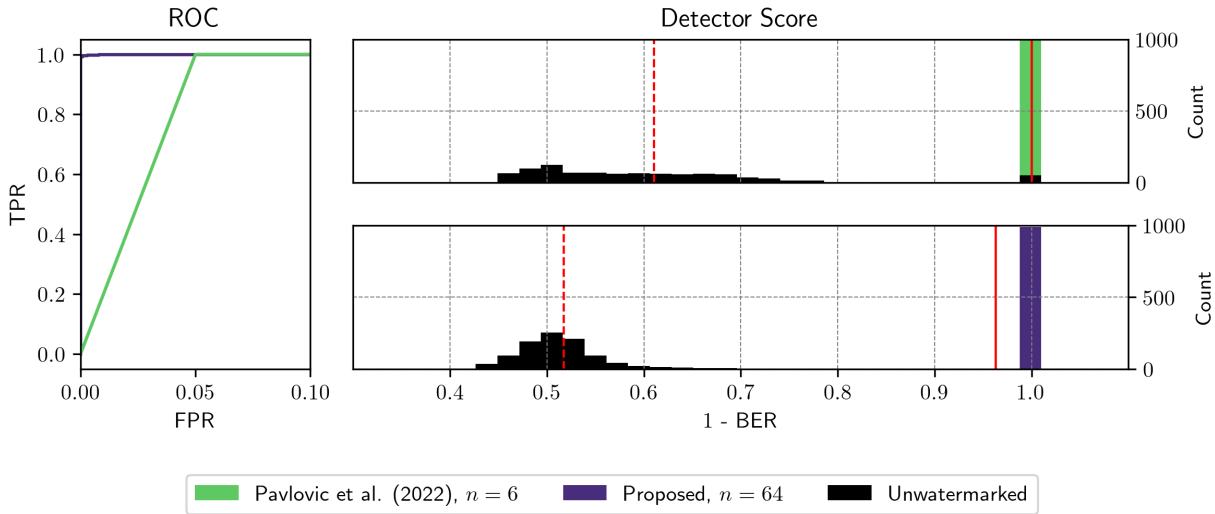


Figure 4.4. Receiver Operating Characteristic (ROC) curves and detector score histograms for the proposed MaskMark method and the DNN-A method of Pavlović et al. over the DAPS-clean subset. The dotted red line marks the average score of unwatermarked segments, while the solid red line marks the maximum observed score of unwatermarked segments. The method of Pavlović et al. produces markedly higher detection scores for unwatermarked audio, making operation at low false-positive rates impossible.

key vectors with high accuracy, the method can not distinguish between watermarked and unwatermarked audio at sufficiently low false-positive rates to be of practical use. We hypothesize training on $n=6$ keys causes the detector network to “overfit”; as a result, on average 4% of *unwatermarked* utterances are mapped exactly to key vectors (and thus perfect detector scores) even in the absence of any transformations – an impractical lower bound on the operating false positive rate. This behavior is further illustrated in Figure 4.4.

By contrast, MaskMark achieves competitive or better robustness under every evaluated transformation when compared to Eigen. MaskMark performs significantly better in the average and worst case over each set of transformations with the exception of neural speech enhancement (in which both Eigen and MaskMark show near-perfect performance) and neural codec compression, for which Eigen performs better on the VITS dataset in the average case. MaskMark demonstrates strong robustness to transformations for which other methods fail completely at low false positive

rates (time-stretching, neural vocoding), and generalizes far outside its relatively small training dataset and limited training distribution of signal-processing transformations.

4.3.3 Perceptual Transparency

When deployed as a cooperative intervention alongside a speech synthesis system, a watermark must remain perceptually transparent in order to avoid degrading the quality of generated speech (and thus the user experience). We assess the perceptual transparency of MaskMark through standard speech quality metrics and a human listener study. Following prior work [32], we compute SNR and PESQ [237] scores over 500 pairs of watermarked and unwatermarked utterances on each dataset. We sample 5-second excerpts from the DAPS dataset and full recordings from the VITS dataset to ensure sufficiently long utterances. Additionally, we conduct a human listening study on Prolific¹ in which 119 participants are asked to distinguish between pairs of watermarked and unwatermarked utterances. In this format, a 50% identification rate corresponds to a perfectly transparent watermark, while higher identification rates correspond to diminished transparency. We select 50 utterance pairs from each dataset, yielding a total of $n=800$ trials per method. Each utterance is assigned to 16 different participants from the United States. The participants are required to wear headphones and pass a listening test to ensure they can distinguish subtle differences in audio samples. Each assignment covers 4 utterances processed by all testing methods and 4 secret validation tests to ensure participants are paying attention. Participants can complete as many unique assignments as they please, and are paid at \$15 per hour. Table 4.1 shows the results of our evaluations; listening examples are available on a supplementary webpage².

While MaskMark produces the lowest average SNR, we note that SNR is designed to measure the magnitude of additive perturbations of the original signal, while our method produces highly correlated filtering-like perturbations through the use of spectral masking. This lets MaskMark

¹<https://www.prolific.co/>

²<https://interactiveaudiolab.github.io/project/maskmark.html>

Speech dataset	SNR $\uparrow$		PESQ $\uparrow$		Perceptual ID $\downarrow$	
	DAPS	VITS	DAPS	VITS	DAPS	VITS
DNN-A [32]	30.06	38.43	4.39	4.38	51.4%	51.2%
Eigen [275], $\beta = 1.0$	23.40	21.10	4.24	4.12	78.2%	76.3%
Eigen [275], $\beta = 0.5$	28.82	26.43	4.44	4.40	61.4%	60.5%
MaskMark	19.55	22.01	4.56	4.59	63.9%	60.8%

Table 4.1. Objective speech quality metrics and human identification rates for the proposed MaskMark and baseline watermarking methods. DNN-A values are shaded to indicate it fails to detect watermarked speech at low false-positive rates ($< 4\%$).

introduce large-magnitude perturbations while also maintaining the highest PESQ scores of all evaluated methods.

In the human evaluation, we find that all methods, save DNN-A, are distinguishable from clean audio at $p < 0.001$. While DNN-A obtains the best perceptual transparency, this comes at the expense of detection performance too poor for practical use: it fails to detect true watermarks when the false positive rate is $< 4\%$. We note that, even when given a clean reference and explicitly instructed to listen for the watermark, listeners are only able to correctly distinguish between MaskMark and clean audio roughly 60% of the time, where chance performance is 50%. Via the Chi-Squared independence test, Eigen ($\beta=1.0$) is significantly more perceptible than the proposed method at $p \ll 0.001$, while Eigen ($\beta=0.5$) does not differ significantly ($p = 0.3057$ for DAPS and $p = 0.918$ for VITS). Overall, MaskMark is not significantly more perceptible than the “weak” Eigen configuration ($\beta=0.5$) while achieving superior robustness to the “strong” Eigen configuration ($\beta=1.0$).

4.4 Results and Discussion

Our experimental results show that MaskMark can discriminate between watermarked and un-watermarked speech more robustly than prior signal-processing and neural network-based speech watermark baselines, while matching or exceeding the perceptual transparency of the only baseline

to achieve nontrivial robustness. As with VoiceBlock, we find a frequency-domain perturbation strategy advantageous, as it allows our embedder to introduce large-magnitude but perceptually inconspicuous modifications to the carrier audio to hide the watermark. Crucially, we find that robust watermarks for synthetic speech can be trained end-to-end on a small dataset of non-synthetic speech and challenging audio transformations, offering a path towards cooperative interventions around arbitrary speech synthesis systems.

4.4.1 Limitations and Future Work

While the proposed MaskMark improves significantly on prior audio watermarking methods in the task of synthetic speech detection, we identify a number of areas for improvement in future work.

The results of our robustness evaluation (Section 4.3.2) show that neural network-based transformations such as vocoder resynthesis and neural codec compression can prove especially challenging for watermarks. While MaskMark achieves near-perfect performance against neural network denoisers and fares much better than baselines against neural network vocoder reconstruction, this advantage diminishes under processing with neural network codecs. Notably, no single watermark achieves higher than 30% TPR@1%FPR against Descript Audio Codec [64], making it the most disruptive transformation evaluated. This observation provided the inspiration for the study of watermark robustness presented in Chapter 5, which evaluates MaskMark and subsequent watermarking methods across a broader array of neural network-based audio transformations. Based on the results presented in Section 4.3.2), future work may incorporate such transformations directly into the training of neural network-based watermarks to improve robustness, and explore the effectiveness of stronger optimization-based adversarial attacks on watermark detectors.

MaskMark is also limited in its information-carrying capacity, as it is trained to embed only 64 key vectors rather than the full set of 2^{512} possible binary vectors. This limitation is in part a result of our choice to build on the prior method of Pavlović et al. – which encourages random detector prediction over unwatermarked audio through the use of a small subset of possible high-

dimensional key vectors rather than the full extent of a lower-dimensional set – as subsequent neural network-based watermarking works suggest that expanding this information capacity may be straightforward [294, 295].

4.5 Impact and Follow-On

My work on MaskMark represents a strong initial step towards robust, practical neural network-based watermarks that can intervene cooperatively around deployed speech synthesis systems to identify AI-generated audio. MaskMark’s development reflected a growing awareness of the potential harms of AI-generated media and the corresponding value of reliable identification methods. During the late stages of the project, initial works for watermarking the outputs of text generation [33] and image generation [288] models captured widespread attention and sparked additional interest in watermarking as a research area relevant to deep learning. This interest spread to audio as well: while there existed very few prior neural network-based audio watermarking methods when MaskMark was submitted for publication – none practical, for reasons mentioned in Section 4.1.2 – the following year saw the development and publication of dozens more, the vast majority of which leveraged the same fundamental end-to-end embedder-detector approach [294, 296–300]. Some of these subsequent works have explicitly built on MaskMark’s frequency-domain multiplicative masking mechanism [301] and exploration of neural network-based codec transformations [302].

Since presenting MaskMark as an oral at ICASSP 2024, I have continued researching the design of cooperative watermarking interventions for AI audio generation systems. In September 2025, I presented a tutorial on speech watermarking at the InterSpeech conference in Rotterdam³, covering the core design of MaskMark and other popular algorithms. I also conducted a follow-up research collaboration with Adobe beginning in spring 2024. While initially intended to develop additional watermarking techniques, this collaboration also uncovered key limitations of existing audio watermarking paradigms. In the following chapter, I present a study produced by this

³https://github.com/oreillyp/wm_tutorial/

collaboration and discuss its implication for algorithmic interventions around audio AI systems more broadly.

CHAPTER 5

LIMITS OF ALGORITHMIC INTERVENTION

In this chapter, I present a study identifying key limitations of existing algorithmic interventions such as watermarking that encode hidden information in audio signals. This work was published at the ICLR 2025 workshop on GenAI Watermarking (WMARK) as Deep Audio Watermarks are Shallow: Limitations of Post-Hoc Watermarking Techniques for Speech, and is presented in edited form here with additional discussion and analysis. I use the pronoun “we” in this chapter to acknowledge the contributions of my co-authors Zeyu Jin, Jiaqi Su, and Bryan Pardo.

The adversarial and cooperative interventions presented in the previous chapters aim to steer the outputs of neural network systems through the application of imperceptible perturbations to audio signals consumed by these systems. For adversarial interventions, these perturbations are designed to induce incorrect predictions (such as the speaker identities output by a biometric surveillance system); for cooperative watermarking interventions, these perturbations are designed to induce correct predictions of encoded information (such as the watermark keys recovered by a detector neural network).

Such information-hiding methods have a number of advantages. As demonstrated in both VoiceBlock and MaskMark, concealing information through small perturbations of pre-existing audio rather than from-scratch synthesis allows for the use of lightweight network architectures. Perturbation imperceptibility can be straightforwardly enforced through the optimization of signal-level metrics (e.g., spectral distances) between perturbed and unperturbed audio, preventing perturbations from interfering with genuine uses of surveilled communication channels or marked generative models. Perhaps most importantly, because these methods intervene *post-hoc* around existing neural network-based systems (e.g., speaker identification, speech synthesis) without requiring direct

access to their internals, perturbation-based information hiding methods can be deployed widely without bespoke modifications.

However, these strengths come at a cost. Recall that in Section 4.4.1 of the previous chapter, we found that the effectiveness of imperceptible post-hoc audio watermarks – including MaskMark – could be reduced through neural network-based transformations of the watermarked audio without noticeably degrading quality. This chapter expands these initial findings through a systematic study of a larger set of watermarks and audio transformations. Focusing on speech audio, we (1) unify and extend existing evaluations of the effect of audio transformations on watermark detectability, (2) demonstrate that state-of-the-art post-hoc audio watermarks can be removed with no knowledge of the watermarking scheme and minimal degradation in audio quality, and (3) discuss the implications of this demonstrated vulnerability for watermarks and general algorithmic interventions that rely on imperceptible information hiding. The results of our study motivate the development of more robust design-level interventions such as *generative watermarks* that operate proactively within AI systems, and which serve as the focus of Chapter 6.

5.1 Preliminaries

In this section, we discuss the limitations of prior evaluations of audio watermarking robustness and provide background information necessary to ground the contributions of our study.

5.1.1 (Post-Hoc) Digital Audio Watermarking

We next provide a condensed review of relevant watermarking methods in order to highlight their shared post-hoc methodology. For a detailed introduction to digital audio watermarking, we refer the reader to Chapter 4.

Digital audio watermarking methods typically consist of two algorithms – an *embedding* algorithm that hides a signature in audio, and a *detection* algorithm that determines whether given audio contains a signature. Commonly, the embedding algorithm encodes a message within the

embedded signature (e.g. a binary string of fixed length), and the detection algorithm attempts to recover this message; a decision as to whether audio contains a signature can then be rendered based on the similarity of the recovered message to a known watermark message [42, 296, 297], under the assumption that messages “recovered” from un-watermarked audio will be uniformly distributed over the set of all possible messages of the chosen length. Alternatively, a detection algorithm may directly produce a score indicating confidence that given audio bears a signature [303], or apply both approaches simultaneously [294, 295].

Traditional audio watermarking methods utilized interpretable signal-processing techniques for embedding and detecting signatures, e.g. via pseudorandom modulation of the spectrogram and correlation measures [275, 278, 304]. More recently, researchers have proposed methods to learn embedding and detection algorithms directly from data using deep neural networks [32, 42, 276, 294, 296, 297, 300]. In this formulation, an *embedder network* processes a given audio recording to embed a signature, while a *detector network* predicts the presence of a signature in a given audio recording and/or the message contained within the signature. Both networks are typically trained together to cooperatively embed and detect signatures. In general, deep neural network-based watermarking methods outperform signal-processing methods in their ability to embed human-imperceptible signatures that remain detectable when watermarked audio is transformed, in part due the incorporation of simulated transformations into the training process [42, 276, 296].

Crucially, existing signal-processing audio watermarks and almost all deep neural network-based audio watermarks operate in a *post-hoc* manner by embedding signatures via low-magnitude perturbations of existing audio at waveform or spectrogram representations [32, 42, 276, 294–298, 300]. While attempts have been made to integrate watermarks more deeply into the process of speech audio generation, the resulting methods can still often be considered post-hoc. For instance, a number of works incorporate watermarks into neural network vocoders that map low-dimensional audio representations predicted by a generative model to high-fidelity waveforms [299, 302, 303, 305]; however, because the vocoders are trained with signal-level objectives to penalize the

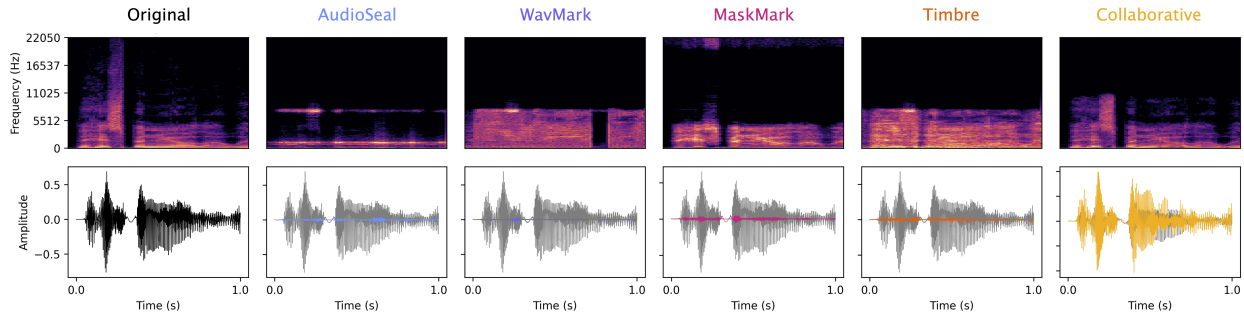


Figure 5.1. Watermark signatures. Given a short audio excerpt (“Original”), we apply five watermarking methods described in Section 5.3.1 (“AudioSeal,” “WavMark,” “MaskMark,” “Timbre,” and “Collaborative”) and visualize the difference between original and watermarked audio at the spectrogram (top row) and waveform (bottom row). For each watermarking method, the original waveform is shown in grey and the difference in color.

divergence between watermarked audio and the original audio from which vocoding representations are extracted, the resulting signatures still manifest as low-magnitude perturbations or are otherwise highly correlated with the source audio – they can not alter high-level speech attributes such as pitch, pronunciation, or timing ¹. Other works embed existing post-hoc watermarks in either the training data of audio generative models or in the latent decoders used to map generated instances to the data space [306–310], and thus inherit the detection limitations of these post-hoc watermarks. Finally, Ji et al. [311] use a masked language model operating on spectrogram tokens to replace selected audio frames with tokens drawn from a watermarked subset. While this approach hypothetically allows for the embedding of signatures within high-magnitude but realistic perturbations of an audio signal, in practice the signature is concealed in only a small subset of audio frames to avoid degrading audio quality (and is thus low-magnitude). In Figure 5.1, we visualize the signatures embedded in audio by selected neural network-based watermarks; we describe these watermarks in more depth in Section 5.3.1.

¹While we find the method of Juvela & Wang [302] results in larger perturbations as measured by waveform difference (see Figure 5.1), the watermark signature is highly correlated with the source, especially at the spectrogram.

5.1.2 Evaluations of Audio Watermark Robustness

While the aforementioned watermarking works conduct evaluations of detection performance under various audio transformations, the choice of transformations and evaluation metrics differs significantly from work to work. For example, some works consider only signal-processing transformations such as additive noise and gain scaling [32, 296], while others consider neural network-based transformations [42, 295, 297]. Additionally, some works measure detection performance in terms of message recovery accuracy [32, 296], and others in terms of the accuracy of discrimination between watermarked and un-watermarked audio [42, 302, 303]. Benchmarks such as AudioMarkBench [312] and OmniSealBench [313] partially bridge these differences but still focus mainly on signal-processing transformations and optimization-based attacks. Specifically, AudioMarkBench evaluates 15 transformations: 8 signal-processing transformations, 2 signal-processing codecs, 3 adversarial optimization attacks, and 2 first-generation neural codecs; OmniSealBench adds to these 6 signal-processing transformations. Our work differs from these previous works in the following key aspects: **(1)** we focus on watermark detection rather than message recovery; **(2)** we consider a much larger number of neural network-based transformations (14 vs. 2), including vocoder and denoiser transformation categories not included in AudioMarkBench or OmniSealBench; **(3)** we extend the codec bitrate evaluation from AudioMarkBench to more recent low-bitrate neural codecs developed specifically for speech, which better preserve audio quality; **(4)** we do not include optimization attacks, as we find that adversaries can more easily remove watermarks while retaining audio quality through the aforementioned single-pass neural network-based transformations; **(5)** we set detection score thresholds (Section 5.3.4) individually per transformation rather than on clean audio, preventing saturated error rates and allowing for clearer distinctions between transformation strengths [42, 288]. Through these choices, we hope to focus attention away from high-effort optimization-based attacks, and towards neural network-based transformations that can passively remove existing post-hoc watermarks while maintaining excellent audio quality.

5.2 Audio Transformations

This work evaluates the efficacy of a variety of audio transformations in removing watermarks. These transformations vary in the degree to which they modify a given audio recording, the manner in which this modification is performed, and the amount of effort and computational expense involved in their application. We try to focus our evaluations on realistic transformations that may be passively applied to watermarked audio as it is transmitted and stored, as well as to attacks that can be undertaken by adversaries with reasonable motivation and resource constraints. We now describe and motivate the transformations considered in this work.

5.2.1 Signal-Processing Transformations

The majority of audio watermarking works consider simple signal-processing transformations such as playback speed change, filtering, waveform dropout, additive noise, reverberation, pitch and timescale modification, phase shift, and nonlinear distortion [42, 294, 296, 297]. These transformations are easy to implement and apply efficiently while preserving the quality and intelligibility of transformed audio – for instance, small phase shifts and changes in playback speed are often imperceptible. Signal-processing transformations may also occur naturally, e.g. re-recording of watermarked audio will likely introduce reverberation, background noise, and filtering. These factors mean that in practice, signal-processing transformations may be encountered as cascades of many individual transformations, potentially increasing the difficulty of watermark detection.

5.2.2 Audio Codec Compression

Audio data is often compressed with codecs for efficient transmission and storage. To maximize perceived audio quality at high compression ratios, codecs are designed to discard perceptually irrelevant information within an audio signal while preserving perceptually relevant information.

This process may remove post-hoc watermarks that conceal low-magnitude signatures in perceptually irrelevant regions of audio, and as a result recent works have explored the detection performance of audio watermarks under both traditional signal-processing codecs [314, 315] and neural network-based codecs [63, 64]. In general, existing watermarks appear to fare better against signal-processing codecs than neural codecs [42, 294, 297, 303], although direct comparisons of state-of-the-art watermarks on multiple neural codecs are scarce in the literature. In addition to the general-purpose neural codecs studied in previous works, we consider for the first time recent *low-bitrate* neural codecs developed for high-ratio compression of speech.

5.2.3 Neural Vocoders

Similar to codecs, neural network-based vocoders reconstruct speech audio from compressed representations such as mel-spectrograms [292]. While vocoders are an important component of many speech generative models [57, 316], it is unlikely that watermarked generated audio will be transformed via additional vocoding unless an adversary specifically uses a vocoder to attempt to remove watermarks. Compared to codecs, a smaller number of works have explored watermark robustness to vocoders [42, 297].

5.2.4 Neural Denoisers

Neural network “denoisers” are often employed to restore noisy or otherwise corrupted audio [58, 284, 293, 317]. An attacker might seek to remove audio watermarks by adding noise to watermarked audio and then processing with a denoiser, in hopes the denoiser will remove the low-magnitude watermark signature along with the noise while preserving the perceptual quality of the original recording. While there exist a number of conventional signal-processing algorithms for removing noise, recent neural network-based denoisers show improved tolerance for high noise levels and are capable of removing complex nonstationary perturbations from speech signals [58, 284] – both potentially useful properties for watermark removal. A small number of audio watermarking works

have explored robustness to neural denoisers [42, 318].

5.2.5 Watermark-Aware Transformations

While recent works have explored the vulnerability of neural network-based audio watermarks to overwriting attacks [297] that flood watermarked audio with multiple signatures and optimization-based attacks [294, 295] that craft adversarial examples to fool detectors, both approaches assume some degree of knowledge of the attacked watermarking algorithms. We find that this knowledge is not necessary for removing watermarks in practice, and instead focus on the watermark-agnostic transformations listed in the previous subsections.

5.3 Experiments

We collect examples of the transformation types listed in Subsections 5.2.1 - 5.2.4 from the audio watermarking literature and conduct experiments measuring their efficacy in removing audio watermarks while preserving the overall quality of watermarked audio. In addition to these existing transformations, we also evaluate more recent and as-yet unstudied codecs, vocoders, and denoisers. Finally, we demonstrate how the audio degradation incurred through transformations can be reduced through band-splitting to target the frequency regions where existing watermarks operate.

5.3.1 Watermarking Methods

We evaluate five state-of-the-art neural network-based watermarking methods for audio, described here.

MaskMark, as described in Chapter 4, embeds watermarks via a multiplicative mask at the magnitude spectrogram and performs detection using a small spectrogram convolutional network. At the time of publication, MaskMark outperformed leading signal-processing and deep neural network-based watermarking methods in robust detection performance while maintaining strong perceptual transparency.

Timbre-Watermark [297], published concurrently with MaskMark, embeds watermarks via a convolutional neural network operating on the magnitude spectrogram rather than a multiplicative mask and aggregates predictions from a convolutional detector network via frame-level averaging rather than strided downsampling. Similarly to MaskMark, detection operates by measuring the bit accuracy of a predicted message against a known watermark message.

WavMark [296], developed concurrently with MaskMark, embeds watermarks via a residual waveform perturbation predicted using an invertible convolutional neural network whose weights are shared with a detector network. Detection is performed by measuring the bit accuracy of a recovered message against a known watermark message. Robustness to temporal distortions (e.g. time shift, speed change) can be improved via a localization scheme using additional encoded message bits.

AudioSeal [294], published subsequent to MaskMark, embeds watermarks via a residual (i.e. additive) waveform perturbation predicted using an Encodec-like [63] neural network, and is trained for simultaneous detection and message recovery with sample-level temporal resolution. Detection is performed by averaging sample-level detector scores over a segment of audio. For Timbre-Watermark, WavMark, and AudioSeal, we use the implementations provided in OmniSealBench [313].

Finally, **Collaborative Watermark** [303], published subsequent to MaskMark, embeds watermarks by synthesizing audio waveforms from a mel-spectrogram representation using a watermarked HiFiGan [292] vocoder. We use a variant [302] trained to be robust to processing with Descript Audio Codec [64].

5.3.2 Dataset

We embed watermarks in 200 five-second voiced segments from the clean subset of the DAPS dataset [287], consisting of clean speech from 10 male and 10 female speakers. The DAPS dataset provides a reasonable approximation of high-quality synthetic speech for our purposes, as all

Transformation		Quality Preservation			Watermark Detection (TPR@1%FPR) ↓				
		ASR-CER ↓	SIM ↑	SQUIM-MOS ↑	Timbre	WayMark	AudioSeal	MaskMark	Collaborative
Sig.-Proc.	Speed	0.03	0.93	4.81	0.27	0.00	0.15	1.00	1.00
	Reverb	0.01	0.93	3.73	1.00	0.98	0.96	0.93	0.83
	Pitch shift	0.01	0.86	4.67	0.17	0.00	0.04	1.00	0.71
	Low-pass	0.00	0.79	3.93	1.00	1.00	1.00	1.00	0.00
	High-pass	0.01	0.80	2.53	1.00	1.00	1.00	1.00	1.00
	Time stretch	0.03	0.90	4.67	1.00	0.44	0.35	1.00	0.74
	Equalizer	0.01	0.97	3.91	1.00	1.00	0.99	0.85	0.86
	Noise (0dB SNR)	0.08	0.71	2.66	0.00	0.00	0.85	0.10	0.00
	Noise (10dB SNR)	0.02	0.83	2.83	0.85	0.00	1.00	0.91	0.02
	Noise (20dB SNR)	0.01	0.90	4.58	1.00	0.04	1.00	1.00	0.06
Codec	MP3	0.00	0.87	4.54	1.00	1.00	1.00	1.00	0.64
	OPUS	0.00	0.97	4.80	1.00	1.00	1.00	1.00	0.79
	DAC	0.00	0.93	4.77	0.92	0.00	1.00	0.22	0.69
	ESC	0.00	0.96	4.81	0.99	0.00	1.00	0.25	0.69
	SpectralCodec	0.00	0.95	4.76	0.56	0.00	0.01	0.64	0.59
	Encodec	0.00	0.92	4.79	0.68	0.00	1.00	0.15	0.26
	Encodec-VoiceCraft	0.01	0.88	4.72	0.06	0.00	0.01	0.07	0.02
	SpeechTokenizer	0.01	0.85	4.59	0.03	0.00	0.00	0.04	0.37
	TiCodec	0.01	0.77	4.66	0.02	0.00	0.01	0.03	0.41
	FACodec	0.01	0.88	4.79	0.01	0.00	0.00	0.03	0.44
Vocoder	BigVGAN	0.00	0.99	4.78	1.00	0.00	0.29	1.00	0.69
	DiffWave	0.01	0.82	3.97	0.98	0.00	0.00	0.93	0.04
	HiFi-GAN	0.01	0.95	4.71	0.80	0.00	0.03	0.38	0.54
	BigVGAN2	0.00	0.99	4.82	1.00	0.00	0.47	1.00	0.81
	Vocos	0.00	0.98	4.77	1.00	0.00	0.14	1.00	0.48
Denoiser	Denoiser (0dB SNR)	0.10	0.63	4.60	0.00	0.00	0.02	0.01	0.00
	Denoiser (10dB SNR)	0.02	0.73	4.71	0.01	0.00	0.03	0.01	0.00
	Denoiser (20dB SNR)	0.01	0.81	4.75	0.04	0.00	0.04	0.03	0.01

Table 5.1. Watermark detection under transformations. Under *Quality Preservation*, we evaluate the audio degradation incurred by each transformation in terms of intelligibility (“ASR-CER”), speaker identity similarity (“SIM”), and overall quality (“SQUIM-MOS”). Cells are shaded darker where degradation is minimized. Arrows in column headers indicate the optimal direction to minimize audio degradation. Under *Watermark Detection*, we evaluate the effect of each transformation on the true-positive detection rate of watermarks, given a fixed 1% false-negative rate (“TPR@1%FPR”). Lower values indicate stronger watermark removal. Cells are shaded darker as the true positive detection rate decreases. The transformations that most effectively and inconspicuously remove watermarks are therefore shaded in every column.

utterances in the clean subset were recorded professionally at 44.1kHz with very little audible noise. DAPS is not present in the training data of any of the evaluated watermarks.

We choose a five-second segment length to allow for fair comparisons, and to ensure that transformation strength, not segment length, is the dominant factor in detection performance. While some works [294] consider larger evaluation sets, we are not interested in fine-grained distinctions between true positive rates but rather in quantifying the poor overall performance

of existing watermarking methods on speech recordings under transformations. In this case, comparing detection scores on 400 examples (200 un-watermarked, 200 watermarked) for each transformation and watermarking approach suffices.

The evaluated watermarks were not all designed to operate at the same sample rate. To account for differences in sample rate, we adopt the method used to evaluate MaskMark in the previous chapter:

1. Audio is split into two bands via low-pass and high-pass filters at the Nyquist frequency of the watermarking method’s native sample rate
2. The low band is resampled to the watermarking method’s native sample rate and the watermark applied
3. The watermarked low band is resampled to the original audio rate, normalized to account for gain change, and summed with the high band

In this manner, a watermark can be applied to audio beyond the watermark’s native sample rate. Detection can then be performed by resampling audio to the watermark’s native sample rate, discarding un-watermarked high-frequency content. We find that this method has essentially no effect on the detection performance of the evaluated watermarks, while allowing for embedding and detection in high-quality speech.

5.3.3 Transformations

We consider the following transformations from the categories described in Sections 5.2.1 - 5.2.4.

Signal-processing transformations: we apply Gaussian noise at 20dB, 10dB, and 0dB SNR; equalization across 6 bands with random gains sampled in $[-1, 1]$ dB; low-pass filtering at 4kHz; high-pass filtering at 500Hz; random pitch shift sampled in $[-1, 1]$ semitones; reverb using impulse responses sampled from the MIT dataset [283]; and playback speed change and phase-vocoding time-stretch at factors sampled in $[-0.95, 1.05]$. We find that a large number of signal-processing

transformations explored in previous works such as phase shift, waveform dropout, gain scaling, and sample-level quantization have virtually no effect on detection performance and therefore omit results to save space [42, 296, 312].

Audio codec compression: we apply the signal-processing codecs MP3 [314] and OPUS [315] at 24 kbps. We also apply the following neural network-based codecs: Encodec [63] at 24 kbps; ESC [319] at 9.0 kbps; Descript Audio Codec (DAC) [64] at 8 kbps; Spectral-Codec [104] at 6.9 kbps; FACodec [320] at 4.8 kbps; SpeechTokenizer [321] at 4 kbps; TiCodec [322] at 3 kbps; and the “VoiceCraft” Encodec variant of Peng et al. [323] at 2.2 kbps. Notably, ESC, Spectral-Codec, FACodec, SpeechTokenizer, TiCodec, and Encodec-VoiceCraft were developed specifically for speech audio. For codecs that operate below 44.1kHz (namely Encodec, ESC, FACodec, SpeechTokenizer, TiCodec, and Encodec-VoiceCraft), we use the band-splitting method described in Section 5.3.2.

Neural vocoders: we apply the mel-spectrogram vocoders BigVGAN [269], BigVGAN2 [324], DiffWave [291], HiFi-GAN [292], and Vocos [290]. With the exception of BigVGAN2, all evaluated vocoders are processed with the aforementioned band-splitting method to account for sample rates below 44.1kHz.

Neural denoisers: we build on the denoising attack of López-López et al. [318], which was shown to be effective against the neural network-based audio watermarks of San Roman et al. [294] and Chen et al. [296]. While López-López et al. utilize a discriminative DCCRN [317] denoiser model and a single noise level, we utilize a more recent generative denoiser model [58] and consider multiple noise levels of 20dB, 10dB, and 0dB SNR to examine the trade-off between attack strength and audio quality degradation.

5.3.4 Metrics

We are primarily interested in discrimination between real and synthetic audio. Although information recovery and attribution are desirable capabilities for watermarking methods, we find that

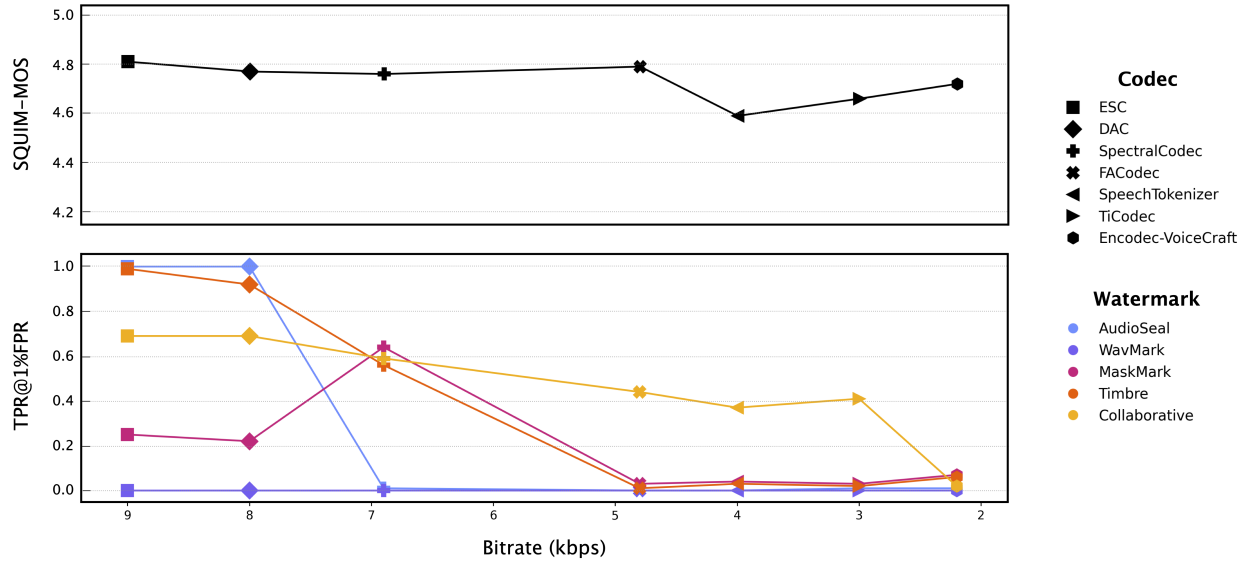


Figure 5.2. Robustness to low-bitrate codecs. We plot the achievable true-positive detection rate at a fixed 1% false-positive rate (“TPR@1%FPR”) for each watermarking method under transformation via low-bitrate neural codecs. Watermark detection declines at low bitrates while audio quality is generally preserved, as indicated by SQUIM-MOS scores.

the prerequisite task of simply detecting the presence of a watermark is already a high a hurdle to clear when watermarked audio is transformed. Therefore, we adopt the standard approach of measuring the achievable true-positive detection rate at a fixed false-positive rate based on detection scores obtained over a balanced dataset of watermarked and un-watermarked audio [42, 295]. While watermarking methods may be required to operate at false-positive rates below 0.1% to function practically in large-scale production environments [294, 325], we find that existing methods fail to detect watermarked audio under many transformations even when allowing much larger false-positive rates. We therefore use the true-positive rate at a fixed 1% false-positive rate (“TPR@1%FPR”).

Whereas previous works [318] quantify the audio degradation incurred by transformations using invasive metrics such as PESQ [237], these metrics do not correlate well with human perception for transformations that compromise precise alignment (e.g. speed change). Instead, we measure the audio degradation incurred by transformations along three interpretable axes: **speech intelligibility**,

as measured by the character error rate between transcripts predicted by a HuBERT-based [326] speech recognition system for watermarked audio before and after transformation (“ASR-CER”); **speaker identity preservation**, as measured by the embedding-space cosine similarity between embeddings produced by a speaker recognition system [327] for watermarked audio before and after transformation (“SIM”); and **audio quality similarity** as measured through mean opinion scores (MOS) predicted by the automated SQUIM [328] metric, with watermarked audio serving as a non-matching quality reference for transformed watermarked audio (“SQUIM-MOS”). Lower ASR-CER implies better intelligibility, higher SIM implies better speaker identity preservation, and higher SQUIM-MOS implies that the transformation introduces less-perceptible artifacts or changes to channel characteristics. ASR-CER has a lower bound of 0, SIM ranges from -1 to 1 , and SQUIM-MOS ranges from 1 to 5 .

The results of our evaluation are presented in Table 5.1. We plot a subset of these results for transformations that leverage low-bitrate neural audio codecs in Figure 5.2. In Figure 5.3, we illustrate the application of the denoiser attack to audio watermarked with the method of Liu et al. [297].

5.4 Results and Discussion

We first summarize key trends in our experimental results, then discuss the implications of these results for watermarks and algorithmic interventions that leverage post-hoc imperceptible information-hiding more generally. Listening examples for selected neural network-based transformations are provided on a supplementary webpage².

5.4.1 Watermark Removal

Our results show that post-hoc watermarks lack general robustness under transformations. No evaluated watermark demonstrates general robustness across all classes of transformation. While

²<https://deep-watermark.github.io/>

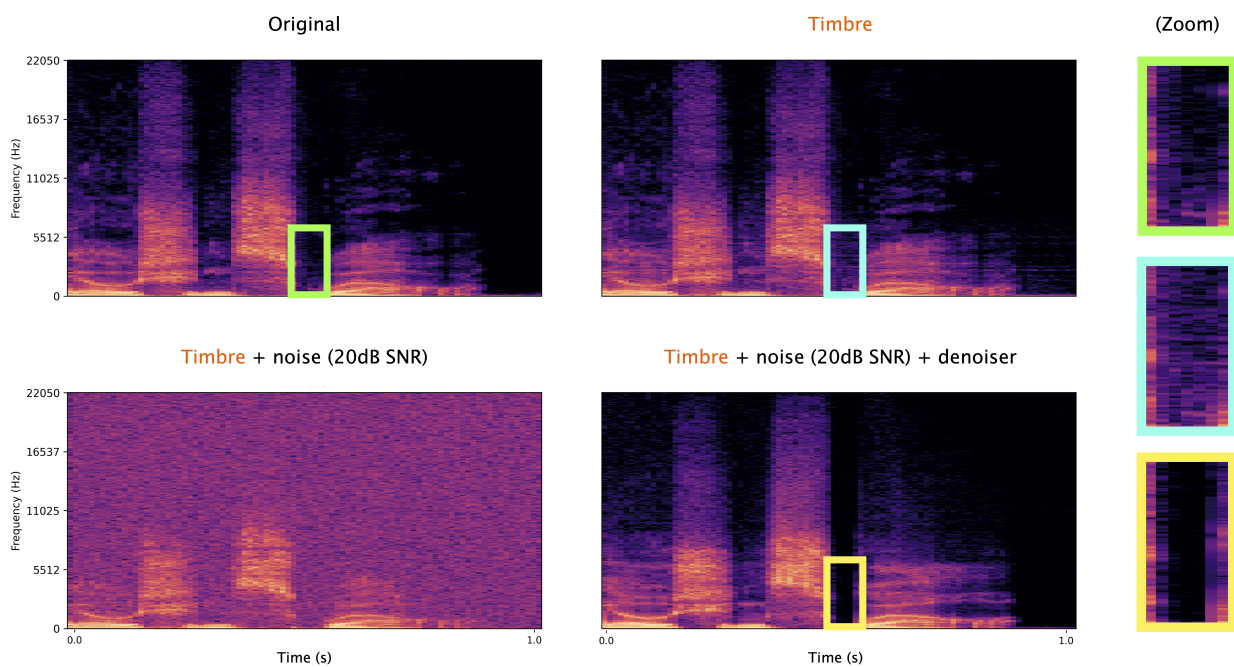


Figure 5.3. Denoiser attack. We illustrate the application of the denoiser attack to Timbre-Watermark with spectrogram plots. Artifacts introduced by the watermark (blue rectangle) can be seen more clearly when compared to the original spectrogram (green rectangle); these are removed by the application of noise followed by the denoiser (yellow rectangle).

some watermarks fare well against some classes (e.g. Timbre-Watermark and MaskMark against vocoders and signal-processing transformations), all watermarks have clear weaknesses that can be exploited by adversaries to thwart detection. Moreover, complementary weaknesses can be leveraged to construct combined attacks. For example, an adversary could apply speed change to foil Timbre-Watermark, WavMark, and AudioSeal, and then apply DAC to additionally foil MaskMark.

We find neural network-based transformations to be especially strong watermark removers. Neural network-based codecs, vocoders, and denoisers are capable of significantly reducing watermark detectability while preserving audio quality as quantified by our metrics. For example, our denoiser attack brings detection rates to near zero for all watermarks even when operating in its weakest configuration (20dB SNR) and allowing a generous 1% false-positive rate. Such transformations can be applied by leveraging off-the-shelf neural network models without any re-training or modification, making them a convenient option for even low-resource adversaries.

Of the neural network-based transformations evaluated, low-bitrate neural audio codecs pose a particular challenge for speech watermarking. Neural audio codecs increasingly leverage the structured nature of human speech to achieve large compression ratios and correspondingly low bitrates. To maintain high perceptual quality as bitrate decreases, these codecs must discard perceptually irrelevant details (e.g. fine-grained spectral structure) during the encoding process, and then “hallucinate” plausible reconstructions of these details during the decoding process. Existing post-hoc watermarks generally embed signatures through low-magnitude perturbations of perceptually irrelevant details, and are thus erased through an encoding-decoding pass. In Figure 5.2, we show that watermark detection performance under neural codec reconstruction generally decreases with the bitrate, even as codecs maintain high audio quality. This class of transformation is worth exploring further, as researchers continue to develop neural codecs capable of operating at even lower bitrates than those evaluated in this work, often with comparable or better audio quality [329, 330]. Moreover, if such codecs are integrated into digital media platforms due to their strong

performance, they may inadvertently remove post-hoc watermarks applied to track synthetic audio.

5.4.2 Implications for Post-Hoc Algorithmic Interventions

Our experimental results also hold implications for broader classes of algorithmic interventions that conceal imperceptible information in audio signals, such as the adversarial interventions discussed in Chapter 3. Similar to post-hoc watermarks, adversarial interventions like VoiceBlock operate by perturbing pre-existing audio signals imperceptibly to induce desired predictions from neural network systems. Such adversarial interventions face two main barriers to deployment in real-world contexts:

1. **Transferability:** many classes of adversarial interventions, such as those intended to protect user privacy, must contend with multiple unseen systems at the time of deployment and must remain effective against *future unseen systems*. For instance, adversarial perturbations intended to thwart facial recognition on personal photos uploaded to social media platforms must resist all future facial recognition systems used by the platforms, including possible systems trained explicitly to negate these adversarial perturbations [331]. The experiments presented in Chapter 3 demonstrate the difficulty of this scenario, as the proposed VoiceBlock system shows degraded effectiveness against an unseen speaker recognition system (Table 3.1) and against the enrollment of adversarial utterances (Table 3.3).
2. **Robustness:** post-hoc imperceptible perturbations must maintain effectiveness even under processing designed to remove these perturbations. Concurrent with the work presented in this chapter, researchers have found that analogous image-modality neural network-based processing can negate both adversarial [332] and cooperative [333] imperceptible post-hoc interventions.

My research has sought to address these issues of transferability and robustness in the audio modality. After the publication of VoiceBlock, I conducted follow-up work with a group of researchers

exploring an adversarial intervention designed to impair the function of music generation models when conditioned on the works of artists without consent [249]. By optimizing perturbations to shift the distribution of extracted latents on which generative models operated, our approach aimed to provide musicians with a protective processing tool that would degrade the outputs of generative models performing style transfer or continuation on perturbed recordings. However, we found that optimization-based approaches struggled to simultaneously maintain perturbation imperceptibility and effectiveness against multiple generative models (i.e., transferability). While crafting transferable adversarial interventions remains challenging, there have been indications that convergence of deep learning practitioners towards a narrow range of architectures and pretrained foundation models may make such interventions more feasible [334].

If there remains some slim hope for transferability, the experiments presented in this chapter show that robustness is another issue entirely. In my prior work *Effective and Inconspicuous Over-the-Air Adversarial Examples with Adaptive Filtering* [181], I found that adversarial perturbations could be optimized to obtain some degree of robustness against realistic audio transformations such as noise and reverberation while maintaining reasonable imperceptibility – just as with post-hoc watermarks like MaskMark. In VoiceBlock, I found that carefully chosen perturbation mechanisms (e.g., frequency-domain filtering) can improve robustness even in the absence of explicit optimization, as indicated by our VoiceBlock evaluation in Table 3.2. However, the advent of powerful neural network-based audio codecs and general reconstruction models threatens to nullify the robustness advantages these methods established against acoustic and signal-processing transformations. The results of our watermark robustness evaluations above indicate a strong trend in the development of neural network-based codecs: stronger modeling techniques allow for decreases in bitrate (or equivalently, increases in compression ratio) without corresponding reductions in audio quality. Thus, codecs can discard ever more perceptually-irrelevant information – including perturbations produced by post-hoc adversarial and cooperative interventions – during encoding and reconstruct perceptually plausible information during decoding. This holds not just for neural codecs, but more

broadly for systems trained with generative reconstruction objectives such as neural denoisers. These developments place imperceptible post-hoc interventions directly at odds with progress in generative modeling of audio – far from an advantageous position.

5.5 Impact and Follow-on

The work presented in this chapter demonstrates the fragility of imperceptible post-hoc perturbation methods leveraged by adversarial and cooperative interventions when faced with neural network-based processing such as codecs and denoisers. This does not mean that all such interventions are without utility. For example, Fernandez et al. [325] argue convincingly that watermarks need not be maximally robust to serve as detection mechanisms and deterrents against the misuse of generative models. Moreover, in large-scale production environments where high false-positive rates can incur significant costs, watermark-based systems for synthetic media detection may have to balance many considerations beyond robustness to removal attacks [294].

Rather, the fragility of such methods should motivate the development of more robust techniques that can better support algorithmic interventions such as watermarking in challenging real-world circumstances. The study presented in this chapter has been cited in a number of subsequent watermarking, synthetic audio detection, and provenance tracing works in exactly this context [166, 335–340], and has provided a baseline for rigorous evaluations of post-hoc watermark robustness [341, 342].

Going further, the limitations of post-hoc methods naturally suggest a shift to pre-hoc methods that operate *within* rather than *around* neural network-based systems, affording greater flexibility to shape their behavior beneficially. To this end, in the following chapter I introduce *design interventions* and present research laying the groundwork for the embodiment of this approach in generative watermarking schemes for audio.

CHAPTER 6

TOWARDS DESIGN INTERVENTION

In this chapter, I introduce a novel neural codec training method that lays the groundwork for *design interventions* in the form of generative watermarks that operate within deployed audio AI systems. This work was published at ICASSP 2025 as "Code Drift: Towards Idempotent Neural Audio Codecs" [43], and is presented in edited form here with additional discussion and analysis. I use the pronoun "we" in this chapter to acknowledge the contributions of my co-authors Prem Seetharaman, Jiaqi Su, Zeyu Jin, and Bryan Pardo.

In the following sections, I describe the advantages of formulating watermarking as a design intervention; the limitation that *audio codec non-idempotence* places on this formulation; and finally, the proposed method for overcoming this limitation through idempotent codec fine-tuning, opening the door to future work in generative watermarking for audio.

6.1 Watermarking as Design Intervention

The study presented in the previous chapter demonstrated the fragility of post-hoc information-hiding techniques under neural network-based audio transformations, raising questions about the viability of adversarial and cooperative interventions in challenging real-world conditions. In general, these post-hoc methods are restricted to concealing information via low-magnitude perturbations of pre-existing audio, often at signal-to-noise ratios (SNR) on the order of 20-30dB, to avoid degrading perceived quality. As shown through our experiments, this limits the robustness of such methods, as perturbations can be essentially erased through neural network processing without affecting the perceived quality of audio. In summary, **post-hoc perturbations – functioning as adversarial and cooperative interventions – do not modify audio sufficiently to withstand**

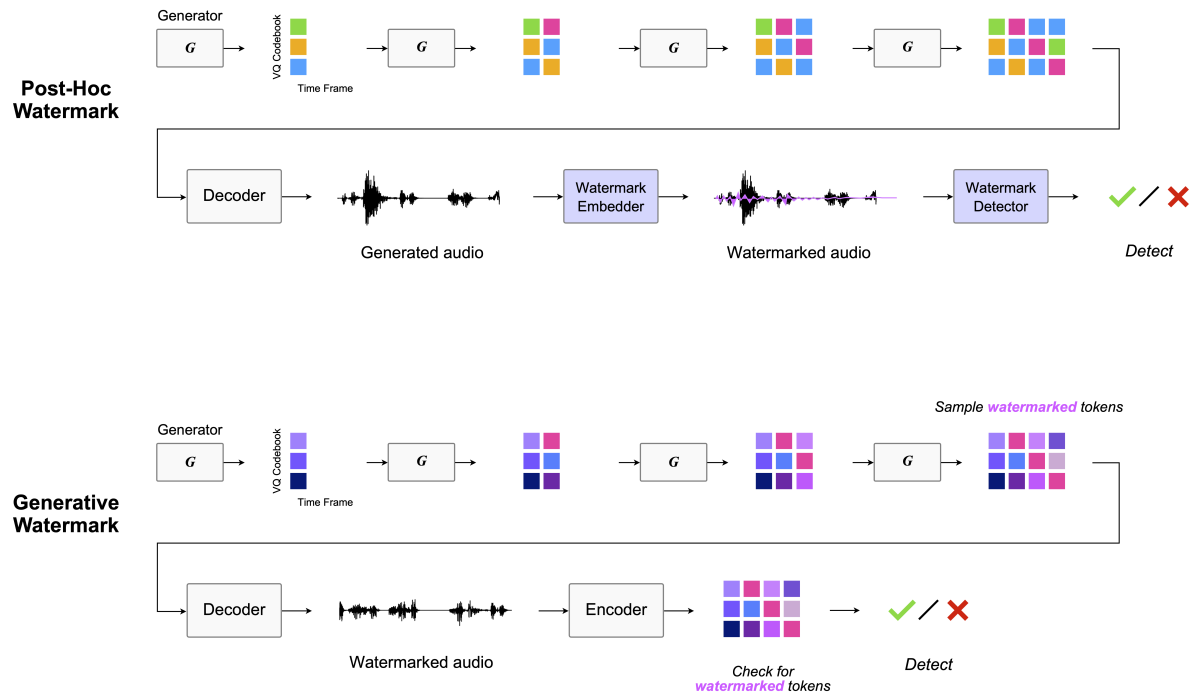


Figure 6.1. Generative watermarking for audio: whereas post-hoc watermarking methods embed signatures in generated audio through imperceptible perturbations (top row), generative watermarks intervene within the generative process to embed statistical signatures (bottom row). Pictured here, a generative watermark intervenes in the sampling of tokens by an autoregressive audio language model; the watermark can then in theory be detected by mapping generated audio back to tokens and performing a statistical test. *Note that if we cannot detect the presence of watermarked tokens after decoding and re-encoding, we cannot detect the generative watermark.*

imperceptible transformations.

How, then, might algorithmic interventions leverage information hiding without inheriting the robustness limitations of post-hoc perturbation methods? Adversarial interventions such as VoiceBlock seek to influence the behavior of an unknown AI system of interest from the outside; cooperative watermarking interventions such as MaskMark seek to influence the behavior of a known AI system (a detector network) while operating alongside a possibly opaque AI system of interest (a generative model). Can we move our intervention one step further, *inside* the AI system of interest, to more substantially influence its inference-time behavior?

This is the role of *design interventions*: algorithmic interventions that incorporate governance objectives directly into the development of AI systems. In contrast to reactive, extrinsic measures such as perturbation-based adversarial attacks that contend with unknown AI systems, design interventions approach harm mitigation from a proactive, intrinsic standpoint by shaping the behavior of AI systems from within. In this chapter, I lay the groundwork for design interventions in the specific context of watermarking for AI-generated media identification; Chapter 7 then discusses the application of design interventions for audio AI systems more broadly.

Watermarking for audio generation – implemented in Chapter 4 as a cooperative intervention in the form of the proposed MaskMark system – can function as a design intervention by modifying the outputs of a generative model *during the generative process* to encode a detectable statistical signature. Unlike post-hoc watermarking techniques, this requires access to the underlying generative model at training and/or inference time in order to enforce constraints on its outputs. While this *generative watermarking* approach sacrifices the the simplicity and modularity of post-hoc methods – which can be applied to the outputs of arbitrary, unknown generative models – it offers a route to greatly improve robustness. As noted above, post-hoc watermarks are restricted to embedding signatures via low-magnitude perturbations of generated instances to avoid degrading quality. By contrast, generative watermarks are free to embed signatures in the structure of high-magnitude, “high-level” features of instances as they emerge through the generation process, *so long as these*

instances remain plausible to users of the generative model. For instance, a post-hoc audio watermark might add filtering-like artifacts to a generated speech recording, while a generative audio watermark might produce a separate speech recording with subtly different stress, timing, and inflection given the same prompt. Because these features are high-magnitude and perceptually relevant, they will likely survive neural network-based transformations such as those evaluated in Chapter 5; and if these features are sampled plausibly, the user of the generative model will not notice any deterioration in the quality of its outputs. A hypothetical generative watermarking scheme for audio language models is pictured in Figure 6.1.

This strategy of watermarking as design intervention has already borne fruit in non-audio modalities. Researchers have demonstrated a number of generative watermarking techniques for text and image generation models [33, 288] with promising robustness under transformations known to remove existing post-hoc watermarks [333], indicating that “plausible sampling” is a more flexible constraint on the watermarking process than “imperceptible perturbation.” However, *attempts to adapt such generative watermarking techniques to the audio modality have generally failed.* In the following sections, I identify a major factor in this failure – arising from the deep representation learning methods that underlie modern audio generation architectures – and demonstrate a novel method for overcoming it, paving the way for work on the adoption of popular generative watermarking techniques from the text modality for audio language models.

6.2 Codec Idempotence: A Barrier to Generative Audio Watermarks

Numerous works have shown that watermarks can be embedded in the text produced by an autoregressive language model by biasing or otherwise constraining its sampling of tokens according to a secret watermarking criterion [33, 343–345], akin to the method illustrated in Figure 6.1. The application of these watermarking methods relies on the assumption that a significant fraction of the tokens observed during watermark detection for a generated piece of text will match those sampled during its (watermarked) generation. While this is trivially true for text – where tokens generally

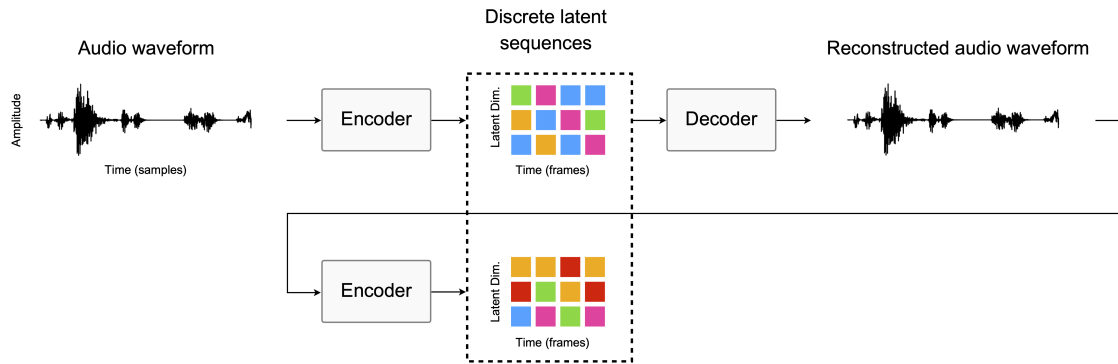


Figure 6.2. Non-idempotence in neural codecs: decoding the tokens produced by a neural codec to audio and subsequently re-encoding may result in a substantially different set of tokens, hampering watermarking methods for audio generative models that rely on the sampling and recognition of specific tokens.

correspond to word-like units, and conversion between raw text and tokens involves a simple lookup – we find that it is not the case for audio. Recall from Section 2.1.3 that audio language models rely on neural network-based tokenizers or *codecs* to map between audio waveforms and a compressed set of discrete tokens. As we show in the following sections, these mappings are *non-idempotent* – the repeated application of decoding and encoding results in different sets of tokens for the same starting audio recording. Thus, while a generative watermarking algorithm may carefully sample a sequence of tokens to conceal a statistical signature, once these tokens are decoded (to obtain audio for use) and re-encoded (to detect the watermark), the detection algorithm may observe an entirely different sequence of tokens and therefore fail to recognize the watermark. In practice, this renders common language model watermarking methods ineffective for audio, even in the complete absence of transformations like those evaluated in Chapter 5 [273, 346]. Moreover, this lack of idempotence has implications beyond generative watermarking, as unstable token mappings can result in degraded audio reconstructions after multiple encodings – potentially rendering neural audio codecs unsuitable for practical file compression despite their ostensible strengths.

To overcome this limitation and facilitate the development of generative watermarks in audio, we analyze the (non-)idempotence of neural audio codecs and propose a novel method for improving

codec idempotence through a simple fine-tuning process. Our experiments show that:

1. State-of-the-art neural audio codecs exhibit varied degrees of idempotence, as some codecs preserve fewer than 50% of tokens in a round-trip mapping and degrade audio outputs significantly after as few as three encodings.
2. The idempotence of neural audio codecs can be improved significantly through fine-tuning with a tailored objective while maintaining strong perceptual quality.
3. This improved idempotence can be achieved without compromising the suitability of a codec for downstream generative modeling, potentially enabling the implementation of language model-style generative watermarks in audio.

The following section discusses relevant prior work in generative watermarking for audio and the design of neural codecs. We then describe our experiments measuring codec idempotence, improving codec idempotence via fine-tuning, and evaluating the effects of idempotence on a codec’s suitability for downstream generative modeling.

6.3 Preliminaries

In this section, we discuss the limitations of prior attempts at generative watermarking for audio and provide background information on neural audio codec design necessary to ground our contributions.

6.3.1 Generative Watermarking for Audio

While a number of attempts were made to incorporate watermarks directly into audio generative models prior to the work presented in this chapter, the resulting methods do not operate at high-level representations but instead induce small signal-level distortions akin to post-hoc watermarks. Juvela et al. propose to incorporate a watermark directly into a HiFi-GAN [292] neural network vocoder; however, because the vocoder only performs "last-mile" generation given strong conditioning

information in the form of a spectrogram, the watermark is still essentially hidden in imperceptible signal-level attributes and as a result lacks robustness [303]. Follow-up work has demonstrated improvement through the use of codec-based transformations during training but still fails to achieve satisfactory robustness [302]. Cho et al. similarly propose to incorporate watermarks into WaveGAN [347] and MelGAN [89] vocoders; however, their watermarks are constrained to the beginning of generated signals, support only a small number of messages, and lack robustness [305]. Liu et al. embed watermarks within a frozen pretrained diffusion vocoder by modifying the initial noise latent and optimizing through multi-step diffusion inference [299]. While the resulting watermark is competitive with state-of-the-art post-hoc approaches, it is constrained to manipulate inaudible signal-level attributes through the use of a spectrogram loss, and the authors do not evaluate its robustness to difficult neural network-based distortions. Other works embed post-hoc watermarks in either the training data of audio generative models or in the latent / codec decoders used to map generated instances to the data space [306–310]. Finally, Ji et al. [311] used a generative model operating on a spectrogram-based VQ-VAE to replace a subset of an audio recording’s encoded latents with "watermarked" latents, then decode to obtain watermarked audio. The generative model responsible for embedding the watermark is not trained with any explicit robustness or message recovery objective; instead, a separate detector network is trained to recover information about the watermarked latents from audio. Possibly due to the lack of cooperative robust training and the independent embedding of watermark bits in latents corresponding to very short slices of audio, the authors only demonstrate robustness to a small set of audio transformations. Fundamentally, the aforementioned methods still embed watermarks via the inaudible modification of signal-level attributes and thus share the robustness limitations of post-hoc approaches.

Concurrent with this work, Défossez et al. [273] and Wu et al. [346] examined the application of language model watermarking methods to audio generative models; both works similarly found that existing methods were ineffective due to a lack of codec idempotence (referred to as the “re-encoded mismatch” by Wu et al.). Thus, codec non-idempotence represents a significant hurdle to

the implementation of generative watermarking as a design intervention.

6.3.2 Neural Audio Codecs

Neural codecs have shown promise in lossy compression of digital media across modalities such as video, images, and audio [63, 348, 349]. Whereas signal-processing codecs rely on hand-crafted algorithms for identifying and removing redundant information within a signal, neural codecs use deep neural networks to learn to encode and decode signals from data directly [62–64]. In the audio modality, neural codecs typically encode waveform audio into a hierarchy of token sequences through residual vector quantization (RVQ), and are trained end-to-end with objectives for both reconstruction quality and regularization of the encoded representations. High compression ratios can be achieved with a combination of small RVQ codebooks (e.g. $2^{10} - 2^{12}$ possible tokens) and low token sequence rates (e.g. 100 – 1000 times less than the source audio). For more details on the design of RVQ codecs, see Section 2.1.3.

Recent works have sought to modify the standard RVQ codec design to disentangle “semantic” attributes of audio such as prosody and linguistic content in encoded representations [268, 321], and to augment encoded token sequences with time-invariant global information to better capture qualities such as speaker identity [268, 322]. While the aforementioned codecs operate directly on waveform audio representations, other works have found advantages in modeling magnitude and phase-magnitude spectrogram representations [104, 319]. In this work, we evaluate neural audio codecs from each of the above categories – RVQ waveform codecs, disentangled “semantic” codecs, and spectrogram codecs.

Because codec token representations are far more compact than waveform audio, while retaining most perceptually relevant information, it is common to train generative language models of audio directly on token sequences [3, 87]. Predicted token sequences can then be mapped to waveform audio via the decoder of the chosen codec, enabling efficient modeling of music [3, 78], speech [58, 106, 321], and sound effects [350]. In this work, we evaluate the performance of fine-tuned

idempotent codecs on a simple conditional speech generation task (vocoding) in which tokens are predicted from a heavily compressed mel-frequency cepstrum coefficient (MFCC) representation.

6.3.3 Studies of Neural Codec Properties

Recent works have evaluated the utility of neural audio codec representations for both discriminative and generative downstream tasks [351–354]. Others have focused on identifying specific properties within these representations and understanding their impact on downstream performance. Lemercier et al. [355] find that the codebooks learned by Encodec [63] exhibit conditional dependence and show that downstream generative modeling performance can be improved by reducing this dependence via an additional training objective. Similarly, we evaluate one specific codec property (idempotence) and examine its relationship to downstream generative modeling performance.

Neural codec idempotence has been previously studied in the image modality. Zhang et al. [356] introduce the notion of “strong idempotence” – stability to re-encoding across different bandwidths – and evaluate both traditional and neural image codec models. Xu et al. [357] explore the relationship between perceptual image compression and codec idempotence, and show that high-quality compression can be achieved by pairing a neural codec with an unconditional generative model through an “inversion” optimization. Li et al. [358] relax invertibility constraints on normalizing flow-based image compression to obtain a fully idempotent codec and improve the trade-off between compression ratio and perceptual transparency in near-idempotent variants. Kim et al. [359] evaluate the idempotence of neural image codecs and propose an auxiliary training objective to improve idempotence; we evaluate similar objectives in our experiments. In this work, we extend the study of neural codec idempotence to the audio modality. We note that while ours is the first published work to identify and mitigate non-idempotence in neural audio codecs, concurrent works identified this phenomenon in the context of watermarking for speech generation [273, 346].

6.4 Measuring Idempotence

We begin by evaluating the idempotence of state-of-the-art neural audio codecs in terms of the stability of both audio outputs and encoded representations under multiple rounds of encoding.

Codecs: We consider the waveform RVQ codecs Encodec (in the original configurations, as well as the 16kHz speech-domain variant trained by Peng et al.) [63, 323] and Descript Audio Codec (DAC) [64]; the “semantic” codecs SpeechTokenizer [321] and FACodec [268]; and the spectrogram-based spectral codecs of Langman et al. [104] and ESC [319]. We use the available pre-trained weights for all codecs, and evaluate each codec at the maximum supported bitrate. We note that these codecs differ substantially in native sample rate, compression ratio, training data, and architecture, making exact performance comparisons difficult; nevertheless, we find strong shared trends in the idempotence of audio outputs and encoded representations. Finally, as baselines we evaluate the signal-processing MP3 [314] and OPUS [315] codecs at low bitrates (see Table 6.1).

Datasets While many codecs are trained for both speech and general-purpose audio compression [63, 64], we focus on speech data to facilitate comparisons. We evaluate codecs on the full-bandwidth speech datasets Espresso [360], containing spontaneous and read speech across four speakers; and VCTK [279], containing read speech across 109 speakers. For each dataset, we select 500 one-second excerpts for evaluation and resample audio to the native sample rate of each codec. For the signal-processing baselines, we select 100 excerpts, due to longer encoding times.

Metrics: To measure the stability of output audio under multiple rounds of encoding, we use the full-reference metrics *Scale-Invariant Signal-to-Distortion ratio (SI-SDR)* [361] and *Perceptual Evaluation of Speech Quality (PESQ)* [237]. PESQ estimates “perceptually-weighted” similarity between the original audio and encoded/decoded audio (up to 8kHz), while SI-SDR operates on the available bandwidth. Given the original audio input as a reference, codecs with high idempotence should produce reconstructions with high PESQ and SI-SDR scores even after multiple encoding iterations. To analyze drift in the internal token representations of neural codecs after multiple

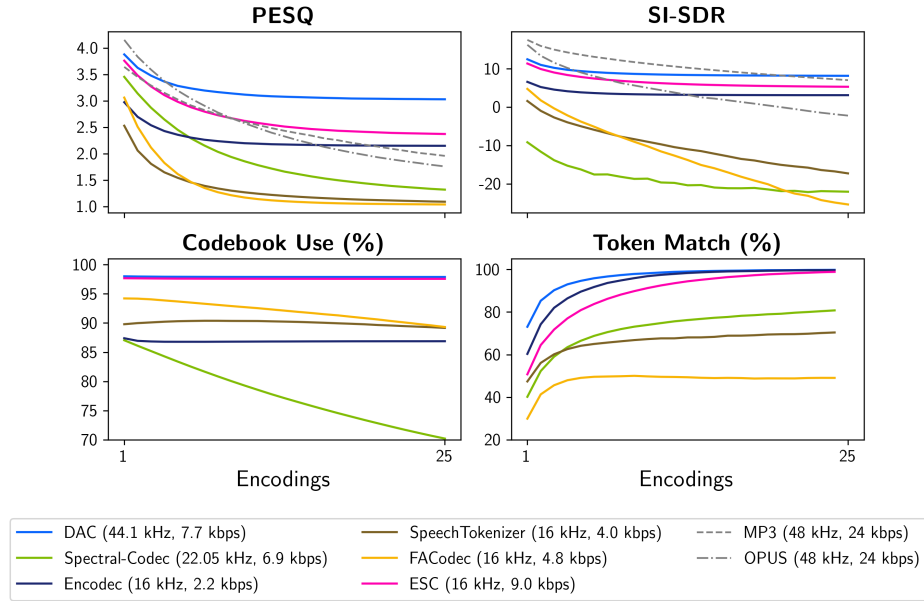


Figure 6.3. Audio and code stability over encodings: audio quality (in PESQ and SI-SDR) of selected neural codecs by encoding iteration, average codebook use by encoding iteration, and match rate between tokens encoded at the n^{th} and $(n + 1)^{\text{th}}$ iterations.

encoding iterations, we compute the *match rate* between tokens at each separate codebook level of the encoded representation after n and $(n + 1)$ encoding iterations. This value ranges from 100% when all tokens match to 0% when none do. While this is a coarse measure that does not account for codebook size or latent similarity, it allows us to observe the convergence of certain codecs towards “stable tokenizations” over multiple iterations. We also measure *codebook use* (the entropy of encoded token distributions as a percentage of the maximum achievable codebook entropy) to study whether codecs converge to small subsets of tokens after multiple encoding iterations. Higher entropy indicates broader codebook use.

In Table 6.1, we compare the performance of our selected codecs in terms of PESQ scores after a single encoding (indicating perceptual transparency) and after 25 encodings (indicating idempotence). Values are grey for datasets seen during codec training. Where we evaluate multiple variants of a codec (e.g. at different bitrates) we separate entries with slashes. In all experiments we normalize codec outputs to match the input volume. In Figure 6.3 we select the variant of each

Table 6.1. Comparison of Codecs

Codec	Sample Rate (kHz)	Bitrate (kbps)	VCTK		Expresso	
			PESQ @ 1 ↑	PESQ @ 25 ↑	PESQ @ 1 ↑	PESQ @ 25 ↑
MP3 [314]	48	24 / 12 / 6	3.79 / 3.79 / 3.41	1.99 / 1.99 / 2.29	3.49 / 3.49 / 3.32	1.93 / 1.93 / 2.13
OPUS [315]	48	24 / 12 / 6	4.17 / 3.72 / 2.30	1.82 / 1.40 / 1.04	4.13 / 3.58 / 2.12	1.69 / 1.33 / 1.04
DAC [64]	44.1 / 24 / 16	7.7 / 24 / 6	3.97 / 4.30 / 4.04	3.17 / 1.28 / 1.32	3.78 / 4.29 / 3.91	2.89 / 1.24 / 1.27
Encodec [63, 323]	48 / 24 / 16	24 / 24 / 2.2	2.96 / 3.65 / 3.03	1.05 / 1.05 / 2.21	2.87 / 3.34 / 2.91	1.05 / 1.05 / 2.08
SpeechTokenizer [321]	16	4	2.65	1.10	2.41	1.07
FACodec [268]	16	4.8	3.09	1.03	3.03	1.04
Spectral-Codec [104]	44.1 / 22.05	6.9 / 6.9	3.24 / 3.49	1.18 / 1.36	3.16 / 3.41	1.14 / 1.27
ESC [319]	16	9	3.85	2.50	3.67	2.24

neural codec with the highest demonstrated idempotence and plot audio quality, codebook use, and token match rate as a function of the number of encoding iterations. In Figure 6.4 we plot the token match rate between encodings of an input under small time shifts, indicating the sensitivity of the token representation to phase, and compute the correlation between average observed sensitivity and idempotence as measured by PESQ score after 25 encoding iterations.

The results of our experiments show that the idempotence of neural audio codecs varies greatly; DAC, ESC, and the Peng et al. variant of Encodec fare best, with DAC in particular demonstrating high audio quality even after 25 encoding iterations. Codec type alone (RVQ, semantic, or spectrogram) does not dictate performance. As can be seen in Figure 6.3, the behavior of token representations under multiple encodings also varies significantly across codecs. In codecs exhibiting the highest idempotence (DAC, ESC), codebook use remains virtually unchanged, while token match rate increases – indicating that even as these codecs converge to nearly fixed tokenizations, they do not collapse to a small subset of possible tokens. By contrast, Spectral-Codec and FACodec converge more slowly or not at all, while codebook use declines noticeably.

The observed differences in idempotence are not completely explained by the single-encoding perceptual transparency of codecs, as indicated by both Figure 6.3 and Table 6.1. For instance, DAC 24kHz and DAC 16kHz have high perceptual transparency (as indicated by PESQ and SI-SDR scores after one encoding iteration) but exhibit very low idempotence after 25 iterations.

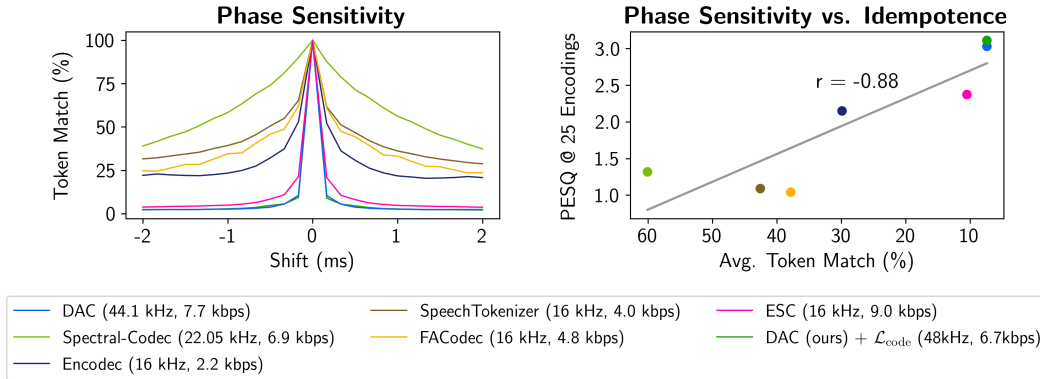


Figure 6.4. Phase sensitivity correlates with idempotence: For each selected codec, we compute the token match rate between encodings of audio inputs under time shifts between ± 2 ms. We then compute the correlation between average observed match rate and PESQ score after 25 encoding iterations. In general, codecs with higher phase sensitivity (as indicated by lower match rates) exhibit higher idempotence (as indicated by higher PESQ scores).

Phase sensitivity seems to play a larger role: as shown in Figure 6.4, the sensitivity of token representations to phase shift positively correlates with idempotence. This is counterintuitive, as we might expect more idempotent codecs to be invariant to imperceptible perturbations such as phase shift. However, from our experiments it appears that fine-grained encoding of phase information is linked to the preservation of audio quality over multiple encoding iterations.

6.5 Encouraging Idempotence

Given the wide range of idempotence observed across neural audio codecs, we explore whether it is possible to improve idempotence without sacrificing perceptual transparency or fundamentally altering the compression scheme. We focus our efforts on DAC, as it shows the highest overall idempotence of audio and encoded representations in our experiments.

For a neural codec f and input signal x , idempotence can be expressed as the relationship

$$f(x) = f(f(x))$$

A neural codec f typically consists of an encoder network f_e , decoder network f_d , and quantization module f_q such that its output can be expressed as $f(x) = f_d \circ f_q \circ f_e(x)$, where $f_q(x)$ is the quantized encoded representation. For DAC codec f and input x , we can decompose the compression and decompression process into individual steps. Taking $x' := f(x)$, we have:

$$x' = f_d \circ f_{dp} \circ f_{dq} \circ f_q \circ f_p \circ f_e(x) \quad (6.1)$$

where f_e is a neural network encoding waveform audio x into frame-level continuous latents; f_p is a linear projection into a low-dimension space for codebook lookup; f_q is a L_2 -normalized nearest-neighbors lookup over codebook vectors, returning codebook indices; f_{dq} indexes the codebook and to obtain codebook vectors; f_{dp} is a linear projection of the indexed codebook vectors back to the latent space; and f_d is a neural network that upsamples the resulting latent sequence to obtain waveform audio. Because the vector quantization operation $f_{dq} \circ f_q$ is non-differentiable with respect to the input, the straight-through gradient estimator is used during backpropagation. DAC is trained with a combination of adversarial and spectrogram losses to encourage perceptual transparency as well as the standard codebook and commitment losses [72]. For additional details we refer the reader to Kumar et al. [64].

Similar to Kim et al. [359], we encourage idempotent representations through the use of an additional regularizing objective during codec training. We explore three regularizing losses for enforcing idempotence at different stages of the compression process.

1. **Encoder idempotence**, the distance between continuous latents produced by the first and second encoding of a signal:

$$\mathcal{L}_{\text{enc}} = \|f_e(x') - f_e(x)\|_2$$

2. **Projected idempotence**, the distance between projected latents produced by the first and second encoding of a signal:

$$\mathcal{L}_{\text{proj}} = \|f_p \circ f_e(x') - f_p \circ f_e(x)\|_2$$

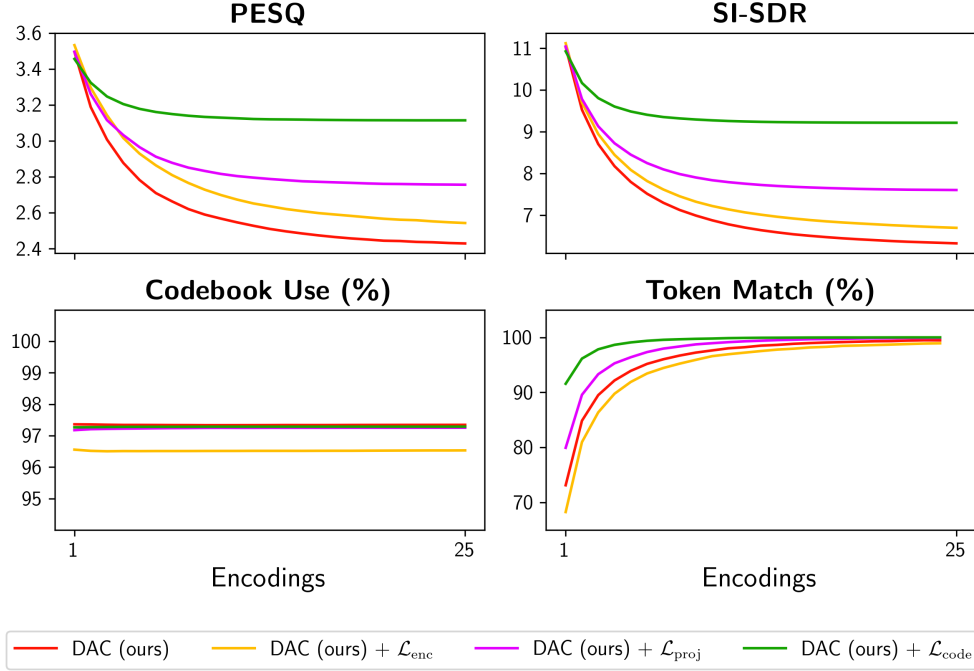


Figure 6.5. Effect of idempotent training: audio quality (in PESQ and SI-SDR), average codebook use, and match rate between tokens encoded at the n^{th} and $(n + 1)^{\text{th}}$ iterations as a function of encoding iterations for our DAC reproduction and fine-tuned variants with idempotence losses described in Section 6.5. Fine-tuning with $\mathcal{L}_{code}$ results in the highest overall idempotence through 25 encoding iterations, while preserving codebook use.

3. **Token idempotence**, the distance between projected latents from the second encoding and the codebook vectors selected in the first encoding:

$$\mathcal{L}_{code} = \|f_p \circ f_e(x') - f_{dq} \circ f_q \circ f_p \circ f_e(x)\|_2$$

All losses are averaged over the time dimension of the representation at which they are computed. We note that the encoder idempotence loss is most similar to the idempotence loss proposed by Kim et al. [359], which is computed on the quantized latents of the first encoding; we find this degrades audio quality significantly and thus use the un-quantized latents instead.

To evaluate the effect of our proposed idempotence losses, we train a reproduction of DAC from scratch for 500k steps on a mixture of openly-licensed speech data (CommonVoice [362], LibriVox [363], VoxPopuli [364]) and proprietary licensed sound effects and music data. Our

DAC operates at 48kHz with a hop length of 640 samples, resulting in a token rate of 75Hz and bitrate of 6.7kbps; all other aspects of the model and training are identical to that of Kumar et al. After pretraining we freeze the quantizer module, disable codebook dropout, introduce one of the idempotence losses described above weighted by a constant λ , and fine-tune for a further 50k steps. To select λ we sweep candidate values of $\{1, 10, 100, 1000\}$ and select the largest value for which the pretrained model’s audio quality is preserved (as measured by validation loss). We use $\lambda = 1$ for $\mathcal{L}_{enc}$, $\lambda = 10$ for $\mathcal{L}_{proj}$, and $\lambda = 100$ for $\mathcal{L}_{code}$. In Figure 6.5 we perform the multiple-encoding experiment from Section 6.4 with our fine-tuned DAC variants and find that the choice of loss has a significant effect on the stability of output audio and encoded representations. All idempotent fine-tuned variants exhibit improved stability in audio quality under re-encoding when compared to our DAC reproduction. The codebook-space objective $\mathcal{L}_{code}$ performs best, resulting in superior audio quality to all evaluated codecs (including the original DAC model of Kumar et al.) after 25 encoding iterations. The projected objective $\mathcal{L}_{proj}$ is less effective, followed by the latent objective $\mathcal{L}_{enc}$.

6.6 Downstream Generative Modeling

While idempotence is in itself valuable for applications such as file compression, we are interested in its implications for generative watermarking of language models trained to predict codec tokens. An idempotent codec that significantly degrades downstream modeling performance, regardless of its suitability for watermarking, will not be adopted widely within generative models. Accordingly, we evaluate the impact of our idempotent training configurations on a simple generative modeling task. Using token representations obtained from DAC, our DAC reproduction, and our fine-tuned variants, we train SoundStorm-like [106] masked generative transformer models to generate speech conditioned on 30-bin MFCC representations computed at the codec frame rate. This can be thought of as a vocoding task, albeit conditioned on a more compressed speech representation than the typical 80-120 band mel-spectrogram features [269, 270], making token prediction difficult.

Table 6.2. Generative Modeling on Codec Tokens

Codec	SQUIM-MOS $\uparrow$	PESQ $\uparrow$	ASR-CER $\downarrow$
DAC (ours)	4.39	1.68	0.07
DAC (ours) + $\mathcal{L}_{\text{enc}}$	4.34	1.66	0.06
DAC (ours) + $\mathcal{L}_{\text{proj}}$	4.36	1.67	0.07
DAC (ours) + $\mathcal{L}_{\text{code}}$	4.26	1.65	0.06

We apply a proprietary bandwidth extension algorithm to the LJSpeech dataset [365] to obtain roughly 24 hours of high-quality single-speaker read speech at 48kHz, allowing us to fully utilize the bandwidth of our DAC-based codecs. We train all models for 100k iterations on 5-second random excerpts at a batch size of 32 using the AdamW [366] optimizer with learning rate $1e-4$, and perform conditioning dropout with probability 0.2. We use small 43m-parameter transformer models consisting of 12 standard transformer encoder blocks with 8 attention heads each and apply rotary positional encodings in each block. During inference, we proceed from the lowest to highest codebook level following the “MaskGIT” strategy used by Yang et al. [58]; we use the confidence-based masking scheme from García et al. [78] with mask temperature 10.5, sample with temperature 1.0 and nucleus 0.85 [367], and apply classifier-free guidance with weight 2.0 [368]. For each model we generate 100 samples using MFCC feature sequences from unseen recordings and compute the following metrics: *SQUIM Mean Opinion Score (SQUIM-MOS)* [328], a deep neural network-based estimate of speech quality that does not require precise alignment to the reference audio; *Character Error Rate (ASR-CER)* of a Wav2Vec2.0 [77] speech recognition system’s predicted transcripts on generated audio, which measures the intelligibility of generated speech; and PESQ.

The results of our evaluations are presented in Table 6.2. Due to the information bottleneck of the MFCC representation, generations do not match the precise spectral structure of the source audio, as indicated by low PESQ scores. However, we find the generations retain high audio quality and intelligibility, as indicated by SQUIM-MOS and ASR-CER scores. Generative models trained on codecs fine-tuned with idempotence objectives produce outputs of similar quality to a model

trained on the base codec, indicating that the proposed idempotence losses do not have a strong negative effect on generative modeling.

6.7 Discussion

The experimental results presented above show that (1) state-of-the-art neural audio codecs lack idempotence, exhibiting unstable token mappings and audio degradation under multiple rounds of encoding; (2) codec idempotence can be increased through fine-tuning with our proposed token idempotence loss $\mathcal{L}_{\text{code}}$ to achieve round-trip token match rates above 90%; and (3) this improved idempotence does not come at significant cost to perceptual transparency (as indicated through signal reconstruction metrics) or downstream generative modeling performance (as demonstrated through masked language modeling of codec tokens for MFCC vocoding).

These results lay a strong foundation for future work applying generative watermarking techniques to idempotence-enhanced neural audio codecs, as our fine-tuning method offers a simple way to increase token match rates without major architectural or training modifications to existing codecs and without compromising desirable distributional properties of encoded token representations. While future generative watermarking methods may benefit from additional codec properties (e.g. some degree of transformation invariance or robustness in token mappings), there is evidence that merely stabilizing tokenizations via fine-tuning provides robustness benefits for language model-style watermarks [369], including against unseen neural network-based codecs. In the following chapter, I discuss follow-up work in generative watermarking and design interventions within audio AI systems more generally.

Finally, beyond potentially supporting design interventions in the form of generative watermarks, our proposed fine-tuning method also offers to improve the practicality of neural audio codecs in low-bitrate lossy file-compression scenarios, audio editing tasks [58], and iterative creative interfaces [117] where multiple encodings may arise naturally.

CHAPTER 7

CONCLUSION

This dissertation introduces a number of contributions to the study of algorithmic interventions intended to mitigate harms posed by highly capable audio AI systems. In Chapter 3, I presented VoiceBlock, a novel adversarial intervention for anonymizing voice communications in real-time to evade surveilling speaker recognition systems. In Chapter 4, I presented MaskMark, a novel cooperative intervention for watermarking speech generated by audio AI systems to facilitate provenance tracing. In Chapter 5, I presented a study analyzing the robustness of cooperative watermarking interventions and identifying key limitations of post-hoc information hiding methods. Finally, in Chapter 6, I identified a technical limitation of neural audio codecs hindering the application of generative watermarking methods in the audio modality and presented a codec fine-tuning method to overcome this limitation, paving the way for robust watermarks that operate as design interventions within audio generation systems.

These works were undertaken from 2022 – 2025 in roughly the order in which they are presented here. Viewed narrowly, they developed novel techniques to constrain, evaluate, or improve the function of neural network-based audio processing systems in information hiding scenarios, often leveraging domain knowledge (e.g. frequency domain filtering mechanisms) to find performant and efficient solutions within academic resource constraints. More broadly though, these works serve as snapshots of the research community’s reactions to rapidly expanding AI capabilities and corresponding emergent harms.

In the following sections, I (1) review the research contributions embodied in these works, (2) highlight important themes within these works that may be relevant to future research in the development of algorithmic interventions, and (3) provide closing remarks addressing the three

main questions raised in the opening chapter of the dissertation.

7.1 Summary of Research Contributions

The **VoiceBlock system presented in Chapter 3** was developed amid a wave of interest in adversarial interventions against biometric surveillance systems [9, 208, 248]. My proposed method outperformed state-of-the-art adversarial attacks in imperceptibly evading speaker recognition systems, demonstrated the effectiveness of filter-based perturbation methods for frequency-domain information hiding, and took an encouraging step towards the practical application of adversarial interventions for speech communications via a streaming implementation capable of operating in real-time on a laptop CPU. This work also identified key criteria for real-world deployment of adversarial interventions, including robustness to neural network-based processing, transferability to unseen systems, and resilience against adaptive systems that incorporate adversarially perturbed data to weaken protections. As subsequent works (including my follow-up research on music data protection [249]) continued to explore adversarial interventions against irresponsibly deployed AI systems of increasing capability, researchers began to better understand the difficulty of satisfying these criteria. Fundamentally, adversarial interventions face an asymmetrically difficult task: they must maintain their effectiveness against previously unseen AI systems, developed and updated by motivated actors, in unpredictable real-world scenarios, *in perpetuity* [41, 331]. Adversarial interventions that fail to grapple with this reality risk serving as “token fixes,” potentially misleading vulnerable users as to the degree of protection they afford from AI harms [41].

In response to mounting evidence of this asymmetric difficulty in my own research and the broader literature, I sought to develop algorithmic interventions that could operate under more favorable conditions. **MaskMark, presented in Chapter 4**, represents a pivot from the adversarial posture of VoiceBlock towards a cooperative posture in two ways. First, the proposed intervention is intended to function alongside a responsibly deployed AI system (such as a text-to-speech model) rather than in opposition to an irresponsibly deployed system. Second, at the implementation

level, MaskMark modifies the two-network training scheme of VoiceBlock such that a frequency-domain information-hiding network (here, an embedder) is optimized *collaboratively* with a second network to steer its outputs (here, a detector) rather than against a fixed “adversary.” This means that information hiding can be optimized to target a *known* neural network that is itself *explicitly trained to facilitate information recovery* even in the presence of audio transformations, easing robustness and removing the requirement of transferability. By combining this scheme with a VoiceBlock-inspired frequency-domain information hiding mechanism and various optimization improvements, MaskMark outperformed state-of-the-art signal-processing and neural network-based watermarks in robust detection performance while maintaining the perceptual quality of watermarked audio. MaskMark’s strong performance and lightweight architecture demonstrated the viability of end-to-end neural network-based methods for robust watermarking of AI-generated speech, and a large number of concurrent and subsequent works have further advanced the development of watermarking methods for synthetic audio. By the time I followed up this work with a tutorial on speech watermarking at InterSpeech 2025, I had to sift through dozens of analogous two-network watermarking approaches when writing introductory materials.

Cooperative watermarking interventions such as MaskMark promise to aid the safe deployment of audio AI systems through the identification of AI-generated audio. However, the benefits of such methods rely on their ability to accurately discriminate between watermarked and un-watermarked recordings under realistic conditions and in the presence of motivated adversaries who may seek to conceal the provenance of AI-generated recordings. Just as with insufficiently robust adversarial interventions, insufficiently robust watermarking methods risk serving as token fixes that gesture towards responsibility while providing a false sense of security. Based on my initial evaluations of MaskMark’s detection performance under neural network-based audio transformations – which ranged from strong (vocoders) to suspect (neural codecs) – I conjectured that general post-hoc watermarking methods might share vulnerabilities to certain transformation classes.

I explored this phenomenon more fully in the work **Deep Audio Watermarks Are Shallow**,

presented in Chapter 5. Examining a large set of audio transformations and state-of-the-art watermarks including MaskMark, I demonstrated the fragility of imperceptible post-hoc watermarking methods when faced with neural network-based processing. In particular, this work foregrounded the adverse dynamic between neural codecs – which seek to discard perceptually irrelevant details as part of a compression process – and post-hoc audio watermarks, which seek to conceal watermark signatures in exactly these perceptually irrelevant details. I also found that improved modeling techniques allowed for the continued development of neural codecs with higher compression ratios (and correspondingly stronger watermark removal performance) without significant tradeoffs in perceptual quality. While post-hoc watermarking methods such as MaskMark can still be broadly useful for harm mitigation even with clear robustness limitations [325], this trend suggested a need for more robust watermarking methods capable of surviving even highly disruptive neural network-based processing.

During the development of the filter-based attack on which VoiceBlock was built [181], I had briefly conducted experiments concealing adversarial perturbations within high-magnitude but realistic audio transformations such as reverberation and textured noise, with the goal of improving attack robustness by moving beyond imperceptible perturbation strategies. Returning to this idea, I sought methods for concealing high-magnitude watermarks within plausible variations of audio generation processes. I was encouraged by early works on generative watermarking of autoregressive language models [33, 370, 371], as I believed that if applied to audio such methods could embed robust signatures at the the level of perceptually relevant content (e.g., timing and pronunciation variations in speech) rather than perceptually irrelevant noise. In late 2023, I began work applying these methods to audio language models [3]. However, I immediately encountered limitations with such methods due to the non-idempotence of the neural audio codecs used to obtain tokens for audio language modeling.

This realization led to the **Code Drift work presented in Chapter 6**, in which I analyzed the non-idempotence of neural audio codecs with the goal of enabling more robust language model-style

generative watermarks for audio. This was the first published work to identify non-idempotence in neural audio codecs, and introduced a novel fine-tuning method capable of improving codec idempotence without significant loss of perceptual transparency or downstream generative modeling performance. Through this work, I sought to remove a barrier – also described but not yet overcome in parallel works [273, 346] – to the implementation of watermarking as a design intervention operating within audio AI systems. Subsequent audio modality works have furthered this line of research by demonstrating cluster-based techniques for overcoming codec non-idempotence, allowing for the application of robust language-modeling watermarks [372, 373]. While Code Drift serves as the final publication component of this dissertation, it marks a first step in my research towards proactive design interventions that shape the behaviors of audio AI systems from within.

7.2 Common Research Themes

Taken together, the research projects presented in this dissertation provide a number of lessons for the design and evaluation of neural network-based information hiding systems in the audio modality, and for the development algorithmic interventions more broadly. Below, I highlight important themes that may be relevant to future research in these areas.

Adversarial interventions face an uphill climb. Adversarial interventions seek to constrain or impair the operation of irresponsibly deployed AI systems by modifying the data ingested by these systems at training or inference time. As discussed in Section 5.4.2, this places a heavy burden on the developers of adversarial interventions. “Protective” algorithms such as VoiceBlock must contend with unknown AI systems and defenses designed to remove adversarial perturbations, and in settings such as de-identification and intellectual property protection *must remain effective indefinitely* in order to prevent the misuse of user data released to malicious actors [41, 331]. While cooperative interventions and design interventions must still contend with malicious actors (e.g., those seeking to remove watermarks), these approaches benefit by operating in a configuration

in which both the intervention algorithm and AI system of interest are known and controlled by (assumed) responsible actors.

Perturb the right representation. Information-hiding interventions such as VoiceBlock and MaskMark operate by perturbing audio signal representations. Whereas prior works had optimized additive perturbations at a waveform representation [9, 374], such perturbations tend to manifest as uncorrelated, easily perceptible noise unless heavily regularized with auxiliary loss functions or multi-stage optimization schemes [181, 375]. VoiceBlock and MaskMark show that *multiplicative* perturbations at a *frequency-domain* representation can be made to manifest as smooth, inconspicuous modifications of the underlying audio without relying on complicated optimization procedures because frequency-domain multiplication produces perturbations highly correlated with the source audio. In effect, the choice of representation acts as a powerful regularizer, allowing for scaling perturbation magnitude (and correspondingly, strength) beyond additive approaches while remaining perceptually inconspicuous [42, 181]. Subsequent watermarking research has further illustrated the benefits of perturbing naturally regularized representations, such as latent spaces within generative models [288].

Intervene within the model where possible. Similarly to the choice of perturbation representation, the “location” of an algorithmic intervention can have implications for its robustness and effectiveness. The post-hoc information-hiding approaches explored in this dissertation seek to preserve the utility of the voice communications and AI audio system outputs on which they act by ensuring that any perturbations introduced remain human-imperceptible. However, this induces a tight constraint on the magnitude of perturbations and renders them vulnerable to neural network-based audio transformations, as discussed in Chapter 5. While post-hoc perturbation approaches are fundamentally “stuck” with the preexisting audio they are tasked with modifying, design interventions within generative models offer more leeway by affording the ability to proactively shape generated audio – loosening the action constraint from highly restrictive “imperceptible modification” to more fuzzy “quality preservation.” For example, a generative watermark within a

text-to-speech system may still be required to render the exact words specified by a user, but can render those words in a variety of ways that differ substantially in their acoustic realization (e.g., as measured by audio distances such as SNR). Works in the image and audio modalities have shown that such interventions *within* generative models can allow for more robust, high-magnitude watermark signatures that survive transformations known to remove imperceptible post-hoc watermarks [288, 373]. This general approach may also be beneficial beyond information-hiding techniques, as “upstream” intervention within AI systems allows for shaping behaviors to ends that may be difficult to achieve through “downstream” modification of their outputs.

7.3 Closing Remarks: Building Better Interventions

The studies of audio-modality algorithmic interventions undertaken in this dissertation seek to answer three main questions:

1. Where and why have existing algorithmic interventions fallen short?
2. Where have algorithmic interventions shown promise?
3. How can algorithmic interventions be made more effective going forward?

While the research presented in the preceding chapters focuses specifically on information hiding methods and problem settings of biometric surveillance and synthetic media detection, it points towards answers relevant to broader types of algorithmic intervention in the audio modality.

In my works on adversarial and cooperative interventions, the most common observed failure mode of preexisting and proposed methods was a *lack of robustness to the conditions under which interventions must operate in the real world*. Such robustness limitations can be magnified by the underlying intervention strategy, as with adversarial interventions that must contend asymmetrically with unknown AI systems, and by the choice of techniques with which interventions are carried out, as with imperceptible perturbation approaches that by definition must be vulnerable to capable

audio compression models. In developing algorithmic interventions, it is therefore critical to pick the right battles and the right tools with which to fight them. Synthetic media identification is likely a battle worth fighting, and post-hoc watermarking methods such as MaskMark can still serve as valuable tools [325] – Meta and Google still leverage post-hoc watermarks for AI-generated visual media [376, 377] – but it is also worth exploring methods that more natively support robust operation.

In this respect, design interventions have shown promise in beneficially steering the outputs of AI systems in audio and other modalities while offering stronger robustness to real-world operating conditions. Within synthetic media identification, generative watermarks for text [33] and image generation models [288] have demonstrated practical performance at scale [55] and improved robustness over post-hoc methods [333]. Since the publication of Code Drift, works on generative watermarking in the audio modality have also demonstrated encouraging robustness against neural codec transformations [372, 373]. Beyond synthetic media identification, audio researchers have proposed a number of design interventions that incorporate attribution and compensation mechanisms directly into the inference of AI audio generation systems [166, 167], as well as design interventions that reshape generative interfaces to center expressive control, expose interpretable and manipulable representations, and encourage creative interactions [79]. These works benefit from a proactive, intrinsic approach to intervention within AI systems, allowing them to influence outcomes in ways inaccessible to reactive, extrinsic interventions that sit outside of AI systems.

On the final question – how can algorithmic interventions be made more effective going forward? – the research presented in this dissertation suggests a number of avenues for future work. Initial generative watermarks for audio are applicable only within autoregressive language models, necessitating the development of new methods applicable to other paradigms such as diffusion modeling. Existing language model watermarking approaches may be strengthened by idempotent codec fine-tuning as proposed in Chapter 6. Proposed design interventions for inference-time data attribution and compensation within audio generation models have not yet been evaluated under

realistic large-scale use or attacks by malicious actors, necessitating further robustness studies.

Most importantly, novel algorithmic interventions should be pursued alongside regulatory and economic interventions that provide the conditions for adoption and enforcement where necessary. From the embedding of watermarks to trace AI-generated media to the compensation of artists for participation in generative media ecosystems, future audio-modality algorithmic interventions will rely on the participation of numerous stakeholders to operate effectively: the providers of AI systems for music and audio generation, the users of digital media platforms on which AI-generated media may appear, artists and creative laborers whose livelihoods are impacted by the spread of AI-generated media, and the legal bodies regulating the use, identification, and attribution of AI-generated media. This broad participation will be key to ensuring that future interventions function as more than token fixes for the harms of increasingly capable audio AI systems.

REFERENCES

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML*, 2023.
- [2] J. Thickstun, D. Hall, C. Donahue, and P. Liang, “Anticipatory music transformer,” *TMLR*, 2024.
- [3] J. Copet et al., “Simple and controllable music generation,” in *Neural Information Processing Systems (NeurIPS)*, 2024.
- [4] A. Caillon and P. Esling, “Rave: A variational autoencoder for fast and high-quality neural audio synthesis,” *arXiv preprint arXiv:2111.05011*, 2021.
- [5] Y. Yang et al., “Streamvc: Real-time low-latency voice conversion,” in *ICASSP*, 2024.
- [6] S. Rouard, F. Massa, and A. Défossez, “Hybrid transformers for music source separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [7] E. Moliner et al., “Automatic music mixing using a generative model of effect embeddings,” in *ICASSP*, 2026.
- [8] J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, M. Mancusi, and Y. Mitsufuji, “Ito-master: Inference-time optimization for audio effects modeling of music mastering processors,” in *ISMIR*, 2025.
- [9] M. Chiquier, C. Mao, and C. Vondrick, “Real-time neural voice camouflage,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [10] P. O’Reilly, A. Bugler, K. Bhandari, M. Morrison, and B. Pardo, “Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models,” in *Neural Information Processing Systems (NeurIPS)*, 2022.
- [11] O. Solon, *U.S. Prisons Are Building Biometric Databases of Their Inmates’ Voices*, MIT Technology Review, [Online]. Available: <https://www.technologyreview.com/2019/01/31/137633/us-prisons-are-building-biometric-databases-of-their-inmates-voices/>, Jan. 2019.
- [12] M. Moon, “A new ai voice tool is already being abused to make deepfake celebrity audio clips,” *Engadget*, 2023.
- [13] B. Edwards, “Bruce willis denies selling deepfake rights to deepcake,” *Ars Technica*, 2022. [Online]. Available: <https://arstechnica.com/information-technology/2022/10/bruce-willis-denies-selling-deepfake-rights-to-deepcake/>.
- [14] J. Coscarelli, “An a.i. hit of fake ‘drake’ and ‘the weeknd’ rattles the music world,” *The New York Times*, 2023.

- [15] Deezer, *Deezer: AI-generated tracks now represent 44% of all new uploaded music*, Deezer Newsroom, Accessed: May 11, 2026, Apr. 2026. [Online]. Available: <https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>.
- [16] Bandcamp, *Keeping Bandcamp Human*, Bandcamp Updates, Accessed: May 11, 2026, Jan. 2026. [Online]. Available: <https://blog.bandcamp.com/2026/01/13/keeping-bandcamp-human/>.
- [17] CISAC and PMP Strategy, “Study on the Economic Impact of Generative AI in the Music and Audiovisual Industries,” International Confederation of Societies of Authors, Composers, and PMP Strategy, Tech. Rep., Dec. 2024, Accessed: May 11, 2026. [Online]. Available: <https://www.cisac.org/services/reports-and-research/cisacpmp-strategy-ai-study>.
- [18] B. Brittain, *Voice actors can pursue some claims over AI voiceovers, US court says*, Reuters, Accessed: May 11, 2026, Jul. 2025. [Online]. Available: <https://www.reuters.com/legal/litigation/voice-actors-can-pursue-some-claims-over-ai-voiceovers-us-court-says-2025-07-10/>.
- [19] U.S. Copyright Office, “Copyright and Artificial Intelligence, Part 3: Generative AI Training,” U.S. Copyright Office, Tech. Rep., May 2025, Pre-publication version. Accessed: May 11, 2026. [Online]. Available: <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>.
- [20] New York Comic Con, *AI-Generated Products Policy*, New York Comic Con, Accessed: May 11, 2026, 2026. [Online]. Available: <https://www.newyorkcomiccon.com/en-us/about/ai-generated-products-policy.html>.
- [21] Future of Life Institute, *Pause Giant AI Experiments: An Open Letter*, Future of Life Institute, Accessed: May 11, 2026, Mar. 2023. [Online]. Available: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
- [22] J. Stempel, *Baltimore sues Elon Musk’s xAI over Grok sexual ‘deepfakes’*, Reuters, Accessed: May 11, 2026, Mar. 2026. [Online]. Available: <https://www.reuters.com/legal/government/baltimore-sues-elon-musks-xai-over-grok-sexual-deepfakes-2026-03-24/>.
- [23] Neural Information Processing Systems Foundation, *NeurIPS 2025 Policy on the Use of Large Language Models*, NeurIPS 2025, Accessed: May 11, 2026, 2025. [Online]. Available: <https://neurips.cc/Conferences/2025/LLM>.
- [24] YouTube Help, *Disclosing use of altered or synthetic content*, Google Help, Accessed: May 11, 2026, 2026. [Online]. Available: <https://support.google.com/youtube/answer/14328491>.

- [25] TikTok Shop, *AI-Generated Content Restrictions and Requirements*, TikTok Shop Academy, Accessed: May 11, 2026, Apr. 2026. [Online]. Available: https://seller-us.tiktok.com/university/essay?knowledge_id=491489038501663.
- [26] Human Artistry Campaign, *Principles for Artificial Intelligence Applications in Support of Human Creativity and Accomplishment*, Human Artistry Campaign, Accessed: May 11, 2026, 2023. [Online]. Available: <https://www.humanartistrycampaign.com/>.
- [27] Fairly Trained, *Certified Models*, Fairly Trained, Accessed: May 11, 2026, 2026. [Online]. Available: <https://www.fairlytrained.org/certified-models>.
- [28] J. Barnett, H. F. Garcia, and B. Pardo, “Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model,” in *ISMIR*, 2024.
- [29] J. Deng et al., “Computational copyright: Towards a royalty model for music generative ai,” *arXiv preprint arXiv:2312.06646*, 2025. [Online]. Available: <https://arxiv.org/abs/2312.06646v5>.
- [30] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, “Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models,” in *Proc. 32nd USENIX Security Symposium*, 2023. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity23/presentation/shan>.
- [31] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, “Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models,” in *Proc. IEEE Symposium on Security and Privacy*, 2024. [Online]. Available: <https://arxiv.org/abs/2310.13828>.
- [32] K. Pavlović, S. Kovacević, I. Djurović, and A. Wojciechowski, “Robust speech watermarking by a jointly trained embedder and detector using a dnn,” *Digital Signal Processing*, vol. 122, p. 103 381, 2022, ISSN: 1051-2004. DOI: <https://doi.org/10.1016/j.dsp.2021.103381>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1051200421004206>.
- [33] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *International Conference on Machine Learning (ICML)*, 2023.
- [34] Stanford Institute for Human-Centered Artificial Intelligence, *Labeling AI-Generated Content May Not Change Its Persuasiveness*, Stanford HAI, Accessed: May 11, 2026, Jul. 2025. [Online]. Available: <https://hai.stanford.edu/policy/labeling-ai-generated-content-may-not-change-its-persuasiveness>.
- [35] The Dais, *Human or AI? Evaluating Labels on AI-Generated Social Media Content*, The Dais, Accessed: May 11, 2026, Mar. 2025. [Online]. Available: <https://dais.ca/reports/human-or-ai/>.

- [36] K. Paul, *Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says*, Reuters, Accessed: May 11, 2026, Jun. 2024. [Online]. Available: <https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/>.
- [37] R. Bommasani et al., “The Limits of AI Data Transparency Policy: Three Disclosure Fallacies,” *arXiv preprint arXiv:2601.18127*, 2026. [Online]. Available: <https://arxiv.org/abs/2601.18127>.
- [38] D. Shepardson, *Trump revokes Biden executive order on addressing AI risks*, Reuters, Accessed: May 11, 2026, Jan. 2025. [Online]. Available: <https://www.reuters.com/technology/artificial-intelligence/trump-revokes-biden-executive-order-addressing-ai-risks-2025-01-21/>.
- [39] H. Liu, Y. Chen, A. Narayanan, A. Balachandran, P. J. Moreno, and L. Wang, “Can deepfake speech be reliably detected?” *arXiv preprint arXiv:2410.06572*, 2024.
- [40] P. O’Reilly, Z. Jin, J. Su, and B. Pardo, “Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech,” in *ICLR Workshop on GenAI Watermarking*, 2025. [Online]. Available: <https://arxiv.org/abs/2504.10782>.
- [41] R. Hönig, J. Rando, N. Carlini, and F. Tramèr, “Adversarial perturbations cannot reliably protect artists from generative ai,” in *ICLR*, 2025.
- [42] P. O’Reilly, Z. Jin, J. Su, and B. Pardo, “Maskmark: Robust neural watermarking for real and synthetic speech,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.
- [43] P. O’Reilly, P. Seetharaman, J. Su, Z. Jin, and B. Pardo, “Code drift: Towards idempotent neural audio codecs,” in *ICASSP*, 2025.
- [44] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.50. [Online]. Available: <https://doi.org/10.1109/TPAMI.2013.50>.
- [45] L. Ruthotto and E. Haber, “An introduction to deep generative modeling,” *GAMM-Mitteilungen*, vol. 44, no. 2, e202100008, 2021. DOI: <https://doi.org/10.1002/gamm.202100008>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/gamm.202100008>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/gamm.202100008>.
- [46] R. Salakhutdinov, A. Mnih, and G. Hinton, “Restricted boltzmann machines for collaborative filtering,” in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML ’07, Corvallis, Oregon, USA: Association for Computing Machinery, 2007, pp. 791–798, ISBN: 9781595937933. DOI: 10.1145/1273496.1273596. [Online]. Available: <https://doi.org/10.1145/1273496.1273596>.

- [47] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, “An introduction to deep reinforcement learning,” *Found. Trends Mach. Learn.*, vol. 11, no. 3–4, pp. 219–354, Dec. 2018, ISSN: 1935-8237. DOI: 10.1561/22000000071. [Online]. Available: <https://doi.org/10.1561/22000000071>.
- [48] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.
- [49] M. Johnson et al., “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. DOI: 10.1162/tac1_a_00065. [Online]. Available: <https://aclanthology.org/Q17-1024/>.
- [50] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2014.
- [51] I. J. Goodfellow et al., “Generative adversarial nets,” in *Neural Information Processing Systems (NIPS)*, 2014.
- [52] D. Podell et al., “SDXL: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023. [Online]. Available: <https://arxiv.org/abs/2307.01952>.
- [53] A. Polyak et al., “Movie gen: A cast of media foundation models,” *arXiv preprint arXiv:2410.13720*, 2024. [Online]. Available: <https://arxiv.org/abs/2410.13720>.
- [54] OpenAI, *GPT-5 system card*, Updated August 13, 2025, Aug. 2025. [Online]. Available: <https://cdn.openai.com/gpt-5-system-card.pdf>.
- [55] T. Luong and V. Mirrokni, *Accelerating mathematical and scientific discovery with Gemini Deep Think*, <https://deepmind.google/blog/accelerating-mathematical-and-scientific-discovery-with-gemini-deep-think/>, Google DeepMind blog, Feb. 2026.
- [56] N. Carlini et al., *Assessing claude mythos preview’s cybersecurity capabilities*, <https://red.anthropic.com/2026/mythos-preview/>, Anthropic Red Team blog, Apr. 2026.
- [57] Y. Chen et al., “F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching,” *arXiv preprint arXiv:2410.06885*, 2024. [Online]. Available: <https://arxiv.org/abs/2410.06885>.
- [58] H. Yang, J. Su, M. Kim, and Z. Jin, “Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens,” in *Interspeech*, 2024.
- [59] F. Schmid, T. Morocutti, F. Foscari, J. Schlüter, P. Primus, and G. Widmer, “Effective pre-training of audio transformers for sound event detection,” in *ICASSP*, 2025.
- [60] IBM Corporation and Microsoft Corporation, *Multimedia programming interface and data specifications 1.0*, Defines the RIFF/WAVE file format for waveform audio data, IBM Corporation and Microsoft Corporation, Aug. 1991.

- [61] A. van den Oord et al., “WaveNet: A generative model for raw audio,” in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA: ISCA, 2016, p. 125. doi: 10.21437/SSW.2016-33. [Online]. Available: https://www.isca-archive.org/ssw_2016/vandenoord16_ssw.html.
- [62] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “Soundstream: An end-to-end neural audio codec,” *arXiv preprint arXiv:2107.03312*, 2021.
- [63] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *arXiv preprint arXiv:2210.13438*, 2022.
- [64] R. Kumar, P. Seetharaman, I. K. Alejandro Luebs, and K. Kumar, “High-fidelity audio compression with improved rvqgan,” in *NeurIPS*, 2023.
- [65] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: <https://www.deeplearningbook.org/>.
- [66] Y. LeCun et al., “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [67] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, “Deep learning for audio signal processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019. doi: 10.1109/JSTSP.2019.2908700.
- [68] A. Hannun et al., “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.
- [69] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012, pp. 1097–1105.
- [71] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
- [72] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 6306–6315.
- [73] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>.
- [74] S. Dieleman, C. Nash, J. Engel, and K. Simonyan, “Variable-rate discrete representation learning,” *arXiv preprint arXiv:2103.06089*, 2021.
- [75] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006. doi: 10.1126/science.1127647.

- [76] J. S. Chung et al., “In defence of metric learning for speaker recognition,” in *Interspeech*, 2020.
- [77] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Neural Information Processing Systems (NeurIPS)*, 2020.
- [78] H. Flores Garcia, P. Seetharaman, R. Kumar, and B. Pardo, “VampNet: Music generation via masked acoustic token modeling,” in *Conference of the International Society for Music Information Retrieval (ISMIR)*, 2023.
- [79] H. F. García, O. Nieto, J. Salamon, B. Pardo, and P. Seetharaman, “Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations,” in *ICASSP*, 2025.
- [80] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” *arXiv preprint arXiv:2210.02747*, 2023.
- [81] Y. Wang et al., “Scaling speech tokenizers with diffusion autoencoders,” in *ICLR*, 2026.
- [82] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, “Quickvc: Any-to-many voice conversion using inverse short-time fourier transform for faster conversion,” *arXiv preprint arXiv:2302.08296*, 2023.
- [83] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.
- [84] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, vol. 2, 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.
- [85] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [86] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, *Jukebox: A generative model for music*, 2020. arXiv: 2005.00341 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2005.00341>.
- [87] Z. Borsos et al., “AudioLM: A language modeling approach to audio generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023. DOI: 10.1109/TASLP.2023.3288409.
- [88] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.
- [89] K. Kumar et al., “Melgan: Generative adversarial networks for conditional waveform synthesis,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [90] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2794–2802.

- [91] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [92] C. Szegedy et al., “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014. [Online]. Available: <https://arxiv.org/abs/1312.6199>.
- [93] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6572>.
- [94] N. Carlini and H. Farid, “Evading deepfake-image detectors with white- and black-box attacks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 658–659. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Carlini_Evading_Deepfake-Image_Detectors_With_White-_and_Black-Box_Attacks_CVPRW_2020_paper.html.
- [95] N. Carlini et al., “Hidden voice commands,” in *Proceedings of the 25th USENIX Security Symposium*, USENIX Association, 2016, pp. 513–530. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini>.
- [96] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *2018 IEEE Security and Privacy Workshops*, IEEE, 2018, pp. 1–7. DOI: 10.1109/SPW.2018.00009. [Online]. Available: <https://arxiv.org/abs/1801.01944>.
- [97] K. Eykholt et al., “Robust physical-world attacks on deep learning visual classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634. DOI: 10.1109/CVPR.2018.00175.
- [98] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003. [Online]. Available: <https://www.jmlr.org/papers/v3/bengio03a.html>.
- [99] T. Li, Y. Tian, H. Li, M. Deng, and K. He, “Autoregressive image generation without vector quantization,” in *NeurIPS*, 2024.
- [100] M. Pasini, J. Nistal, S. Lattner, and G. Fazekas, “Continuous autoregressive models with noise augmentation avoid error accumulation,” in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.
- [101] MIDI Manufacturers Association (MMA) and Association of Music Electronics Industry (AMEI), *Midi 1.0 detailed specification*, Available at <https://www.midi.org/specs>, 1983.
- [102] C.-Z. A. Huang et al., “Music transformer: Generating music with long-term structure,” in *International Conference on Learning Representations (ICLR)*, 2019.

- [103] J. Thickstun, D. Hall, C. Donahue, and P. Liang, “Anticipatory music transformer,” *arXiv preprint arXiv:2306.08620*, 2023.
- [104] R. Langman, A. Jukić, K. Dhawan, N. R. Koluguri, and B. Ginsburg, *Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis*, 2024.
- [105] C. Wang et al., “Neural codec language models are zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.
- [106] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, “Soundstorm: Efficient parallel audio generation,” *arXiv preprint arXiv:2305.09636*, 2023.
- [107] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, “Maskgit: Masked generative image transformer,” in *CVPR*, 2022.
- [108] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.
- [109] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *International Conference on Learning Representations (ICLR)*, 2021.
- [110] T. Salimans and J. Ho, “Progressive distillation for fast sampling of diffusion models,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [111] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *International Conference on Learning Representations*, 2021.
- [112] S. Kumar et al., “Tac: Timestamped audio captioning,” *arXiv preprint arXiv:2602.15766*, 2026.
- [113] T. Labiausse, L. Mazaré, E. Grave, P. Pérez, A. Défossez, and N. Zeghidour, “High-fidelity simultaneous speech-to-speech translation,” *arXiv preprint arXiv:2502.03382*, 2025.
- [114] Seamless Communication et al., “Seamless: Multilingual expressive and streaming speech translation,” *arXiv preprint arXiv:2312.05187*, 2023.
- [115] P. U. Diehl et al., “Restoring speech intelligibility for hearing aid users with deep learning,” *Scientific Reports*, vol. 13, no. 2719, 2023. doi: 10.1038/s41598-023-29871-8. [Online]. Available: <https://www.nature.com/articles/s41598-023-29871-8>.
- [116] J. H. Christensen, H. Whiston, M. Lough, J. C. Gil-Carvajal, J. Rumley, and G. H. Saunders, “Evaluating real-world benefits of hearing aids with deep neural network-based noise reduction: An ecological momentary assessment study,” *American Journal of Audiology*, Feb. 2024. doi: 10.1044/2023_AJA-23-00149. [Online]. Available: https://doi.org/10.1044/2023_AJA-23-00149.
- [117] H. Flores García and S. Moore, “Token telephone,” in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Utrecht, The Netherlands, Sep. 2024, p. 4. [Online]. Available: https://nime.org/proc_installations/nime2024_installations_3/index.html.

- [118] A. M. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, no. 236, pp. 433–460, 1950. DOI: 10.1093/mind/LIX.236.433.
- [119] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, “A proposal for the Dartmouth summer research project on artificial intelligence,” *AI Magazine*, vol. 27, no. 4, pp. 12–14, 2006, Original proposal dated August 31, 1955. DOI: 10.1609/aimag.v27i4.1904.
- [120] A. M. Turing, “Intelligent machinery, a heretical theory,” *Philosophia Mathematica*, vol. 4, no. 3, pp. 256–260, 1996, Lecture given to the ’51 Society at Manchester, c. 1951. DOI: 10.1093/philmat/4.3.256.
- [121] N. Wiener, “Some moral and technical consequences of automation,” *Science*, vol. 131, no. 3410, pp. 1355–1358, 1960. DOI: 10.1126/science.131.3410.1355.
- [122] I. J. Good, “Speculations concerning the first ultraintelligent machine,” in *Advances in Computers*, F. L. Alt and M. Rubinoff, Eds., vol. 6, Elsevier, 1966, pp. 31–88.
- [123] J. Weizenbaum, “ELIZA: A computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966. DOI: 10.1145/365153.365168.
- [124] J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W. H. Freeman and Company, 1976.
- [125] Commission for Racial Equality, “Medical school admissions: Report of a formal investigation into st. george’s hospital medical school,” Commission for Racial Equality, London, Tech. Rep., 1988.
- [126] O. Schwartz, “Untold history of ai: Algorithmic bias was born in the 1980s,” *IEEE Spectrum*, Apr. 15, 2019. [Online]. Available: <https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias>.
- [127] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine bias*, ProPublica, May 2016. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [128] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, vol. 81, PMLR, 2018, pp. 77–91. [Online]. Available: <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- [129] R. Chesney and D. K. Citron, “Deep fakes: A looming challenge for privacy, democracy, and national security,” *California Law Review*, vol. 107, pp. 1753–1819, 2019. DOI: 10.2139/ssrn.3213954.
- [130] S. Feldstein, “The global expansion of AI surveillance,” Carnegie Endowment for International Peace, Tech. Rep., Sep. 2019. [Online]. Available: <https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance>.

- [131] D. Loucaides, "The changing face of protest," *Rest of World*, Mar. 27, 2024. Accessed: May 21, 2026. [Online]. Available: <https://restofworld.org/2024/facial-recognition-government-protest-surveillance/>.
- [132] P. Mozur, "In hong kong protests, faces become weapons," *The New York Times*, Jul. 26, 2019. Accessed: May 21, 2026. [Online]. Available: <https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html>.
- [133] M. Borak, "London police deploy facial recognition during palestine and israel protests," *Biometric Update*, Jan. 15, 2024. Accessed: May 21, 2026. [Online]. Available: <https://www.biometricupdate.com/202401/london-police-deploy-facial-recognition-during-palestine-and-israel-protests>.
- [134] Reuters, *IBM to pause hiring in plan to replace 7,800 jobs with AI*, Reuters, May 2023. [Online]. Available: <https://www.reuters.com/technology/ibm-pause-hiring-plans-replace-7800-jobs-with-ai-bloomberg-news-2023-05-01/>.
- [135] Reuters, *Sweden's Klarna says AI chatbots help shrink headcount*, Reuters, Aug. 2024. [Online]. Available: <https://www.reuters.com/technology/artificial-intelligence/swedens-klarna-says-ai-chatbots-help-shrink-headcount-2024-08-27/>.
- [136] D. Bass, *Duolingo cuts 10% of contractors as it uses more AI to create app content*, Bloomberg, Jan. 2024. [Online]. Available: <https://www.bloomberg.com/news/articles/2024-01-08/duolingo-cuts-10-of-contractors-in-move-to-greater-use-of-ai>.
- [137] M. Heikkilä, *This artist is dominating AI-generated art. and he's not happy about it*, MIT Technology Review, Sep. 2022. [Online]. Available: <https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/>.
- [138] K. Nguyen, *AI is causing student artists to rethink their creative career plans*, KQED Arts, Apr. 2023. [Online]. Available: <https://www.kqed.org/arts/13928253/ai-art-artificial-intelligence-student-artists-midjourney>.
- [139] R. Salkowitz, *AI is coming for commercial art jobs. can it be stopped?* Forbes, Sep. 2022. [Online]. Available: <https://www.forbes.com/sites/robsalkowitz/2022/09/16/ai-is-coming-for-commercial-art-jobs-can-it-be-stopped/>.
- [140] Society of Authors, *SoA survey reveals a third of translators and quarter of illustrators losing work to AI*, Society of Authors, Apr. 2024. [Online]. Available: <https://societyofauthors.org/2024/04/11/soa-survey-reveals-a-third-of-translators-and-quarter-of-illustrators-losing-work-to-ai/>.

- [141] CVL Economics, “Future unscripted: The impact of generative artificial intelligence on entertainment industry jobs,” The Animation Guild, Concept Art Association, Human Artistry Campaign, and National Cartoonist Society Foundation, Tech. Rep., Jan. 2024. [Online]. Available: <https://animationguild.org/wp-content/uploads/2024/01/Future-Unscripted-The-Impact-of-Generative-Artificial-Intelligence-on-Entertainment-Industry-Jobs-pages-1.pdf>.
- [142] E. Tan, K. Huang, and M. Isaac, *Meta lays off 8,000 employees, as A.I. casualties mount*, The New York Times, May 2026. [Online]. Available: <https://www.nytimes.com/2026/05/19/technology/meta-layoffs-ai.html>.
- [143] Illinois General Assembly, *Biometric information privacy act*, 740 ILCS 14; effective October 3, 2008, 2008. [Online]. Available: <https://ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57>.
- [144] United States District Court for the Northern District of California, *In re Facebook biometric information privacy litigation*, Order granting final approval of class action settlement, No. 3:15-cv-03747-JD, Feb. 2021. [Online]. Available: <https://law.justia.com/cases/federal/district-courts/california/candce/3%3A2015cv03747/290385/474/>.
- [145] ACLU of Illinois, *In big win, settlement ensures Clearview AI complies with groundbreaking Illinois biometric privacy law*, ACLU of Illinois, May 2022. [Online]. Available: <https://www.aclu-il.org/big-win-settlement-ensures-clearview-ai-complies-groundbreaking-illinois-biometric-privacy-law/>.
- [146] European Parliament and Council of the European Union, *Regulation (EU) 2016/679, article 22: Automated individual decision-making, including profiling*, General Data Protection Regulation, Apr. 2016. [Online]. Available: <https://gdpr-info.eu/art-22-gdpr/>.
- [147] K. Conger, R. Fausset, and S. F. Kovalski, *San francisco bans facial recognition technology*, The New York Times, May 2019. [Online]. Available: <https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html>.
- [148] Treasury Board of Canada Secretariat, *Directive on automated decision-making*, Government of Canada directive; came into effect April 1, 2019, 2019. [Online]. Available: <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.
- [149] National Institute of Standards and Technology, “Artificial intelligence risk management framework (AI RMF 1.0),” National Institute of Standards and Technology, Tech. Rep. NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1. [Online]. Available: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.
- [150] J. Ding, E. Jäger, V. Ivanova, and M. Bunz, “My voice, your voice, our voice: Attitudes towards collective governance of a choral AI dataset,” in *Workshop on Creativity and Generative AI, Advances in Neural Information Processing Systems*, Presented at the NeurIPS 2024 Workshop on Creativity & Generative AI, 2024. doi: 10.48550/arXiv.2412.01433. [Online]. Available: <https://arxiv.org/abs/2412.01433>.

- [151] S. Cole, *Reddit just shut down the deepfakes subreddit*, Vice Motherboard, Feb. 2018. [Online]. Available: <https://www.vice.com/en/article/reddit-shuts-down-deepfakes/>.
- [152] Vanderbilt Law School, *Taylor swift & AI deepfakes: Swifties join the fight against deepfake pornography*, Vanderbilt Law School, Feb. 2024. [Online]. Available: <https://law.vanderbilt.edu/taylor-swift-ai-deepfakes-swifties-join-the-fight-against-deepfake-pornography/>.
- [153] MultiState, *State deepfake laws in 2026: What's changed and what's next*, MultiState Insider, Feb. 2026. [Online]. Available: <https://www.multistate.us/insider/2026/2/12/how-ai-generated-content-laws-are-changing-across-the-country>.
- [154] J. R. Biden, *Executive order 14110: Safe, secure, and trustworthy development and use of artificial intelligence*, Federal Register, Published November 1, 2023, Oct. 2023. [Online]. Available: <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.
- [155] D. Castelvetti, *What you need to know about Biden's sweeping AI order*, IEEE Spectrum, Oct. 2023. [Online]. Available: <https://spectrum.ieee.org/biden-ai-executive-order>.
- [156] National Conference of State Legislatures, *Summary of artificial intelligence 2025 legislation*, National Conference of State Legislatures, 2025. [Online]. Available: <https://www.ncsl.org/technology-and-communication/artificial-intelligence-2025-legislation>.
- [157] D. J. Trump, *Ensuring a national policy framework for artificial intelligence*, The White House, Dec. 2025. [Online]. Available: <https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/>.
- [158] Coalition for Content Provenance and Authenticity, *C2PA: Coalition for content provenance and authenticity*, Coalition for Content Provenance and Authenticity, 2024. [Online]. Available: <https://c2pa.org/>.
- [159] D. Castro and M. McLaughlin, *Why AI-generated content labeling mandates fall short*, Center for Data Innovation, Dec. 2024. [Online]. Available: <https://datainnovation.org/2024/12/why-ai-generated-content-labeling-mandates-fall-short/>.
- [160] Tennessee General Assembly, *Ensuring likeness voice and image security act of 2024*, ELVIS Act, Tenn. Pub. Ch. 588, 2024. [Online]. Available: <https://publications.tn.gov/files/acts/113/pub/pc0588.pdf>.
- [161] CISAC and PMP Strategy, "Study on the economic impact of generative AI in the music and audiovisual sectors," International Confederation of Societies of Authors and Composers, Tech. Rep., 2024. [Online]. Available: <https://www.cisac.org/services/reports-and-research/cisacpmp-strategy-ai-study>.

- [162] European Parliament, “Generative AI and copyright: Training, creation, regulation,” European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs, Tech. Rep. PE 774.095, Jul. 2025. DOI: 10.2861/9120512. [Online]. Available: [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU\(2025\)774095_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU(2025)774095_EN.pdf).
- [163] F. Pasquale and H. Sun, “Consent and compensation: Resolving generative AI’s copyright crisis,” *Virginia Law Review Online*, vol. 110, pp. 207–247, 2024. [Online]. Available: <https://virginialawreview.org/articles/consent-and-compensation-resolving-generative-ais-copyright-crisis/>.
- [164] Universal Music Group, *Universal music group and Udio announce Udio’s first strategic agreements for new licensed AI music creation platform*, Universal Music Group, Oct. 2025. [Online]. Available: <https://www.universalmusic.com/universal-music-group-and-udio-announce-udios-first-strategic-agreements-for-new-licensed-ai-music-creation-platform/>.
- [165] Spotify, *Universal music group and Spotify announce landmark licensing agreements for fan-made covers and remixes*, Spotify Newsroom, May 2026. [Online]. Available: <https://newsroom.spotify.com/2026-05-21/universal-music-group-spotify-licensing-agreements-fan-made-covers-remixes/>.
- [166] F. Morreale, W. Hutiri, J. Serrà, A. Xiang, and Y. Mitsufuji, “Attribution-by-design: Ensuring inference-time provenance in generative music systems,” *arXiv preprint arXiv:2510.08062*, 2025. [Online]. Available: <https://arxiv.org/abs/2510.08062>.
- [167] W. Kim et al., “From generation to attribution: Music ai agent architectures for the post-streaming era,” in *NeurIPS 2025 AI4Music Workshop*, 2025. [Online]. Available: <https://arxiv.org/abs/2510.20276>.
- [168] S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry, “TRAK: Attributing model behavior at scale,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 27 074–27 113.
- [169] P. W. Koh and P. Liang, “Understanding black-box predictions via influence functions,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1885–1894. [Online]. Available: <https://proceedings.mlr.press/v70/koh17a.html>.
- [170] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” *arXiv preprint arXiv:1606.06565*, 2016.
- [171] V. Krakovna et al., *Specification gaming: The flip side of AI ingenuity*, DeepMind Blog, 2020. [Online]. Available: <https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/>.

- [172] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 35, Curran Associates, Inc., 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- [173] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6572>.
- [174] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=rJzIBfZAb>.
- [175] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography Conference*, Springer, 2006, pp. 265–284. doi: 10.1007/11681878_14.
- [176] M. Abadi et al., “Deep learning with differential privacy,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Association for Computing Machinery, 2016, pp. 308–318. doi: 10.1145/2976749.2978318.
- [177] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016.
- [178] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 28, PMLR, 2013, pp. 325–333. [Online]. Available: <https://proceedings.mlr.press/v28/zemel13.html>.
- [179] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 35, Curran Associates, Inc., 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- [180] Y. Bai et al., *Constitutional AI: Harmlessness from AI feedback*, 2022. arXiv: 2212.08073 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2212.08073>.
- [181] P. O’Reilly, P. Awasthi, A. Vijayaraghavan, and B. Pardo, “Effective and inconspicuous over-the-air adversarial examples with adaptive filtering,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [182] C. Szegedy et al., “Intriguing properties of neural networks,” in *International Conference on Learning Representations (ICLR)*, 2014.

- [183] W. Zhang et al., “Attack on practical speaker verification system using universal adversarial perturbations,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [184] A. S. Shamsabadi, F. S. Teixeira, A. Abad, B. Raj, A. Cavallaro, and I. Trancoso, “Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [185] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *IEEE Security and Privacy Workshops*, 2018.
- [186] A. Défossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in *Interspeech*, 2020.
- [187] D. Ronssin and M. Cernak, “Ac-vc: Non-parallel low latency phonetic posteriorgrams based voice conversion,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.
- [188] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, “Ddsp: Differentiable digital signal processing,” in *International Conference on Learning Representations (ICLR)*, 2020.
- [189] B. Gellman and A. Soltani, “Nsa surveillance program reaches ‘into the past’ to retrieve, replay phone calls,” *The Washington Post*, 2014.
- [190] H. R. Watch, *Secret evidence and the threat of more warrantless surveillance*, <https://www.hrw.org/news/2018/01/11/secret-evidence-and-threat-more-warrantless-surveillance>, 2018.
- [191] H. R. Watch, *Q & a: Us warrantless surveillance under section 702 of the foreign intelligence surveillance act*, <https://www.hrw.org/news/2017/09/14/q-us-warrantless-surveillance-under-section-702-foreign-intelligence-surveillance>, 2017.
- [192] U. Iqbal et al., “Your echos are heard: Tracking, profiling, and ad targeting in the amazon smart speaker ecosystem,” *arXiv preprint arXiv:2204.10920*, 2022.
- [193] E. F. Foundation, *Upstream vs. prism*, <https://www.eff.org/702-spying>, 2018.
- [194] J. Stanley. “On the creation of giant voiceprint databases,” American Civil Liberties Union, Accessed: May 27, 2026. [Online]. Available: <https://www.aclu.org/news/privacy-technology/creation-giant-voiceprint-databases>.
- [195] D. Nathan and G. Joseph. “Prisons across the u.s. are quietly building databases of incarcerated people’s voice prints,” *The Intercept*, Accessed: May 27, 2026. [Online]. Available: <https://theintercept.com/2019/01/30/prison-voice-prints-databases-securus/>.
- [196] B. Lipton and C. Quintin. “The catalog of carceral surveillance: Voice recognition and surveillance,” Electronic Frontier Foundation, Accessed: May 27, 2026. [Online]. Available: <https://www.eff.org/deeplinks/2021/09/catalog-carceral-surveillance-voice-recognition-and-surveillance>.

- [197] D. Kayyali, *The history of surveillance and the black community*, <https://www.eff.org/deeplinks/2014/02/history-surveillance-and-black-community>, 2014.
- [198] H. Giorgis, “When the fbi spied on mlk,” *The Atlantic*, 2021.
- [199] P. J. Barger and S. Sridharan, “On the performance and use of speaker recognition systems for surveillance,” in *International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2006.
- [200] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, “A review of speaker diarization: Recent advances with deep learning,” *Computer Speech & Language*, vol. 72, Nov. 2021.
- [201] L. Fan, Q.-Y. Jiang, Y.-Q. Yu, and W.-J. Li, “Deep hashing for speaker identification and retrieval,” in *Interspeech*, 2019.
- [202] L. Schmidt, M. Sharifi, and I. Lopez-Moreno, “Large-scale speaker identification,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [203] L. Li, C. Xing, D. Wang, K. Yu, and T. F. Zheng, “Binary speaker embedding,” in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016.
- [204] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, “Voice convergin: Speaker de-identification by voice transformation,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2009.
- [205] F. Alegre, G. Soldi, N. Evans, B. Fauve, and J. Liu, “Evasion and obfuscation in speaker recognition surveillance and forensics,” in *International Workshop on Biometrics and Forensics*, 2014.
- [206] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, “Real-time, full-band, online dnn-based voice conversion system using a single cpu,” in *Interspeech*, 2020.
- [207] Y. A. Li, A. Zare, and N. Mesgarani, “Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion,” in *Interspeech*, 2021.
- [208] N. Tomashenko et al., “The VoicePrivacy 2020 challenge: Results and findings,” *Computer Speech & Language*, vol. 74, p. 101 362, 2022.
- [209] M. Chiquier, C. Mao, and C. Vondrick, “Real-time neural voice camouflage,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [210] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, “Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations,” in *CCS*, 2020.
- [211] A. Kassis and U. Hengartner, “Practical attacks on voice spoofing countermeasures,” *arXiv preprint arXiv:2107.14642*, 2021.
- [212] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, “Enabling fast and universal audio adversarial attack using generative model,” in *AAAI*, 2021.
- [213] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” in *International Society for Music Information Retrieval (ISMIR)*, 2018.

- [214] H. Abdullah et al., “Hear “no evil”, see “kenansville”*: Efficient and transferable black-box attacks on speech recognition and voice identification systems,” in *IEEE Symposium on Security and Privacy*, 2021.
- [215] S. Ahmed et al., “Pipe overflow: Smashing voice authentication for fun and profit,” *arXiv preprint arXiv:2107.14642*, 2022.
- [216] Verizon, *Monthly ip latency statistics*, <https://www.verizon.com/business/terms/latency/>, 2022.
- [217] M. Kassim, R. A. Rahman, M. A. A. Aziz, A. Idris, and M. I. Yusof, “Performance analysis of voip over 3g and 4g lte network,” in *2017 International Conference on Electrical, Electronics and System Engineering (ICEESE)*, IEEE, 2017, pp. 37–41.
- [218] M. Morise, H. Kawahara, and H. Katayose, “Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,” *Journal of the Audio Engineering Society*, Feb. 2009.
- [219] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016. doi: 10.1587/transinf.2015EDP7457.
- [220] J. Hsu, *Pyworld*, <https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder>, 2021.
- [221] R. McCurdy, “Tentative standards for sound level meters,” *Electrical Engineering*, vol. 55, no. 3, pp. 260–263, 1936.
- [222] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735.
- [223] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.
- [224] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech model pre-training for end-to-end spoken language understanding,” in *Interspeech*, 2019.
- [225] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi,” in *Interspeech*, 2017.
- [226] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
- [227] T. J. Hazen, W. Shen, and C. White, “Query-by-example spoken term detection using phonetic posteriorgram templates,” in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009.
- [228] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *AAAI*, 2018.
- [229] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the voxceleb speaker recognition challenge 2020,” *arXiv preprint arXiv:2009.14153*, 2020.

- [230] C. Wang, D. Yogatama, A. Coates, T. Han, A. Y. Hannun, and B. Xiao, “Lookahead convolution layer for unidirectional recurrent neural networks,” in *ICLR Workshop*, 2016.
- [231] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, “Upsampling artifacts in neural audio synthesis,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [232] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Neural Information Processing Systems (NeurIPS)*, 2019.
- [233] J. S. Chung et al., “In defence of metric learning for speaker recognition,” in *Interspeech*, 2020.
- [234] G. Zhu, F. Jiang, and Z. Duan, “Y-vector: Multiscale waveform encoder for speaker embedding,” in *Interspeech*, 2021.
- [235] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *Interspeech*, 2018.
- [236] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *Interspeech*, 2017.
- [237] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *ICASSP*, 2001.
- [238] M. Wang, C. Boeddeker, R. G. Dantas, and A. Seelan, *Pesq*, <https://github.com/ludlows/python-pesq>, 2022.
- [239] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011. doi: 10.1109/TASL.2011.2114881.
- [240] M. Pariente, *Pystoi*, <https://github.com/mpariente/pystoi>, 2021.
- [241] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Neural Information Processing Systems (NeurIPS)*, 2020.
- [242] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, “Fast and easy crowdsourced perceptual audio evaluation,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
- [243] M. Morrison, B. Tang, G. Tan, and B. Pardo, “Reproducible subjective evaluation,” in *ICLR Workshop on ML Evaluation Standards*, Apr. 2022.
- [244] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: From phenomena to black-box attacks using adversarial samples,” *arXiv preprint arXiv:1605.07277*, 2016.
- [245] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “The space of transferable adversarial examples,” *arXiv preprint arXiv:1704.03453*, 2017.

- [246] Y. Dong et al., “Boosting adversarial attacks with momentum,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.
- [247] C. Xie et al., “Improving transferability of adversarial examples with input diversity,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2730–2739.
- [248] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, “V-Cloak: Intelligibility-, naturalness- & timbre-preserving real-time voice anonymization,” in *32nd USENIX Security Symposium (USENIX Security 23)*, USENIX Association, 2023. arXiv: 2210.15140 [cs.CR]. [Online]. Available: <https://arxiv.org/abs/2210.15140>.
- [249] J. Barnett et al., “Audio data defenses: Protecting music and speech data from targeted attacks,” in *Extended Abstracts for the Late-Breaking Demo Session of the 25th Int. Society for Music Information Retrieval Conf.*, 2024.
- [250] K. Jain et al., “Compact neural tts voices for accessibility,” in *ICASSP*, 2025.
- [251] S. Nercessian, “Zero-shot singing voice conversion,” in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, ser. ISMIR, 2020, pp. 70–76.
- [252] S. Dowell, “Polish startup wows with ‘deep fake’ voice cloning tool which can make leonardo dicaprio sound like kim kardashian,” *TheFirstNews*, 2023.
- [253] S. Bond, “People are trying to claim real videos are deepfakes. the courts are not amused,” *National Public Radio*, 2023.
- [254] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *ICML*, 2021.
- [255] Coalition for Content Provenance and Authenticity, *C2PA technical specification*, Technical specification, version 2.4, Accessed 2026-06-08, 2026. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html.
- [256] International Press Telecommunications Council, *Digital source type newscodes vocabulary*, Accessed 2026-06-08, 2026. [Online]. Available: <https://cv.iprtc.org/newscodes/digitalsourcetype/>.
- [257] Meta, *Labeling AI-generated images on Facebook, Instagram and Threads*, Accessed 2026-06-08, Feb. 2024. [Online]. Available: <https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/>.
- [258] TikTok, *Partnering with our industry to advance AI transparency and literacy*, Accessed 2026-06-08, May 2024. [Online]. Available: <https://newsroom.tiktok.com/en-us/partnering-with-our-industry-to-advance-ai-transparency-and-literacy>.
- [259] Coalition for Content Provenance and Authenticity, *C2PA implementation guidance*, Technical specification guidance, version 2.4, Accessed 2026-06-08, 2026. [Online]. Available: <https://spec.c2pa.org/specifications/specifications/2.4/guidance/Guidance.html>.

- [260] Content Authenticity Initiative, *Durable content credentials*, Accessed 2026-06-08, Apr. 2024. [Online]. Available: <https://contentauthenticity.org/blog/durable-content-credentials>.
- [261] OpenAI, *C2PA and SynthID in OpenAI-generated images*, Help center article, Accessed 2026-06-08, 2026. [Online]. Available: <https://help.openai.com/en/articles/8912793-c2pa-and-synthid-in-openai-generated-images>.
- [262] P. England et al., “AMP: Authentication of media via provenance,” in *Proceedings of the 12th ACM Multimedia Systems Conference*, Association for Computing Machinery, 2021, pp. 108–121. DOI: 10.1145/3458305.3459599. [Online]. Available: <https://doi.org/10.1145/3458305.3459599>.
- [263] X. Wang et al., “Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale,” *arXiv preprint arXiv:2408.08739*, 2024.
- [264] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, “The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023. DOI: 10.1109/taslp.2022.3233236.
- [265] J. Frank and L. Schönherr, “WaveFake: A Data Set to Facilitate Audio Deepfake Detection,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- [266] X. Li, P.-Y. Chen, and W. Wei, “Sonar: A synthetic ai-audio detection framework and benchmark,” *arXiv preprint arXiv:2410.04324*, 2024.
- [267] H. Ali, S. Subramani, S. Sudhir, R. Varahamurthy, and H. Malik, “Is audio spoof detection robust to laundering attacks?” In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&MMSec ’24, Baiona, Spain: Association for Computing Machinery, 2024, pp. 283–288, ISBN: 9798400706370. DOI: 10.1145/3658664.3659656. [Online]. Available: <https://doi.org/10.1145/3658664.3659656>.
- [268] Z. Ju et al., “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” *arXiv preprint arXiv:2403.03100*, 2024.
- [269] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “Bigvgan: A universal neural vocoder with large-scale training,” in *ICLR*, 2023.
- [270] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, “Chunked autoregressive gan for conditional waveform synthesis,” in *ICLR*, 2022.
- [271] D. Afchar, G. Meseguer-Brocal, and R. Hennequin, “Detecting music deepfakes is easy but actually hard,” *arXiv preprint arXiv:2405.04181*, 2024.
- [272] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense,” in *Neural Information Processing Systems (NeurIPS)*, 2023.

- [273] A. Défossez et al., “Moshi: A speech-text foundation model for real-time dialogue,” *arXiv preprint arXiv:2410.00037*, 2024.
- [274] Y. Cui, L. He, and S. Zhao, “Introducing the watermark algorithm for synthetic voice identification,” *Azure AI services Blog*, 2022. [Online]. Available: <https://techcommunity.microsoft.com/t5/azure-ai-services-blog/introducing-the-watermark-algorithm-for-synthetic-voice/ba-p/3298548>.
- [275] Y. Tai and M. F. Mansour, “Audio watermarking over the air with modulated self-correlation,” in *ICASSP*, 2019.
- [276] C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang, and N. Yu, “Dear: A deep-learning-based audio re-recording resilient watermarking,” in *AAAI*, 2023.
- [277] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Interspeech*, 2020.
- [278] D. Kirovski and H. Malvar, “Spread-spectrum watermarking of audio signals,” *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1020–1033, 2003. doi: 10.1109/TSP.2003.809384.
- [279] J. Yamagishi, *English multi-speaker corpus for cstr voice cloning toolkit*, 2012. [Online]. Available: <https://datashare.ed.ac.uk/handle/10283/2950>.
- [280] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.
- [281] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *ICASSP*, 2017, pp. 5220–5224.
- [282] G. Wichern et al., “Wham!: Extending speech separation to noisy environments,” in *Interspeech*, 2019.
- [283] J. Traer and J. H. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, E7856–E7865, 2016.
- [284] A. Défossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in *Interspeech*, 2020.
- [285] P. Seetharaman, G. Wichern, B. Pardo, and J. Le Roux, “Autoclip: Adaptive gradient clipping for source separation networks,” in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2020.
- [286] T. Dozat, “Incorporating nesterov momentum into adam,” *ICLR Workshop*, pp. 2013–2016, 2016.
- [287] G. J. Mysore, “Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? - a dataset, insights, and challenges,” *IEEE Signal Processing Letters*, vol. 22, no. 8, 2015.

- [288] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, “Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust,” in *Neural Information Processing Systems (NeurIPS)*, 2023.
- [289] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, “Style transfer of audio effects with differentiable signal processing,” *Journal of the Audio Engineering Society*, 2022.
- [290] H. Siuzdak, “Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis,” *arXiv preprint arXiv:2306.00814*, 2023.
- [291] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” in *ICLR*, 2021.
- [292] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *NeurIPS*, 2020.
- [293] S. Fu et al., “Metricgan+: An improved version of metricgan for speech enhancement,” in *Interspeech*, 2021.
- [294] R. San Roman, P. Fernandez, H. Elsahar, A. D’efosse, T. Furon, and T. Tran, “Proactive detection of voice cloning with localized watermarking,” *International Conference on Machine Learning (ICML)*, 2024.
- [295] Y. Liu, L. Lu, J. Jin, L. Sun, and A. Fanelli, “Xattnmark: Learning robust audio watermarking with cross-attention,” *arXiv preprint arXiv:2502.04230*, 2025.
- [296] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, “Wavmark: Watermarking for audio generation,” *arXiv preprint arXiv:2308.12770*, 2023.
- [297] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, “Detecting voice cloning attacks via timbre watermarking,” in *Network and Distributed System Security Symposium*, 2024. doi: 10.14722/ndss.2024.24200.
- [298] M. K. Singh, N. Takahashi, W. Liao, and Y. Mitsufuji, “SilentCipher: Deep Audio Watermarking,” in *Proc. InterSpeech 2024*, 2024.
- [299] W. Liu, Y. Li, D. Lin, H. Tian, and H. Li, “Groot: Generating robust watermark for diffusion-model-based audio synthesis,” in *ACM MM*, 2024.
- [300] P. Li, X. Zhang, J. Xiao, and J. Wang, “Ideaw: Robust neural audio watermarking with invertible dual-embedding,” *arXiv preprint arXiv:2409.19627*, 2024.
- [301] L. Li, L. Jin, Y. Wang, H. Sun, Z. Hu, and C. Maple, “Aura: A stegaformer-based scalable deep audio watermark with extreme robustness,” in *ICASSP*, 2026.
- [302] L. Juvela and X. Wang, “Audio codec augmentation for robust collaborative watermarking of speech synthesis,” *arXiv preprint arXiv:2409.13382*, 2024.
- [303] L. Juvela and X. Wang, “Collaborative watermarking for adversarial speech synthesis,” in *Proc. ICASSP*, arXiv preprint arXiv:2309.15224, 2024.
- [304] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, “Twenty years of digital audio watermarking—a comprehensive review,” *Signal processing*, vol. 128, pp. 222–242, 2016.

- [305] Y. Cho, C. Kim, Y. Yang, and Y. Ren, “Attributable-watermarking of speech generative models,” in *Proc. ICASSP*, 2022.
- [306] J. Zhou et al., “Traceablespeech: Towards proactively traceable text-to-speech with watermarking,” in *Proc. InterSpeech*, 2024.
- [307] J. Zhou, J. Yi, Y. Ren, J. Tao, T. Wang, and C. Y. Zhang, “Wmcodec: End-to-end neural speech codec with deep watermarking for authenticity verification,” *arXiv preprint arXiv:2409.12121*, 2024.
- [308] H. Wang et al., “Ssr-speech: Towards stable, safe and robust zero-shot text-based speech editing and synthesis,” *arXiv preprint arXiv:2409.07556*, 2024.
- [309] R. S. Roman, P. Fernandez, A. Deleforge, Y. Adi, and R. Serizel, “Latent watermarking of audio generative models,” *arXiv preprint arXiv:2409.02915*, 2024.
- [310] M. K. Singh, N. Takahashi, W.-H. Liao, and Y. Mitsufuji, “LOCKEY: A novel approach to model authentication and deepfake tracking,” in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.
- [311] S. Ji et al., “Speech watermarking with discrete intermediate representations,” in *AAAI*, 2025.
- [312] H. Liu, M. Guo, Z. Jiang, L. Wang, and N. Z. Gong, “Audiomarkbench: Benchmarking robustness of audio watermarking,” in *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [313] F. Research, *Omnisealbench*, <https://github.com/facebookresearch/omnisealbench>, 2024.
- [314] *Mp3 (mpeg layer iii audio encoding)*, 1993. [Online]. Available: <https://www.iso.org/standard/22412.html>.
- [315] J.-M. Valin, K. Vos, and T. B. Terriberry, “Definition of the opus audio codec,” *IETF RFC 6716*, 2012. [Online]. Available: <https://tools.ietf.org/html/rfc6716>.
- [316] S. E. Eskimez et al., “E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2024.
- [317] Y. Hu et al., “Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement,” in *Proc. InterSpeech 202*, 2020.
- [318] A. López-López, E. Rosello, and A. M. Gomez, “Speech watermarking removal by DNN-based speech enhancement attacks,” in *Proceedings of IberSPEECH*, 2024.
- [319] Y. Gu and E. Diao, *Esc: Efficient speech coding with cross-scale residual vector quantized transformers*, 2024.
- [320] Z. Ju et al., “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” *arXiv preprint arXiv:2403.03100*, 2024.
- [321] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Spechtokenizer: Unified speech tokenizer for speech large language models,” in *ICLR*, 2024.

- [322] Y. Ren et al., “Fewer-token neural speech codec with time-invariant codes,” in *ICASSP*, 2024.
- [323] P. Peng, P.-Y. Huang, D. Li, A. Mohamed, and D. Harwath, “Voicecraft: Zero-shot speech editing and text-to-speech in the wild,” *arXiv preprint arXiv:2403.16973v1*, 2024.
- [324] S.-g. Lee and R. Valle. “Achieving state-of-the-art zero-shot waveform audio generation across audio types.” NVIDIA Developer Blog. [Online]. Available: <https://developer.nvidia.com/blog/achieving-state-of-the-art-zero-shot-waveform-audio-generation-across-audio-types/>.
- [325] P. Fernandez, A. Level, and T. Furon, “What lies ahead for generative ai watermarking,” in *Proceedings of the ICML 2024 Workshop on Generative AI and Law*, 2024.
- [326] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. DOI: 10.1109/TASLP.2021.3122291.
- [327] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the voxceleb speaker recognition challenge 2020,” *arXiv preprint arXiv:2009.14153*, 2020.
- [328] A. Kumar et al., “Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio,” in *ICASSP*, 2023.
- [329] S. Ji et al., “Wavtokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling,” *arXiv preprint arXiv:2408.16532*, 2024.
- [330] J. D. Parker et al., “Scaling transformers for low-bitrate high-quality speech coding,” *arXiv preprint arXiv:2411.19842*, 2024.
- [331] E. Radiya-Dixit, S. Hong, N. Carlini, and F. Tramèr, “Data poisoning won’t save you from facial recognition,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=B5XahNLmna>.
- [332] R. Hönig, J. Rando, N. Carlini, and F. Tramèr, “Adversarial perturbations cannot reliably protect artists from generative ai,” in *International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=yfW1x7uBS5>.
- [333] Y. Liu et al., “Image watermarks are removable using controllable regeneration from clean noise,” *arXiv preprint arXiv:22410.05470*, 2024.
- [334] R. Olivier, H. Abdullah, and B. Raj, “Watch what you pretrain for: Targeted, transferable adversarial examples on self-supervised speech recognition models,” in *Proceedings of the 2nd AdvML Frontiers Workshop at the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, Honolulu, Hawaii, USA: PMLR, 2023.
- [335] Z. Liu and M. Černak, “StreamMark: A deep learning-based semi-fragile audio watermarking for proactive deepfake detection,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Accepted, 2026.

- [336] L. Cuccovillo, X. Wang, M. Gerhardt, and P. Aichroth, “Neural encoding detection is not all you need for synthetic speech detection,” in *Proceedings of the IEEE International Workshop on Biometrics and Forensics (IWBf)*, 2026, pp. 1–6.
- [337] A. Pujari and A. Rattani, “AudioAuth: A dual-watermarking framework for robust audio integrity and source attribution,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2026. doi: 10.1109/TBIOM.2026.3679274.
- [338] H. Li et al., “Feature-aligned speech watermarking for robustness to reconstruction distortions,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Accepted, 2026. arXiv: 2606.11828 [cs.SD].
- [339] W. Ge, X. Wang, and J. Yamagishi, “FakeMark: Deepfake speech attribution with watermarked artifacts,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2026, pp. 10 687–10 697.
- [340] L. Yao et al., “Yours or mine? overwriting attacks against neural audio watermarking,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, 2026, pp. 27 756–27 763. doi: 10.1609/aaai.v40i33.39997.
- [341] Y. Özer, W. Ge, Z. Zhang, X. Wang, and J. Yamagishi, “Self voice conversion as an attack against neural audio watermarking,” in *IEICE Technical Report*, SP2025-53, vol. 125, 2026, pp. 165–170.
- [342] K. Li, X. Hu, I. Grishchenko, and D. Lie, *HarmonicAttack: An adaptive cross-domain audio watermark removal*, Under review, 2025. arXiv: 2511.21577 [cs.SD].
- [343] A. Liu, L. Pan, X. Hu, S. Meng, and L. Wen, “A semantic invariant robust watermark for large language models,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [344] A. B. Hou et al., “Semstamp: A semantic watermark with paraphrastic robustness for text generation,” in *NAACL*, 2024.
- [345] E. Giboulot and T. Furon, “Watermax: Breaking the llm watermark detectability-robustness-quality trade-off,” in *Neural Information Processing Systems (NeurIPS)*, 2024.
- [346] Y. Wu, R. Chen, G. Milis, J. Guo, and H. Huang, “A watermark for auto-regressive speech generation models,” in *Proceedings of Interspeech 2025*, 2025, pp. 3474–3478. doi: 10.21437/Interspeech.2025-2328.
- [347] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” in *ICLR*, 2019.
- [348] J. Li, B. Li, and Y. Lu, “Deep contextual video compression,” in *Neural Information Processing Systems*, 2021.
- [349] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, “Conditional probability models for deep image compression,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [350] F. Kreuk et al., “Audiogen: Textually guided audio generation,” in *ICLR*, 2023.

- [351] P. Mousavi, L. D. Libera, J. Duret, A. Ploujnikov, C. Subakan, and M. Ravanelli, “Dasb - discrete audio and speech benchmark,” *arXiv preprint arXiv:2406.14294*, 2024.
- [352] H. Wu et al., “Codec-superb: An in-depth analysis of sound codec models,” *arXiv preprint arXiv:2402.13071*, 2024.
- [353] H. Wu et al., *Towards audio language modeling – an overview*, 2024.
- [354] H. Wu et al., *Codec-superb @ slt 2024: A lightweight benchmark for neural audio codec models*, 2024.
- [355] J.-M. Lemerrier, S. Rouard, J. Copet, Y. Adi, and A. Défossez, “An independence-promoting loss for music generation with language models,” in *ICML*, 2024.
- [356] Q. Zhang, T. Xu, Y. Li, and Y. Wang, “Evaluating strong idempotence of image codec,” *arXiv preprint arXiv:2304.08269*, 2023.
- [357] T. Xu et al., “Idempotence and perceptual image compression,” in *ICLR*, 2024.
- [358] Y. Li, T. Xu, Y. Wang, J. Liu, and Y.-Q. Zhang, “Idempotent learned image compression with right-inverse,” in *NeurIPS*, 2024.
- [359] J.-H. Kim, S. Jang, J.-H. Choi, and J.-S. Lee, “Instability of successive deep image compression,” in *ACM MultiMedia*, 2020.
- [360] T. A. Nguyen et al., “Expresso: A benchmark and analysis of discrete expressive speech resynthesis,” *arXiv preprint arXiv:2308.05725*, 2023.
- [361] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “Sdr – half-baked or well done?” In *ICASSP*, 2019.
- [362] R. Ardila et al., “Common voice: A massively-multilingual speech corpus,” in *LREC*, 2020.
- [363] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [364] C. Wang et al., “VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 993–1003. [Online]. Available: <https://aclanthology.org/2021.acl-long.80>.
- [365] K. Ito and L. Johnson, *The lj speech dataset*. [Online]. Available: <https://keithito.com/LJ-Speech-Dataset/>.
- [366] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.
- [367] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” in *ICLR*, 2020.
- [368] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

- [369] N. Jovanović, I. Labiad, T. Souček, M. Vechev, and P. Fernandez, “Watermarking autoregressive image generation,” in *Advances in Neural Information Processing Systems*, 2025.
- [370] S. Aaronson, *My AI safety lecture for UT effective altruism*, Shtetl-Optimized, Accessed: 2026-06-30, Nov. 2022. [Online]. Available: <https://scottaaronson.blog/?p=6823>.
- [371] J. Kirchenbauer et al., “On the reliability of watermarks for large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- [372] Y. Wu, G. Milis, R. Chen, and H. Huang, “Robust distortion-free watermark for autoregressive audio generation models,” in *Advances in Neural Information Processing Systems*, Accepted, 2025. arXiv: 2510.21115 [cs.SD].
- [373] G. Milis, Y. Qin, Y. Wu, and H. Huang, “Hidden in plain tokens: Simply robust, gradient-free watermark for synthetic audio,” in *Proceedings of the International Conference on Machine Learning (ICML)*, Accepted, 2026. arXiv: 2605.25967 [cs.SD].
- [374] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, “Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations,” *ACM Conference on Computer and Communications Security (SIGSAC)*, 2020.
- [375] Y. Qin, N. Carlini, I. J. Goodfellow, G. Cottrell, and C. Raffel, “Imperceptible, robust, and targeted adversarial examples for automatic speech recognition,” in *ICML*, 2019.
- [376] W. Castro and Z. Yalniz. “Video invisible watermarking at scale,” Engineering at Meta. [Online]. Available: <https://engineering.fb.com/2025/11/04/video-engineering/video-invisible-watermarking-at-scale/>.
- [377] S. Goyal et al., *SynthID-Image: Image watermarking at internet scale*, 2025. arXiv: 2510.09263 [cs.CR].