ROOM FOR ERROR: LARGE-SCALE SIMULATION OF OVER-THE-AIR ACOUSTIC ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Studies of the risks facing AI models follow a well trodden path. However, acoustic models provide a unique challenge: utterances exist in both the physical and digital world. However, across the adversarial literature this challenge is either ignored with digital only evaluations, or tested within a single lab environment. This is a methodological and metrological shortcoming that undermines our understanding of risk. Through real-world testing, conceptual discussions, and a novel, high-throughput acoustic simulation framework, we demonstrate a new path for adversarial acoustics. This includes formalizing and operationalizing a novel *Dual-Form Signal to Noise Ratio* to decouple source stealth from victim attack efficacy, resolving a crucial limitation in current works. We support this with experiments that demonstrate that attackers with coarse room knowledge achieve relative Word Error Rate increases of between 26 and 71%, rising to 94.5% under an oracle upper bound. This work lays the groundwork for repeatable, verifiable research that embraces, rather than abstracts, the acoustic environment.

1 INTRODUCTION

The study of AI risk was kick started by the observation that computer vision systems are susceptible to adversarial perturbations—carefully crafted perturbations that, while imperceptible to humans, elicit anomalous behavior in vision models (Goodfellow et al., 2015; Carlini & Wagner, 2017). However, the AI landscape has fundamentally shifted since these early works, with text and voice user interfaces emerging as the central form of AI interaction. Such modalities are increasingly integrated with sensitive or private scenarios like smart devices or scribes, which makes understanding the risks facing models in these domains so crucial.

Carlini & Wagner (2018) were the first to demonstrate that Automatic Speech Recognition (ASR) are similarly susceptible to inaudible adversarial perturbations. However, the research that this has inspired primarily considers signals that often exist strictly in the digital domain—where the sample, perturbation, attacker, and model all exist as digital audio signals at the moment before they are passed to a model, suggesting a level of device compromise that would obviate the need to attack.

Digital only attacks abstract away most of the parts that make acoustic environments interesting, including attenuation, reflections, and the presence of distinct, physically located attackers, victims, and audio sources. As such, Over-the-Air (OTA) attacks have been considered, where the perturbation must survive playback through speakers, propagation through air, and recording by microphones (Yuan et al., 2018; Li et al., 2020; Qin et al., 2019; Szurley & Kolter, 2019; Olivier & Raj, 2023). However, these attacks are almost exclusively conditioned upon a single acoustic environment, complicating implementation and limiting generalization and replication.

This is inherently problematic, as the physical acoustic space is a harsh environment for adversarial perturbations. It acts as a complex, frequency-selective filter characterized by reverberation, attenuation, and background noise. While research has attempted to somewhat resolve these issues (Athalye et al., 2018; Chen et al., 2024; Wang et al., 2024), we have identified three critical flaws in the current evaluation methodologies for OTA attacks:

1. **The SNR Ambiguity:** Attack stealth is reported using a single Signal-to-Noise Ratio (SNR) scalar (e.g., "20 dB"). In a physical room, this is an ill-defined unless the measurement location is specified, and opens additional questions. Should this ratio exist at the victim's microphone

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(where the attack must overpower the clean signal to succeed) or at the attacker's speaker (where the attack must be quiet to remain undetected by the original speaker)? We argue that it is crucial that the community begins to standardize acoustic-awareness in metrics.
2. **The White-Box Geometry Fallacy:** Attacks either assume the attacker can either ignore, or precisely model, the room's acoustics properties to craft their perturbation. However precise measurement is rarely possible, and ignoring it obscures a key driver of performance. Fundamentally the degradation of attack success as this geometric knowledge fades is poorly understood.
3. **Statistical Insignificance:** Evaluations are typically conducted either purely digitally (no physical environment) or over-the-air in a single laboratory configuration (one environment). Where simulated reverberation has been used (Qin et al., 2019; Yakura & Sakuma, 2019), it has served as augmentation during attack *generation*, rather than as a distribution of environments over which attacks are *evaluated*. Neither of these captures the diversity of real-world acoustic environments.

Our Contributions. To address these gaps, our work makes the following contributions:

1. We formalize the **Knowledge Gradient**, a spectrum of attacker information ranging from naive digital attacks to white-box omniscience, exposing the "Information Cost" of room uncertainty.
2. We introduce **Dual-Form SNR Metrics** to explicitly decouple source stealth from target interference, revealing the energetic cost of projection often hidden in single-scalar reporting, formalising a preliminary approach initially articulated by Abdullah et al. (2021).
3. We demonstrate **high-throughput simulations of acoustically-aligned environments** for modeling of the impacts of simulated, known or unknown rooms.

These results are supplemented by a novel formalism of acoustic adversarial attacks, that allows us to articulate the opportunities for significant new, acoustically-aligned research in the Adversarial Machine Learning (AML) space. We stress that our contributions are model- and attack-agnostic, with the attacks (FGSM, PGD) and ASR models (Whisper, wav2vec2) evaluated here as representative demonstrations of why future work must account for acoustic alignment.

2 Acoustic Attacks

Formally, an acoustic attack against an ASR system can be defined as follows. Let $\mathcal{X} \subseteq \mathbb{R}^T$ denote a fixed length T stored representation of an acoustic waveform and $\mathcal{Y}$ the space of token sequences. An ASR model is then the mapping $M : \mathcal{X} \to \mathcal{Y}$, which assigns to an input waveform $x \in \mathcal{X}$ a transcription $y = M(x)$. An adversarial attack seeks a perturbation $\delta \in \mathcal{X}$ such that the perturbed transcription $y' = M(x+\delta)$ diverges from y, typically measured via the Word Error Rate $\mathrm{WER}(y, y')$ through either targeted or untargeted mechanisms.

We can generalize this scenario to consider the *information available to the attacker* through the information operator $H : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ denotes the attacker's observation space and $H(x)$ denotes the observation of the waveform available to the attacker. This formulation allows us to taxonomically classify attacks into three key classes: the case where $H(x) = x$, denoting perfect information; $H(x) = \emptyset$, where there is no sample specific information, leading to universal attacks; or $H(x; t) = (x_1, \ldots, x_t)$; a partial information case where the attacker observes a prefix of the waveform up to time t. The attack is also classified in terms of its strength, limiting δ to within some ℓ_p norm (typically $p \in \{2, \infty\}$) or perceptual measures (Qin et al., 2019; Schönherr et al., 2019; Szurley & Kolter, 2019; Sun et al., 2024), typically through the Signal-to-Noise Ratio (SNR)

$$\mathrm{SNR}(\delta, x) = 10 \log_{10}\big(\|x\|_p^p \,/\, \|\delta\|_p^p\big) \;\geq\; \mathrm{SNR}_{\mathrm{tgt}}. \tag{1}$$

While some perceptual measures have been employed in prior works, they were designed to assess speech quality and intelligibility under benign channel distortions, rather than to quantify adversarial impact or detectability, and all abstract away the physical nature of acoustics in ignoring the physical positions of the source p_s, the target/victim p_v, and the attacker p_a.

Relativity of Observations These limitations speak to the background of those performing adversarial machine learning—where expertise is concentrated upon the mechanics of machine learning—and the difficulties associated with real-world environmental testing. However, this has lead the physical nature of sound to be broadly ignored. Even when it has been noted in works like Abdullah et al. (2021), research repeatedly ignores or abstracts away the difference between sources of acoustic signals (be they the original audio source or the attacker), and the nature of perception across different

parts of the acoustic space, and how this attenuates victim information—assuming that H transfers a clean version of x. This observation leads us to believe that a diverse, acoustically-aligned evaluation suite is necessary for truly understanding the risks facing physically deployed ASR systems. Acoustic-alignment will allow these developments to respond to the properties of real acoustic channels, rather than simply transplanting techniques from computer vision to a new data modality.

As such, we explicitly consider SNR as a property at both the victim/target, which is correlated with attack performance, and at the source, which correlates with detectability. While it is true that target detectability may also be possible (especially using automated systems), we focus on the implicit threat model behind ASR research, in which the highest likelihood of detectability involves the original speaker observing the modified signal. To achieve this, we consider H to be a lossy physical transfer function $H_{p_i \to p_j}$ from a source location p_i to a target location p_j, that depends upon the environment ω and measurement infrastructure ψ. This reflects the fundamental physical reality, that there exists a **spatial mismatch** between the source transmission and both the attacker and victim observations of the signal. This in turn makes it clear that there is not one single constraint on the SNR, but potentially multiple, as the perturbed signal risks detection at both the source and victim. We capture this by introducing the **Dual-Form SNR**, which considers the SNR at the source (SNR_s) and victim (SNR_v) in term of the mean squared power P, replacning Equation 1 with

$$\text{SNR}_i = 10 \log_{10} \left(\frac{P_{H_{p_s \to p_i}(x)}}{P_{H_{p_a \to p_i}(\delta)}} \right) \qquad i \in \{s, v\} \ . \tag{2}$$

If the SNR is treated as a proxy for stealth, then evaluating stealth in situ may be impossible without sufficient information about (ω, ψ). Moreover, if the attacker is poorly positioned—due to distance or within a high-reverberation zone—the available information admitted by H may be insufficient to maintain the alignment necessary for a successful attack, regardless of the SNR_s and SNR_v, as the attacker must operate upon the observation $z = H_{p_s \to p_a}(x)$. We believe that this formalism of relativity is crucial, as it clearly underpins the nature of risk in real world systems.

Acoustic Contamination More critically, this formulation assumes H is independent of δ. However, in any real-time or adaptive attack, the **attacker's microphone is co-located with their own source of interference**. Consequently, the observation is not a passive window into x, but the mixture:

$$z = H_{p_s \to p_a}(x) + H_{p_a \to p_a}(\delta) \ . \tag{3}$$

Here $H_{p_a \to p_a}$ represents the near-field feedback of the perturbation. Because the distance $d(p_a, p_a)$ is typically small, the energy of δ dominates the observation space $\mathcal{Z}$ at the attacker's sensor. This creates a **deleterious feedback loop**, in that the moment that the learned mechanism f_θ emits a perturbation to influence the ASR, it blinds itself to the underlying signal x. This self-contamination significantly reduces the effective information (and SNR) for the attacker to a level where subsequent adaptation or continued alignment with the waveform x is physically impossible without sophisticated echo cancellation, which would still modify the available information.

3 Physical Attacks, and the Fundamental Realities of Scale

The content above suggests that there may well be no space for digital-only attacks to provide real insights into the vulnerabilities of ASR models, as they are simply too divorced from the environments in which these models actually exist. This is not to say that physical domain attacks are a viable pathway for uncovering risks either. Consider the mechanism of a typical adversarial attack: an input x is drawn from a distribution $\mathcal{D}$, and is passed through M to construct a perturbation δ via gradient access. However, if we intend to consider the physical domain, this process is constrained by:

- **Temporal Barriers**: Unlike image-based attacks, physical audio ingestion is lower-bounded by the utterance length. As we will discuss in subsequent sections, we consider 8 million experimental evaluations. In a physical environment, replicating this would take 925 days, synchronized playback, and this is before we consider the labor cost of reconfiguring $\mathcal{P}$.
- **Complexity Explosion**: For a true physically aware acoustic attack, the generator can not simply learn the mapping $f_\theta(x)$, but rather $f_\theta(x; \omega, \psi, \mathbf{p})$, where $\omega \in \Omega$ represents the acoustic environment, $\psi \in \Psi$ the hardware infrastructure, and $\mathbf{p} \in \mathcal{P}$ the relative geometry over $\{s, v, a\}$.

This last factor induces a prohibitive growth in sample complexity. It is not enough for a model to simply be robust against random noise. Instead, the physical domain requires robustness to be considered as a product of structured transformations and random elements. However, the number of observations required to generalize across the joint space of rooms Ω, hardware configurations Ψ, and spatial layouts $\mathcal{P}$ is orders of magnitude larger than has been acknowledged in any prior works.

To probe these limitations, we interrogated the performance of ASR systems in a complex L-shaped room (see Appendix A for details). Our empirical results expose a fundamental correlation collapse between digital metrics and physical outcomes. As is detailed in Table 1, while strictly digital perturbations exhibit a strong negative correlation ($r = -0.62$) between SNR and the Word Error Rate (WER), this relationship entirely dissipates in the acoustic environment ($r = -0.07$). These experiments were not designed as a 1-to-1 validation of our simulated evaluations—directly replaying 8 million attack evaluations over-the-air is physically infeasible (Appendix A)—but to test whether digital metrics remain predictive of transcription failure once signals propagate through a real room.

Table 1: Statistical correlations between WER and SNR for purely digital environments and acoustic environments with additive noise.

Condition	**Correlation** (r)	**95% CI**
Purely Digital	-0.62	$[-0.83, -0.24]$
Acoustic Noisy	-0.07	$[-0.46, 0.34]$

We further observe that there are baseline levels of environmental degradation in acoustic environments, that rival the influence of adversarial perturbations in purely digital domains. Even without an active attacker, natural acoustics push *Whisper-Base* to its operational limits, with baseline WERs ranging from 69.8% to 86.9%, compared with a reported clean WER of 2.7% for LibriSpeech (Radford et al., 2023). Spatial layout clearly dictates efficacy; with experimental configurations where the attacker is closer to the victim yielding a mean WER of 90.2%, while equidistant nodes suffer significantly higher path losses without a commensurate increase in transcription failure.

The simulator that we now introduce is not a perfect simulacrum of such environments. Instead, it leverages externally validated systems to bring acoustic evaluation and training meaningfully closer to deployment conditions. This will demonstrate that models deployed in unconstrained physical spaces may already be operating under severe environmental duress, heavily skewing their susceptibility to further perturbation. The inability of raw digital-only metrics to capture these nonlinearities, combined with the impossible human cost of managing large-scale physical configurations even for the simplest $\mathcal{P}$, underscores the necessity of our fast, acoustically-aligned simulation framework.

4 An Acoustic Simulation Framework

The preceding analysis establishes the fundamental tension in acoustic research: digital-only simulations are fast, and align with established workflows; while physical evaluations lack the scale necessary for statistical generalization. The impacts of this tension are only magnified for adversarial acoustics, due to the iterative mechanisms involved with adversarial attacks. To resolve this tension, we contend that researchers must adopt a middle ground that balances physical realism with high-throughput *acoustically-aware* simulations. This is not to say that perfect-realism should be the goal, rather we contend that research methods in spaces with underlying physics should move as close to those true physics as available resources allow.

Our objective is therefore not to develop a 1-to-1 digital twin of any specific room, which would require extensive per-room measurement and validation, and would be nigh on impossible to realize for all but the most permissive attack scenarios. Rather, we aim to provide a high-throughput, diverse acoustic evaluation suite. To draw an analogy, consider distributional robustness benchmarks like ImageNet-C (Hendrycks & Dietterich, 2019), where the aim is not to replicate any single deployment, but rather to systematically widen the distribution of conditions under which robustness is measured.

Modeling adversarial perturbations in physical environments necessitates a selection between wave-based and geometrical acoustics (GA) simulations. Wave-based methods, such as Finite Difference Time Domain (FDTD), provide high-fidelity solutions to the wave equation but scale as

Table 2: Comparison of acoustic simulation methodologies. ISM provides the necessary differentiability and throughput for large-scale adversarial optimization, producing significant alignment dividends over digital-only approximations covering computational cost, accuracy ($A(\cdot)$), gradient accessibility, implementation difficulty, and weaknesses.

Method	**Cost**	**A(Low-f)**	**A(High-f)**	**Grad.**	**Impl.**	**Weakness**
Wave-based	$\uparrow\uparrow\uparrow$	High	High	Difficult	$\uparrow\uparrow\uparrow$	$O(f^4)$ scaling; low iteration.
Ray Tracing	$\uparrow\uparrow$	Low	High	Stochastic	$\uparrow$	Unstable gradients; graininess.
ISM	$\downarrow$	Mod.	High	**Analytic**	$\downarrow$	Specular only; no diffraction.

$O(f^4)$ (Vorländer, 2008), rendering them prohibitively expensive for 16 kHz ASR signals. In contrast, GA methods like the Image Source Method (ISM) of Allen & Berkley (Allen & Berkley, 1979) treat sound as rays, providing a computationally tractable approximation that is among the most widely used and validated models of room acoustics. ISM-generated Room Impulse Responses (RIRs) have been shown to be competitive with measured RIRs for reverberant training (Ko et al., 2017; Kim et al., 2017; Szöke et al., 2019)

One of our aims with this work was to develop a high-throughput acoustic environment that can be applied to AML problems, producing what we dub an *acoustically-aligned* environment that extends this established model to a differentiable, adversarial pipeline. As the backbone of this, we leverage the ISM as implemented in **PyRoomAcoustics** (Scheibler et al., 2018) for its analytical, deterministic modeling of specular reflections, which yields a RIR that can be efficiently incorporated into differentiable network pipelines. The algorithmic details of this are described within Appendix D.

We note that such simulations are not the only viable pathway to incorporate the RIR into acoustic simulations. In fact, there exist several works that present databases of measured RIRs, that could be applied to these circumstances. However, these datasets are intrinsically tied to hardware idiosyncrasies, transducer non-linearity, and measurement noise (Stan et al., 2002). Moreover, these datasets share a similar limitation to physical experiments, in that they lack the capacity for continuous parameterization and dynamic control, inhibiting their capacity for properly understanding and exploring the influence of acoustic information on adversarial attacks.

Our framework is designed to generate a massive distribution of simulated environments. In doing so, we both demonstrate that it is possible to reach the levels of statistical significance in acoustic experimental evaluation, while simultaneously mitigating the risk of overfitting to the artifacts of a single setup. In essence, we trade the localized fidelity of a measured RIR for the global statistical significance of a high-throughput simulation, ensuring our findings describe fundamental properties of reverberant propagation rather than artifacts of a specific laboratory configuration or physical testing environment. For this reason we deliberately leave device-specific artifacts (including transducer nonlinearities or automatic gain control) to future work, as our aim is to widen the space of acoustic environments under which robustness is evaluated, rather than to replicate a particular device.

4.1 The Knowledge Gradient: Quantifying Information Cost

Our high-throughput capability allows us to formalize the **Knowledge Gradient**, which quantifies the Information Cost associated with environmental uncertainty. We explore how an attacker's utility degrades as their knowledge of the transfer function H (and thus the RIR h) fades through

- **Naive ($\mathcal{K}_{naive}$):** Attacker assumes identity channels ($h \approx \delta_{dirac}$), acting as if in a digital domain.
- **Blind ($\mathcal{K}_{blind}$):** Attacker uses an EoT ensemble of 3 rooms sampled from a distribution centered on the target room dimensions with $\pm 20\%$ variance, representing coarse knowledge of scale.
- **Approximate ($\mathcal{K}_{approx}$):** Tighter $\pm 10\%$ variance but only one randomly sampled room, reflecting imperfect measurement without optimization over uncertainty.
- **Oracle ($\mathcal{K}_{oracle}$):** Perfect pairwise RIR knowledge, analogous to white-box AML, serving as an upper bound on attacker performance rather than a realistic attacker. (All levels assume white-box access to the model; they differ only in environmental knowledge.)

Table 3: Comparison of Audio Adversarial Frameworks. **Dual-SNR** refers to separating Source (Stealth) and Victim (Interference) metrics. **Blind** refers to testing on unknown room geometries. **Physics** denotes if the method explicitly models room acoustics.

Method	Attack Type	Stealth Metric	Dual SNR	Blind Eval	Physics Model
C&W (2018)	Digital	L_p Norm	No	No	None
CmdrSong (2018)	Music	User Study	No	No	Air
Qin et al. (2019)	PsychoAc	Masking	No	No	Sim
Yakura (2019)	EoT	SNR	No	Yes	RIR+
Devil's (2020)	BlackBox	User Study	No	Yes	None
AdvPulse (2020)	Burst	User Study	No	Yes	Air
AdvReverb (2024)	RIR-Gen	User Study	No	No	RIR
Ours	**Eval**	**Dual-SNR**	**Yes**	**Yes**	**RIR++**

4.2 Acoustically-Aligned Evaluation Framework

Prior research has pivoted toward naturalistic disguises, such as *AdvReverb* (Chen et al., 2024), which leverages reverberation to hide adversarial perturbations. However, we challenge the practical efficacy of this paradigm, in that while the *format* of the noise is less suspicious, the *energy* required to maintain alignment across a room may still render it detectable. As such, our framework shifts the focus from *generation* to **rigorous evaluation**. Rather than proposing new attacks, we consider why attacks that appear locally stealthy can be energetically visible when accounting for physics. The backbone of our evaluation involves adversarial attacks that—subject to a source-side energy budget c—seek to find a perturbation δ to maximize ASR loss at the victim

$$\max_{\delta} \mathcal{L}_{\text{ASR}}(M(H_{p_s \to p_v}(x) + H_{p_a \to p_v}(\delta)), T) \text{ s.t. } \text{SNR}_s \geq c \ . \tag{4}$$

Ideally, stealth would be measured on the energy actually received at the source's microphone, $\|h_{as} * \delta\|_2 \leq c$. However, projecting onto this set has no closed form and would require an inner iterative solve at every optimization step (and for every room in the Blind ensemble). Since $H_{p_a \to p_s}$ is linear and time-invariant, Young's Convolution Inequality (Young, 1912) gives $\|h_{as} * \delta\|_2 \leq \|h_{as}\|_1 \|\delta\|_2 \leq \|h_{as}\|_1 \sqrt{T}\, \|\delta\|_\infty$, so constraining the broadcast signal ($\|\delta\|_\infty \leq \epsilon$, calibrated as in Appendix G) yields an analytic upper bound on the energy received at the source, with a closed-form projection (clipping). This formulation explicitly treats the RIR as an energetic barrier rather than a disguise, as seen in Table 3. We dub the difference $\Delta\text{SNR} = \text{SNR}_s - \text{SNR}_v$ the **Projection Cost**. Importantly, the relaxation above only restricts the set of perturbations the optimizer may consider: **all reported SNR_s, SNR_v and Projection Cost values are computed on the true convolved signals**, so the spatial decoupling we report is measured, not bounded. As attenuation and reverberation differ along the attacker–source and attacker–victim paths, the Projection Cost is not a constant but a function of room geometry, revealing the energetic cost (or advantage) of acoustic projection.

We achieve these attacks using two core untargeted attack frameworks. The Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) generates δ by taking a single step in the direction of the gradient sign. This approach is generalized by its iterative refinement, Projected Gradient Descent (PGD) (Madry et al., 2018), which traverses the loss landscape of the model f to maximize error by

$$x_{t+1} = \Pi_{x+\mathcal{S}} \left(x_t + \alpha \cdot \text{sign}(\nabla_{x_t} \mathcal{L}_{\text{ASR}}(f(x_t), T)) \right) \tag{5}$$

where $x_0 = x$ an initial known signal, $\mathcal{L}_{\text{ASR}}$ is the ASR loss function (typically Cross-Entropy or CTC/Connectionist Temporal Classification), α is the step size, T is the original transcription, and $\Pi_{x+\mathcal{S}}$ is a projection operator that ensures the adversarial example remains within both an ϵ-bounded constraint set $\mathcal{S}$ around the original signal x and the valid data range, enforcing the c constraint. Single-step FGSM attack simply set the step size to equal the total budget ($\alpha = \epsilon$).

We emphasize that our contributions—which relate to the level of information available to the attacker, the relativity of SNR, and the value of acoustically-aligned evaluation environments—are not predicated upon the specific form of the attack. Changing the attack mechanism does not change the level of information accessible to the attacker, nor does it fundamentally change the limitations of the digital-only environments tested within prior works. Finally, the scale of our experiments

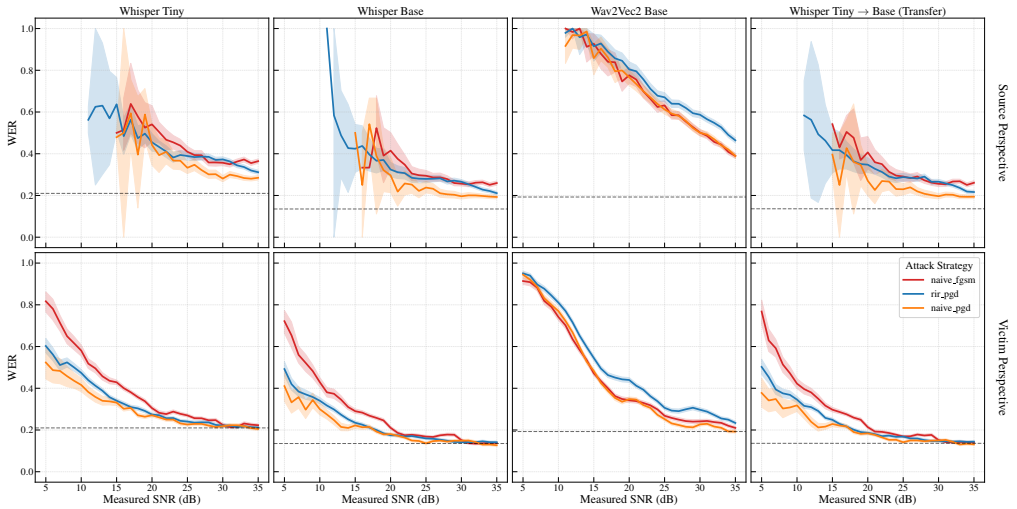

Figure 1: Gradient Misalignment across models, showing SNR at the Source (Top) or Victim (bottom) required to reach a WER. Shading represents bootstrapped 95% confidence intervals.

(which we will now discuss) would make the more complex optimization loops for alternative attacks computationally prohibitive. A discussion of the implementation of our approach, and the computational considerations associated with it can be found in Appendix D.

5 Experimental Results

To statistically cover the long-tail variance of acoustic environments, we conduct this evaluation at massive scale, processing 200 unique audio samples across 50 distinct room geometries. Unlike standard digital evaluations that test a single attack configuration per sample, our framework performs a dense hierarchical sweep for every one of the 10,000 sample–environment pairings:

- **Attack Generation (32 Variants):** For each sample, we generate the Cartesian product of 2 Optimizers (FGSM, PGD), 4 Target SNRs (15–45 dB, see Appendix G for details), and 4 Knowledge Levels (Naive, Blind, Approx, Oracle).
- **Attack Evaluation (8 Contexts):** Each variant tests 4 Transfer Scenarios and 2 Defense States.

256 distinct interference and metric calculations were performed for every sample. Across 4 model configurations, this resulted in over 10 million scheduled attack evaluations (8 million completed, due to runtime limits), providing a statistically rigorous mapping of the acoustic attack landscape. Appendix D contains additional information about the process, and Appendix A demonstrates how physically repeating these simulations may take more than 925 days.

Our attacks are constructed under an ℓ_∞ constraint, where the perturbation budget ϵ is calibrated to an approximate target SNR, based upon a pre-computational analysis of 200 room configurations. The difference between this and the SNR is deliberate: the ℓ_∞ constraint prevents the optimizer from concentrating energy in isolated, highly detectable bursts, while the ℓ_2-based SNR provides a physically meaningful measure of the models perception of environmental impact and interference.

5.1 Gradient Misalignment and Robustness

Table 4 and Figure 1 reveals that rather than physical awareness universally improving attacks, our framework exposes stark difference in model acoustic vulnerabilities. For encoder-only models like wav2vec2, the Knowledge Gradient is effective, with the Oracle RIR (an upper bound) yielding a 94.5% relative increase in the WER under attack. Conversely, for encoder–decoder models like Whisper, we paraxoically observe that the Naive FGSM attack (which ignores room acoustics entirely) performs comparably to, or even outperforms, attacks with perfect room knowledge.

Table 4: Word Error Rates for a range of experimental configurations, relative to a per-model baseline. Here Transfer represents Whisper Tiny $\rightarrow$ Base (Transfer).

		Baseline	Naive (Identity)	Blind (Dist.)	Approximate	Oracle (Known)
Whisper Tiny	FGSM	0.21	0.264 (+25.7%)	0.226 (+7.8%)	0.226 (+7.8%)	0.226 (+7.8%)
	PGD	0.21	0.239 (+13.8%)	0.256 (+22.1%)	0.257 (+22.1%)	0.265 (+26.0%)
Whisper Base	FGSM	0.135	0.176 (+29.9%)	0.146 (+8.0%)	0.146 (+8.0%)	0.146 (+8.1%)
	PGD	0.135	0.155 (+14.8%)	0.17 (+25.9%)	0.17 (+26.0%)	0.176 (+30.4%)
wav2vec Base	FGSM	0.193	0.328 (+70.3%)	0.24 (+24.5%)	0.24 (+24.4%)	0.24 (+24.4%)
	PGD	0.193	0.325 (+69.0%)	0.329 (+71.0%)	0.326 (+69.4%)	0.375 (+94.5%)
Transfer	FGSM	0.136	0.178 (+30.8%)	0.148 (+8.7%)	0.148 (+8.7%)	0.148 (+8.8%)
	PGD	0.136	0.157 (+15.4%)	0.172 (+26.6%)	0.173 (+26.9%)	0.179 (+31.7%)

This one-step inverse failure is a signature of Gradient Misalignment in reverberant channels. To interrogate the drivers behind this counterintuitive phenomenon, we return to the attack mechanism. In the naive case, the ∇_{naive} of Equation 5 is computed on the clean model, producing a resultant perturbation that exhibits broadband phenomenology, across the entire spectral range. While the room h distorts this noise, enough spectral energy survives to disrupt the ASR.

By contrast, the RIR-aware attack in frequency domain backpropagates through the channel by way of $\nabla_{\text{rir}} = \nabla_{\text{loss}} * h_{\text{reverse}}$, so that the gradient is weighted by the conjugate channel response. Real rooms are strongly frequency-selective, which makes the channel inherently ill-conditioned and ensures that gradients in the attenuated bands are heavily suppressed, even without exact zero-gain nulls. The single-step $\text{sign}(\cdot)$ operation then discards this magnitude information, spreading perturbation power across the spectrum, including into bands the room will attenuate before the signal reaches the victim. We hypothesize that this wastes a substantial fraction of the attack budget. Iterative PGD is not subject to the same constraint, as repeated small steps let the optimizer steer away from these attenuated bands, consistent with RIR-aware PGD outperforming Naive PGD (Table 4: 0.176 vs. 0.155 on Whisper Base; 0.375 vs. 0.325 on wav2vec2 Base). While not a total failure, RIR-FGSM provides only a marginal lift (+8% for Whisper Base) compared to Naive FGSM (+29.9%).

5.2 The Acoustic Tax and Projection Cost

Figure 2 visualizes the inherent decoupling of Source and Victim SNR, with our analysis revealing a significant perceptual disconnect between the two locations: an attack that achieves a successful interference level of $\text{SNR}_{\text{victim}} \approx 20$ dB at the target often manifests as a highly stealthy $\text{SNR}_{\text{source}} \approx 40$ dB at the source monitoring microphone. This is reinforced by Figure 7 demonstrating that WER is primarily correlated with victim SNR. This strongly suggests that future geometrically aware attacks—particularly those using directional acoustic devices—could further exploit this decoupling to yield attacks that are **both stealthy and efficacious**.

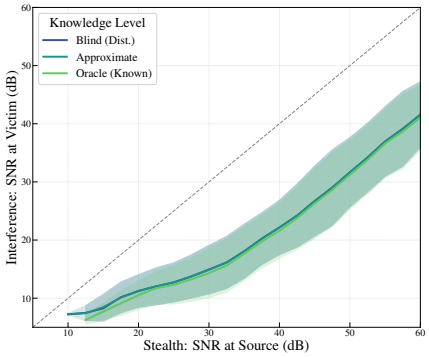

Figure 2: Divergence from the identity line (dashed) quantifies the Acoustic Tax inherent in OTA projection across tested room geometries. Shading: 1σ variance.

The heatmap in Figure 6 quantifies this relationship, demonstrating that the gap between victim interference and source audibility is not a constant but a function of room geometry and reverberation. This result provides a critical physical nuance to prior works on perceptual stealth (Szurley & Kolter, 2019); **while an attack may be masked psychoacoustically at one location, its energetic signature is non-uniform, with a differential of over 30 dB for large, reverberant spaces $\mathbf{RT_{60} > 0.6}$ s)**. This implies that the carrier device's own monitoring system is a poor proxy for model impact, and conversely, **that reporting a single SNR scalar is a physically incomplete measure of an attack's presence**. This spatial invariance shown in Figure 4 further reinforces the disconnect: for an attacker increasing broadcast power to compensate for distance (d_{AV}), the impact on the model remains consistent, confirming that OTA attacks are limited not by placement, but by the attacker's ability to balance minimizing energetic visibility with attack utility.

5.3 Model Architecture and Transferability

As quantified in the Oracle RIR column of Table 4, Wav2Vec2 (encoder-only) proved significantly more brittle to reverberant interference than Whisper (encoder-decoder), showing a 94.5% relative increase in WER compared to Whisper Base's 30.4%. Adversarial examples generated on Whisper-Tiny transferred effectively to Whisper-Base, maintaining a consistent 31.7% lift over baseline, suggesting a robustness against model scale.

Appendixes E and D further contextualize this by considering defences and computational costs, with the physical implications of acoustic modeling, including the Acoustic Tax (Figure 6), the preservation of signal quality (Figure 5), and the impact of spatial properties found in Appendix B.

Implications The initially counterintuitive nature of some of our results emphasizes that acoustic attack performance is clearly highly non-linear in the channel. Environmental knowledge clearly benefits iterative attacks (RIR-aware PGD outperforms Naive PGD across models), but degrades single-step attacks, where channel ill-conditioning undermines RIR-aware FGSM. Meanwhile, Naive FGSM matches Oracle PGD on Whisper Base (0.176 WER), suggesting that brute-force broadband interference remains a potent and low-cost threat. Neither observation would be visible in a digital-only evaluation, nor from a single physical room. Acoustic AML has largely inherited mechanisms from computer vision; our results suggest substantial room for attacks designed around the properties of acoustic channels—for example, optimizing for robustness to environmental uncertainty, or allocating budget to bands that survive propagation—through our framework.

5.4 Limitations

As discussed previously, our simulation framework relies upon prior experimental validations of the ISM method, with physical experiments (Appendix A) restricted to bounded white noise rather than optimized perturbations. However, we stress that our framework is designed as a tool to increase acoustic diversity and evaluating distributional robustness across room configurations, rather than a 1-to-1 digital tin of specific rooms or transducers. As the ISM model simulates propagation without diffraction or background noise (Appendix C) can produce RIRs that deviate from physical measurements, and makes our simulator likely optimistic, with physical baseline WERs (69.8–86.9%) far exceeding simulated ones, suggesting that our results are in fact a *lower bound on environmental difficulty*. Nonetheless, our physical experiments show that digital-only metrics lose their predictive relationship with transcription failure under real propagation ($r = -0.62$ digitally vs. $r = -0.07$ physically), motivating room-aware evaluation.

6 Conclusion

Our high-throughput acoustically-aligned simulation framework has allowed us to elucidate two fundamental properties of acoustic AML: the Knowledge Gradient and SNR relativity. Our analysis reveals that by abstracting away the acoustic environment we may have both overestimated realisable stealth, and underestimated the brute-force simplicity required to influence model predictions.

More broadly, this work demonstrates how a physics-based lens reveals problems that are obscured within digital-only environments. We demonstrate that digital-only insights do not necessarily translate to physical simulations; specifically, the energetic disconnect between emissions and reception at the source and victim creates an acoustic tax that dictates the feasibility of any real-world attack. Our results suggest that while an attack may be masked psychoacoustically at one location, its energetic signature is non-uniform across the acoustic space, rendering single-scalar SNR reports an incomplete and potentially misleading measure of an attack's presence.

We emphasize that our framework does not claim to represent a perfect acoustic digital twin, and replicating these results in a physical environment would take hundreds of days. Rather, we contend that for research in modalities with underlying physical constraints, investigators must move as close as possible to that physical reality as available resources allow, without compromising on statistical viability. Such an approach will provide the metrological foundations for a new era of systemic robustness evaluation, moving the field beyond mathematical convenience toward a deeper understanding of the risks facing deployed ASR.

AI USE STATEMENT

This paper received LLM based support in two different ways. The first has been to provide editorial support in the preparation of the manuscript, which has been used as a secondary point of validation for checking for issues with structure and grammar. It is important to stress though that this was strictly limited to copy-editing, rather than using the LLM to write (or rewrite) key content.

The secondary use was to provide support with generating supplementary code, and ensuring that constructed code was aligned with the content of the paper.

ETHICS STATEMENT

This work explores the impact of changing how we enhance adversarial attacks against acoustic systems through a deeper acoustic simulation layer. Superficially, such advances have the potential to improve the performance and transferability of adversarial attacks against real world systems, which has the potential to deliver real world harms.

However, with this said, any security dividends from focusing upon strictly digital acoustic environments in AML research is illusory. Risks and harms will always be present, and it is a well established principle that there is no security by obscurity. Failing to examine the true nature of acoustic signals obscures our appreciation of risk, while failing to secure the very systems we as researchers purport to be concerned about.

We would also stress that current evaluation practices can cause real harm, due to how we present safety to downstream user communities. To expand upon this, we note that a common precept within the Adversarial Machine Learning community is to assume a particular threat model, with the nature of academic comparisons often incentivizing us to then follow in the footsteps of those who came before us. However, in doing so, we inadvertently create—and, crucially, present—a myopic view of the risk landscape. In essence, we portray to people outside our academic community that any risks they face will be aligned with the areas in which we act as a community. However, the drivers for academic investigations can oftentimes be driven by mathematical convenience and historical precedent, rather than a close alignment with current systemic risks and harms. We fundamentally believe that acoustic AML research has been pursuing mathematical convenience over making forward steps on appropriate evaluations, and that this convenience will inherently limit the value of outcomes produced by the AML community.

We emphasize the above point not just for the risks of erroneous portrayals of risk to downstream practitioners, but also because our focus on these spaces inherently biases real attacker behavior away from these threat models. After all, if an attacker understands that defenses against manipulation have been calibrated on digital-only simulations, then they're naturally incentivized to consider an alternative pathway for model manipulation, that factors in more real world acoustic factors.

Finally, we would also note that there are some applications where a lack of robustness in a model may be positive. In a world where broad scale surveillance is increasingly normalized, it may well be the case that adversarial attacks may induce privacy, creating a net public good.

With these points made, we still believe that research into defenses induces a net societal gain. Improving robustness to natural or adversarial perturbations will improve the performance of systems that are already one of the dominant access portals for AI within the community, and achieving this will only be possible if we improve the quality of our evaluations.

REPRODUCIBILITY STATEMENT

This work was conducted under a funding agreement that places limitations on the public provision of associated resources (including code) until it has undergone a formal review by the funding partner. While this paper has undergone review, the funding body is unable to commit the resources for code review until the paper has been formally accepted.

With that said, we have received confirmation that the code will be approved upon publication. At this point, we will publish a github that contains the code, key running scripts, and details of the employed

evaluation environment, to maximize the likelihood of researchers being able to independently implement our approaches and recreate our results.

REFERENCES

Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. In *IEEE Symposium on Security and Privacy (SP)*, pp. 730–747, 2021.

Jont B. Allen and David A. Berkley. Image Method for Efficiently Simulating Small-Room Acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 4 1979.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing Robust Adversarial Examples. In *International Conference on Machine Learning (ICML)*, pp. 284–293. PMLR, 2018.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12449–12460, 2020.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Nicholas Carlini and David Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018.

Meng Chen, Li Lu, Jiadi Yu, Zhongjie Ba, Feng Lin, and Kui Ren. AdvReverb: Rethinking the Stealthiness of Audio Adversarial Examples to Human Perception. *IEEE Transactions on Information Forensics and Security*, 19:1948–1962, 2024.

Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 2667–2684, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations, ICLR*, volume 3, 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara N Sainath, and Michiel Bacchiani. Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home. In *Proc. Interspeech 2017*, pp. 379–383, 2017.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224. IEEE, 2017.

Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. AdvPulse: Universal, Synchronization-free, and Targeted Audio Adversarial Attacks via Subsecond Perturbations. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pp. 1121–1134, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, volume 6, 2018.

Raphael Olivier and Bhiksha Raj. There is More Than One Kind of Robustness: Fooling Whisper with Adversarial Examples. In *INTERSPEECH*, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In *International Conference on Machine Learning (ICML)*, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*, volume 40, pp. 28448–28467. PMLR, 2023.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A Python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 351–355. IEEE, 2018.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In *Network and Distributed System Security Symposium (NDSS)*, 2019.

Guy-Bart Stan, Jean Jacques Embrechts, and Dominique Archambeau. Comparison of Different Impulse Response Measurement Techniques. *Journal of the Audio Engineering Society*, 50: 249–262, 04 2002.

Zheng Sun, Jinxiao Zhao, Feng Guo, Yuxuan Chen, and Lei Ju. CommanderUAP: Practical and Transferable Universal Adversarial Attacks on Speech Recognition Models. *Cybersecurity*, 7(1): 38, 2024.

Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černockỳ. Building and Evaluation of a Real Room Impulse Response Dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

Joseph Szurley and J Zico Kolter. Perceptual Based Adversarial Audio Attacks. *arXiv preprint arXiv:1906.06355*, 2019.

Michael Vorländer. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer, 2008.

Li Wang, Jiaqi Li, Yuhao Luo, Jiahao Zheng, Lei Wang, Hao Li, Ke Xu, Chengfang Fang, Jie Shi, and Zhizheng Wu. AdvSV: An Over-the-Air Adversarial Attack Dataset for Speaker Verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4555–4559. IEEE, 2024.

Hiromu Yakura and Jun Sakuma. Robust Audio Adversarial Example for a Physical Attack. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5334–5341. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/741.

William Henry Young. On the Multiplication of Successions of Fourier Constants. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(596):331–339, 1912.

Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A Gunter. CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition. In *USENIX Security Symposium*, 2018.

A Motivating Physical Experiments

As Section 3 discussed, practical physical validation of adversarial attacks very quickly becomes nigh-on impossible, when attempting to consider the physical environment. Consider for example

a PGD style attack that is run for n iterations on s samples that is t seconds long, over v different physical configurations. Even a model that can process samples faster than the time it takes to listen to them still must listen to the sample first, so, if r is the real-time-factor associated with processing, naively, the computational time associated with an attack τ is strictly $\tau > (1 + r) \times n \times s \times t \times v$.

Complicating this is the need for the signal environment to be quiet at the beginning of each attack, which requires that there is a delay between iterations of the attack, and the signals at each of the three node locations (source, attacker, victim) must be synchronized. Moreover, each of the v different physical configurations requires physical intervention to manually change the configuration.

When sample lengths are on the order of 30 seconds, a PGD attack involving 50 steps (which may be ultimately conservative due to the nature of acoustic hulls) that requires 5 seconds of silence to reset the acoustic environment will require $\tau > 30(1 + r)$ minutes to attack a single sample, in a single room configuration. Complicating things further is that no realistic environment is acoustically isolated, and as such there is a need to repeat attacks against samples multiple times to ensure rigor, producing a further multiplicative scaling for the costs and associated times.

As such, conducting enough experiments to achieve statistically significant results under such a framework is practically infeasible when considering time and labor costs. This observation underscores the utility of our acoustically-aligned simulations for exploring the robustness and security properties of models that are deployed in the real-world, without requiring the time associated with real-world simulations.

Rather than attempting to validate our 8 million physical experiments directly, we considered two different perspectives. The first was to leverage an externally-validated modeling framework for producing RIRs, as was discussed within Section 4.

The second validation stage was to consider a physical experiment using an over-the-air attack with our three nodes positioned within a complex L-shaped room with maximal dimensions of 5m $\times 6$m, with furniture, open windows, and open doors. Within this environment, rather than testing PGD or FGSM style attacks, we considered strictly the influence of bounded white noise, to limit the temporal utilization of the room environment.

Within this acoustic space, we positioned our three nodes in six different physical configurations, detailed as follows:

- **Scenario 1**: The nodes are positioned in the ordering of Victim, Source, Attacker, with a spacing of 1 m between nodes.
- **Scenario 2**: Same as Scenario 1, however with a spacing of 3 m between nodes.
- **Scenario 3**: The nodes are positioned in the ordering of Victim, Attacker, Source with a spacing of 1 m between nodes.
- **Scenario 4**: Same as Scenario 3, however with a spacing of 3 m between nodes.
- **Scenario 5**: The nodes are positioned in an equidistant manner, with a spacing of 1 m between nodes.
- **Scenario 6**: Same as Scenario 5, however with a spacing of 3 m between nodes.

Over these six different physical configurations, we executed a factorial experimental design across three independent variables: noise magnitude, source-mixing mode, and signal content.

To ensure statistical representativeness and mitigate speaker-dependent bias, we utilized a fixed corpus of 10 randomly sampled utterances from the LibriSpeech *clean-100* dataset. For each spatial configuration, we evaluated two distinct operational modes for the attacker. In **Mode 1 (Remote Interference)**, the source node broadcast clean speech while the attacker node broadcast bounded Gaussian white noise from a distinct spatial coordinate. In **Mode 2 (Local Interference)**, the attacker noise was computationally mixed with the speech signal and broadcast solely from the source node, while the remote attacker node remained silent. This allows for an empirical comparison between local signal corruption and remote acoustic masking.

The noise intensity was varied across three magnitudes ($\epsilon \in \{0.0, 0.5, 1.0\}$), where $\epsilon = 1.0$ represents a peak amplitude normalized to the signal's maximum digital headroom. To ensure temporal synchronization across the distributed nodes, each trial was preceded and followed by a $5,000$ Hz synchronization pulse (50ms duration) separated by 100ms silence buffers. This enabled precise

post-hoc temporal alignment via cross-correlation, allowing us to isolate the content signal and factor out network-induced trigger jitter.

For the final evaluation, we utilized the *Whisper-Base* encoer-decoder-based ASR model. We computed the Word Error Rate (WER) by comparing the victim's captured transcription against a baseline transcription of the digital source files. Simultaneously, we quantified the acoustic channel degradation using two metrics: *Absolute SNR*, calculated relative to the pristine digital reference to measure total path loss and reverberation, and *Relative SNR*, calculated against the $\epsilon = 0.0$ baseline recording to isolate the specific masking efficacy of the added white noise.

In total, each physical session comprised 360 trials (10 samples $\times$ 3 magnitudes $\times$ 2 modes $\times$ 6 physical configurations), producing 1080 distinct physical recordings for every test. To ensure performance, we then repeated each of these tests 3 times. To prevent hardware-level automatic gain control (AGC) or non-linear compression from contaminating the results, the source playback was scaled to a constant 0.75 of the digital maximum.

That each trial required more than 5 hours (15 hours when factoring in the 3 repetitions) of time clearly demonstrates the futility of large scale experimental evaluations, and the critical importance of studying acoustics in acoustically-aligned environments, where real-world physics factor into experimental outcomes. Our high-throughput simulation framework is vital for scaling acoustic AML research beyond toy examples. Replicating the breadth of our experimental results would not be possible under real world conditions.

This is especially so when these scenarios required constant human engagement to change physical configurations, and monitor for intrusive external noises that required re-running a particular experiment. And, crucially, these experiments are not adaptive—incorporating a multi-step attack framework would lead to a multiplicative increase in the associated times and costs.

A.1 Experimental Results

Our physical validation experiments yield several critical insights that challenge prevailing assumptions in adversarial machine learning for acoustic systems, highlighting the stark contrast between digital simulations and physical realities.

The Disconnect Between Digital SNR and Physical Transcriptions Most notably, the data reveals a fundamental collapse in the correlation between Signal-to-Noise Ratio (SNR) and Word Error Rate (WER) when transitioning from digital to physical environments. As detailed in Table 1, purely digital perturbations exhibit a strong negative correlation ($r = -0.62$), suggesting that as SNR decreases, WER predictably increases. However, in physical, noisy acoustic environments, this correlation entirely dissipates ($r = -0.07$). This empirically demonstrates that evaluating acoustic adversarial attacks relying solely on digital-environments is methodologically flawed; real-world variables such as multipath propagation, room impulse response (RIR), and destructive interference exert a non-linear influence on model transcription accuracy that raw digital-only metrics cannot capture. This breaks the implicit assumption in past works that digital only SNR and WER would be meaningfully correlated, and aligns with our acoustically-aligned numerical experiments.

Baseline Environmental Degradation These results clearly demonstrate that the physical space acts as a formidable adversary to acoustic models, underpinning the utility for acoustic-awareness for both attacks and general model performance. Even in the absence of an active attacker (Magnitude 0.0), the baseline WER for the Whisper-Base model ranges from 69.8% to 86.9% in Table 5. The acoustic filtering of our physical testing environment, combined with standard physical obstacles, pushes the ASR model near its operational limits, even before any malicious payload is introduced. This highlights a critical, inherent vulnerability: models deployed in unconstrained physical spaces are already operating under severe environmental duress, heavily skewing their susceptibility to further perturbation.

Spatial Dynamics and the Non-Linear Nature of Vulnerabilities Unsurprisingly, the physical layout of the nodes heavily dictates the efficacy of the attack, although the actual results do not necessarily fully align with mathematical intuition. Table 6 illustrates that configuring the attacker closer to the victim yields the highest mean WER at 90.2%, as the attacker's broadcast suffers less

environmental attenuation than the target speech. Interestingly, the equidistant configurations produce the most severe Mean Absolute SNR (-30.9 dB), but a marginally lower Mean WER (89.3%). This anomaly further corroborates the findings from Table 1, in that severe signal degradation does not strictly guarantee a corresponding maximization of transcription failure, when the acoustic signal is exposed to physical attenuation.

Table 5: SNR, WER, and estimated environmental loss across different spatial configurations and noise scales. Relative SNR compares the heard signal at the victim without noise versus with noise. Absolute SNR compares the heard signal at the victim against the original digital signal. Scenarios 1–2: source is closer to the victim; Scenarios 3–4: attacker is closer; Scenarios 5–6: equidistant. Odd numbered scenarios denote 1m spacing between devices, while even scenarios denote 3m spacing.

Scenario	Metric	Noise Magnitude		
		Mag 0.0	**Mag 0.5**	**Mag 1.0**
1	Abs SNR	-12.5	-16.2	-16.3
	Rel SNR	∞	6.1	-13.8
	WER	69.8%	85.5%	91.3%
2	Abs SNR	-29.6	-41.7	-38.7
	Rel SNR	∞	-45.1	-47.9
	WER	86.9%	87.6%	94.2%
3	Abs SNR	-34.0	-30.3	-29.6
	Rel SNR	59.8	-47.4	-65.1
	WER	78.7%	105.8%	94.1%
4	Abs SNR	-29.3	-31.1	-27.7
	Rel SNR	∞	-3.8	-21.6
	WER	75.9%	87.9%	99.7%
5	Abs SNR	-35.1	-38.5	-33.9
	Rel SNR	∞	-9.6	-2.2
	WER	75.5%	92.5%	96.2%
6	Abs SNR	-23.2	-27.4	-28.4
	Rel SNR	∞	-27.8	-16.7
	WER	81.4%	95.2%	97.2%

Distance Penalties and Remove Masking Efficacy Finally, these experiments highlight the harsh realities of the inverse-square law upon practical attacks, and the crucial importance of physical configuration in determining attack performance. Increasing the node spacing from 1 to 3 meters incurs severe distance penalties. For instance in Scenario 3, the Magnitude 0.5 case produces a WER of 105.8%, indicating severe model hallucinations. However, extending the distance to 3 meters (Scenario 4) allows the room to diffuse the impact of this effect, capping WER at 87.9%. Despite these attentuation penalties, the precipitous drops in Relative SNR across the active attack scenarios proves that remote acoustic masking can aggressively overpower the source signal at the target microphone. This validates over-the-air remote interference as a highly potent attack vector, independent of local signal corruption.

Table 6: Mean Absolute SNR and Mean WER across varying spatial layouts.

Spatial Layout	**Mean Abs SNR**	**Mean WER**
Source Closer (S1, S2)	-25.2 dB	85.7%
Attacker Closer (S3, S4)	-30.4 dB	90.2%
Equidistant (S5, S6)	-30.9 dB	89.3%

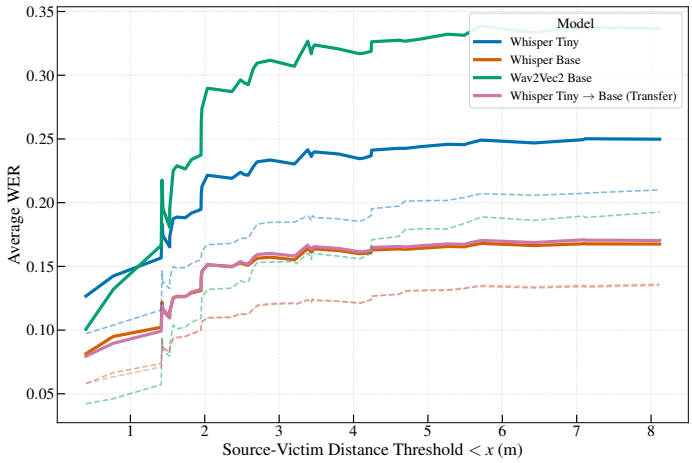

Figure 3: Relationship between average WER and Source-Victim Distance. Attacked performance is represented by solid lines, clean performance by dashed lines.

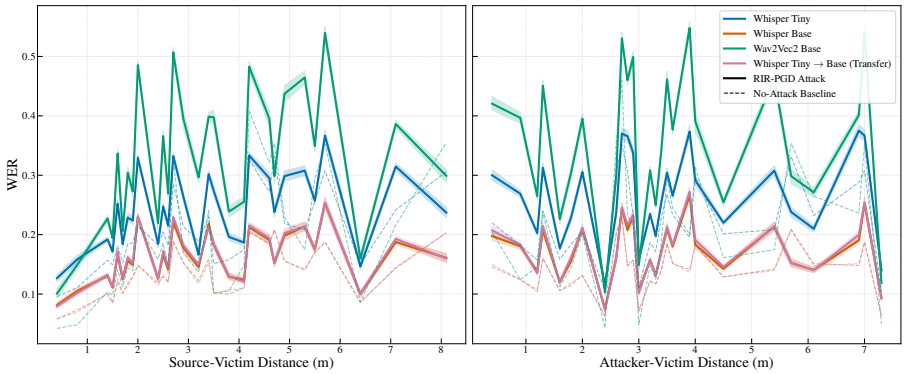

Figure 4: Impact of spatial geometry on WER. Success scales primarily with Source-Victim distance (acoustic distortion of the carrier) but remains largely invariant to Attacker-Victim distance, provided the attacker meets the energetic cost of projection.

B Physical Implications of Acoustic Modeling

While Figure 4 exhibits a relatively chaotic dynamic between individual Source–Victim distances and the resulting WER, the cumulative distribution in Figure 3 reveals a macroscopic physical truth. By smoothing the stochastic interference patterns of specific wall reflections, we observe a clear, monotonic decay in model performance as the acoustic path length increases.

Crucially, the high degree of correlation between the attacked (solid) and clean (dashed) curves indicates that the room's natural reverberation acts as a barrier to model performance that is missed evaluating within a strictly digital context. Channel distortion and reverberations provide a level of added difficulty that the attacker is not bypassing, but is amplifying to successfully attack the model.

That the efficacy of the perturbation is fundamentally linked to the degradation of the carrier signal's quality (due to the influence of distance, and increased scope for reverberation based impacts) underscores the necessity of our room-aware modeling, not just for adversarial research, but for acoustic works more broadly.

Controlling for the SNR at source and victim in Figures 8 and 9 presents a similar, but more complex picture, to that of Figure 1: the Naive FGSM typically outperforms all other approaches - producing higher WER's for the same perceived SNR. The exception to this is wav2vec, which appears to produce more consistent gradients, allowing the attack to navigate around the spectral nulls inherent to the acoustic environment.

C Extensions

We emphasize that our approach focuses upon establishing the metrological basis for rigorous, acoustically-aware explorations of adversarial risk in physically relevant environment. Prior to this, the field of audio AML was caught in a scale trap, where evaluations were either too fast to be physically relevant (in digital-only attacks), or too slow to be statistically rigorous (through physical OTA attack evaluations). By introducing a high-throughput, acoustically-aligned simulation framework, we provide the foundations required to consider more complex physical conditions. We now present options for future work, to expand upon these foundations.

Directivity In an attempt to minimize the impact of observed variance, our simulations strictly considered speakers and microphones with linear responses. However, directional speakers and microphones have the potential to impact upon perceptions of the relative SNR, and the information available at each location. This could theoretically allow a sophisticated attacker to lower SNR_s while maintaining a high SNR_v.

While this could be considered to invalidate the Dual-Form SNR as a fixed environment property, the observation itself is only possible through our framework, which formally separates SNR at different locations. Furthermore, because our framework is modular, future researchers can consider directional radiation patterns directly into the RIR simulation framework we have provided, in order to provide an assessment of stealth under sophisticated attack conditions.

Natural Perturbations Similarly, the presence of non-stationary noise—of the kind seen through background chatter, traffic, or from other similar sources—may influence the Knowledge Gradient for the attacker. However, by providing an acoustically-aligned environment that is stable, repeatable, and scalable, this work enables the creation of such baseline scenarios, enabling researchers to quickly consider the layering of stochastic, real-world noise distributions on top of deterministic RIRs to study systemic robustness at scale, a task that would be physically impossible and digitally meaningless without the framework's architecture.

Scaling defenses Our defensive evaluation is limited to 8-bit quantization, yet the Dual-Form SNR suggests a natural class of spatial defenses. Because the source and attacker occupy distinct positions, multi-microphone arrays can exploit differences in direction of arrival—for example, via beamforming—to attenuate off-axis interference before transcription. This inverts the Knowledge Gradient, as the defender may knows their own array geometry and may measure their own room, holding an informational advantage over the attacker. Conversely, the Projection Cost implies that attacks strong enough to overcome such spatial filtering will be energetically more visible at the source. Because PyRoomAcoustics natively supports microphone arrays and beamforming, our framework can evaluate such acoustically-aware defenses at scale; we leave this to future work.

D Computational Process

An aim of this work is to provide a robust and generalizable framework for evaluating adversarial attacks against ASR systems, the approach for which is outlined within Algorithm 1. Within this work, we performed evaluations against the LibriSpeech `test-clean` English corpora (Panayotov et al., 2015), where all audio was normalized to 16 kHz mono signals, and filtered to ensure that no sample ran longer than 15 seconds. We tested against Whisper (Tiny & Base variants, MIT License, Radford et al. (2023)) and Wav2Vec (Base variant, MIT License, Baevski et al. (2020)). Of these, Whisper is a sequence-to-sequence transformer model, while Wav2Vec Connectionist Temporal Classification (CTC)-based model.

For our tested attacks, PGD was run over 40 iterations, with $\alpha = 0.01$ and random starting, while FGSM was a single-step gradient sign update. The tested Carlini & Wagner style attack employed ℓ_2-loss optimization with 5 binary search steps for the trade-off constant c and 100 iterations per step, with the learning rate set to 0.01.

As was discussed in the main body of the work, the tested room simulations are performed over thousands of randomized shoebox rooms whose lengths and widths are uniformly distributed over $[3, 10]$m, and with a height between $[2.5, 3.5]$m. While our framework can be extended beyond

Algorithm 1 Decoupled RIR-Aware Adversarial Attack. Prior to running this code, the set of h is pre-computed on CPU and converted to GPU tensors. The Conv1D operator allows for gradients ∇_δ to flow through the network.

Input: Source audio x, Target text T, Model f, Knowledge Level $\mathcal{K}$
Output: Adversarial Example x_{adv}

{Phase 1: Offline RIR Pre-computation (CPU)}
if $\mathcal{K} = \mathcal{K}_{oracle}$ **then**
  $H_{batch} \leftarrow \{h_{true}\}$
else if $\mathcal{K} = \mathcal{K}_{blind}$ **then**
  $H_{batch} \leftarrow$ Sample $\{h_1, \ldots, h_N\} \sim \mathcal{D}_{room}$
end if
Convert H_{batch} to GPU Tensors
Sample and convert H_{batch}, comprising the room RIR's h to GPU Tensors

{Phase 2: Online Optimization (GPU)}
$x_t \leftarrow x$ {Initialize $x_0 = x$}

for step $t = 0$ **to** $K - 1$ **do**
  {Calculate gradient to maximize loss relative to original text T}
  $g \leftarrow \nabla_{x_t} \mathcal{L}_{\text{ASR}}(f(x_t), T)$

  {Apply PGD update rule: $x_t + \alpha \cdot \text{sign}(g)$}
  $x_{t+1} \leftarrow x_t + \alpha \cdot \text{sign}(g)$

  {Projection operator $\Pi_{x+\mathcal{S}}$: Clip to ϵ-ball and valid data range}
  $x_{t+1} \leftarrow \text{Clip}(x_{t+1}, x - \epsilon, x + \epsilon)$
  $x_{t+1} \leftarrow \text{Clip}(x_{t+1}, -1, 1)$
end for

$x_{adv} \leftarrow x_K$

shoebox room geometries, we deliberately restricted ourselves to such a modeling framework for two key criteria. The first relates to validation: RIR simulations are most strongly validated within the shoebox geometry context. The second is that non-uniform geometries present significant challenges with regards to identifying the characteristic lengths of acoustic signals, in order to appropriately contextualize results. With both of these points made, we stress that it is still possible for our framework to be extended to more complex geometries, however we leave solving these associated problems for future work.

Performance The high-throughput performance of the system described within this work is a key part of being able to reach 8 million adversarial attacks. Table 7 demonstrates that the workload is heavily skewed toward CPU-bound signal processing (PESQ/STOI) rather than GPU-bound adversarial optimization. Under optimal conditions, the total cost is estimated at 11.75 GPU-Days. However, large-scale deployment on shared clusters resulted in an observed runtime of $\sim$100 GPU-Days. This $8.5\times$ discrepancy highlights a synchronization penalty where GPUs are starved while waiting for CPU-bound metrics.

To again reiterate the implications of digital simulation, consider the time associated with just playing 1.28 million samples directly within a single acoustic environment. Even if each sample was capped at 10 seconds of length, this would require 925 days of active time. That we are able to reach these scales within our acoustically-aligned digital simulation framework represents a significant step-change in the ability to approach physically realistic scenarios for evaluations of not just acoustic AML problems, but of the broader class of ASR inspired problems.

Table 7: Computational cost breakdown for the full evaluation suite. The processing of quality metrics constitutes the majority of the theoretical runtime. Costs incurred on a system with 64GB of ram and a 80GB A100 GPU.

COMPONENT	COUNT	TIME	LOAD
ATTACKS (GEN)	1.28M	2.45 DAYS	20.9%
TRANSCRIPTION	10.2M	1.35 DAYS	11.5%
METRICS (CPU)	40.9M	7.95 DAYS	67.7%
TOTAL	**-**	**11.75 DAYS**	**100%**

E DEFENSES

A common defensive methodology is to consider the impact of a pre-model quantization phase. Quantization inherently reduces the information content of the signal, and thus has been posited to serve as a viable defense against adversarial manipulation. However, as can be seen in Table 8, there is a consistent increase (on the order of 10 $\%$) in the observed WER. We posit that this likely stems from the decreased acoustic energy available at the victim receiver after factoring in h_{SV}, leading to the clean signal occupying only a fraction of the dynamic range. Thus quantization at this point has a disproportionate effect on the signal output, relative to past works that have strictly considered quantization within the digital domain. Counterintuitive to expectations, quantization acts as what is essentially a second broadband noise source, that increases the model's confusion.

These results are reinforced by the differences in performance between wav2vec's CTC model, relative to Whispers Transformer architecture. We believe that quantization introduces variations that elicit a collapse in CTC alignment. By contrast the Transformer's attention mechanism allows for aggregation of information across longer context windows, increasing the models resilience to the impacts of quantization.

Table 8: Word Error Rates when considering both no defence, and 8-bit quantization as a defence.

	DEFENSE	NO DEFENSE	8-BIT QUANT	REL. CHANGE (%)
WHISPER TINY	NAIVE FGSM	0.25	0.278	+11.1%
	NAIVE PGD	0.221	0.257	+16.1%
	RIR PGD	0.25	0.28	+11.9%
WHISPER BASE	NAIVE FGSM	0.168	0.184	+9.6%
	NAIVE PGD	0.144	0.166	+15.5%
	RIR PGD	0.168	0.185	+10.5%
WAV2VEC2 BASE	NAIVE FGSM	0.285	0.371	+30.2%
	NAIVE PGD	0.274	0.377	+37.3%
	RIR PGD	0.337	0.413	+22.6%
WHISPER TINY $\rightarrow$ BASE (TRANSFER)	NAIVE FGSM	0.17	0.187	+10.1%
	NAIVE PGD	0.146	0.169	+15.8%
	RIR PGD	0.17	0.188	+10.5%

F ACOUSTIC MEASURES OF PERCEPTUAL QUALITY

It is important to note that while the performances of these attacks can be similar, there is a definite decoupling in terms of their perceptibility, with Figure 5 demonstrating that Naive PGD consistently produces the largest PESQ and STOI scores – indicating a small degree of added stealth, relative to the other attacks. While the RIR PGD and Naive FGSM are almost equivalent for STOI, RIR PGD is relatively favoured for PESQ, and equals Naive PGD at higher Victim SNR levels. These results suggest that the RIR-aware attacks may appear to be slightly more natural than their FGSM equivalents, although these results must be tempered by our earlier discussions relating to the pitfalls of these perceptual measures.

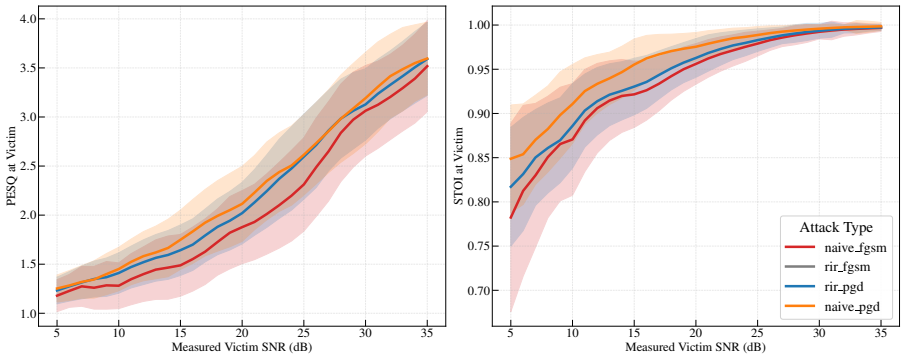

Figure 5: Perceptual measures of quality, broken down by attack type.

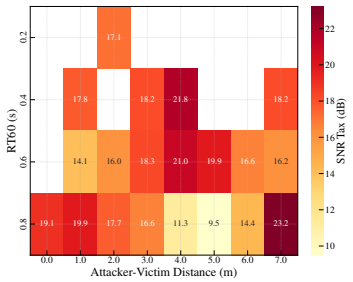

Figure 6: Acoustic Tax Heatmap, covering the SNR differential ($\text{SNR}_{\text{source}} - SNR_{\text{victim}}$) in terms of both RT_{60} and Attacker-Victim distance, revealing how the energetic signature of an attack is non-uniformly distributed across the acoustic space.

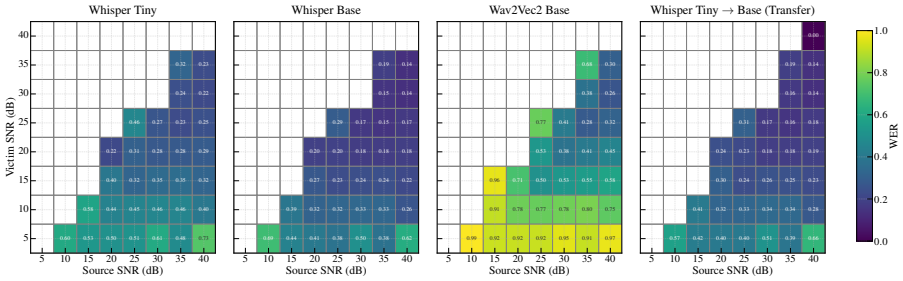

Figure 7: Relationship between the dual-SNR metrics and WER.

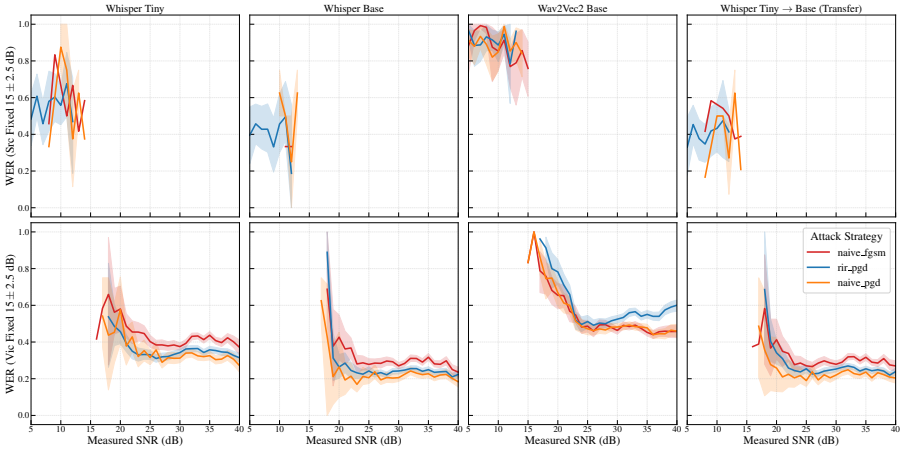

Figure 8: WER when the SNR is fixed at 15 ± 2.5 at either the source (top) or the victim (bottom).

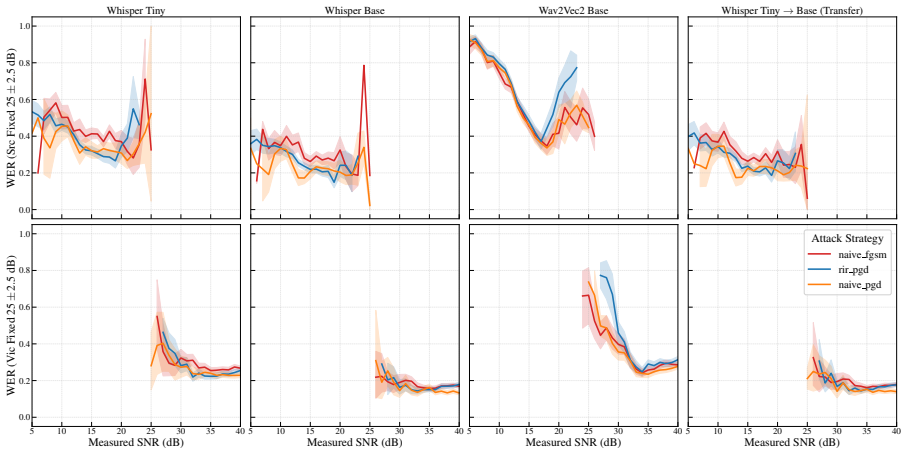

Figure 9: WER when the SNR is fixed at 25 ± 2.5 at either the source (top) or the victim (bottom).

G DEFINING c

Our aim with c in Equation 4 was not to define a hard constraint, that exactly mapped to a particlar source SNR. Rather, our approach was to seed our experimental set with a range of source SNRs, that roughly existed within a range of values. To achieve this, we began by sampling a set of 10 random acoustic environments. Each of these rooms were exposed to 100 samples at varying c, creating a mapping between the source SNR and c. We then selected a range of c values associated with SNRs of $\{15, 25, 35, 45\}$, and used these values to guide our attacks.

We again emphasize that the exact relationship between c and the SNR is not necessary for our analysis, and would, in fact, produce unfavorable outcomes. Our approach involves exploring a large range of acoustic environments, each of which would have their own unique mapping between these two parameters. Thus a single value of c produces a range of SNR values at the source, and, commensurately, at the victim as well. Thus our choice to use a set of c values was a deliberate one, to ensure that our environments demonstrated a sufficient diversity of SNR values.