Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hiding a Voice From Speaker Encoders: Adversarial Protection Trained Over Measured Room Channels

Bhavishay Sheoran
Indian Institute of Technology Roorkee https://orcid.org/0009-0000-5298-5217
Mayank Goswami
mayankfph@iitr.ac.in
Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Research Article

Keywords: speaker verification, adversarial examples, room impulse response, expectation over transformation, voice privacy, physical adversarial attacks

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11671474/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Hiding a Voice From Speaker Encoders: Adversarial Protection Trained Over Measured Room Channels

Evaluated Through Channel Models, Human Listeners and a Real Air Gap

Bhavishay Sheoran [1] and Mayank Goswami, [1,2]

[1]Department of Physics, IIT Roorkee
[2]MF School of AI and DS, IIT Roorkee
mayank.goswami@ph.iitr.ac.in

***Abstract*—Adversarial perturbations can stop a speaker encoder from extracting a talker's identity. The tools are standard—projected gradient descent, Expectation over Transformation (EOT) and room impulse responses—but the closest protection methods optimise without any room response, so it is open whether the channel model used in optimisation matters. We contribute an experimental protocol and its findings. We measured four acoustic channels by exponential sine sweep with consumer devices, trained the protection over three and tested it on the fourth, and judged it with an independently trained encoder and a threshold set under the evaluation conditions. At $\varepsilon = 0.004$ (SNR 23.9 dB, PESQ 2.07, ESTOI 0.86) the independent encoder's identification rate falls from 98% to 28% (20 speakers × 4 held-out channels, 5 runs). Against simulated channels matched in number, reverberation time and post-processing, training on measured channels lowers the judge's similarity by a further 0.045 (95% CI 0.036–0.054) at equal budget and 0.039 at equal SNR. A control shows that about 80% of this advantage comes from a spectral colouring shared by all measured responses, including the held-out test channel, much of it made by our deconvolution: with it added, simulated responses come within 0.007, and a larger simulated set lowers the similarity further than the measured ones. In a feasibility test with one clip re-recorded in four rooms, free of this processing, the protection lowered the judge's similarity in every room, but less than through the channel models, and the adversary still succeeded in three of the four. In a controlled listening test with 28 listeners, the protected speech was rated 1.2 points lower in quality (4.15 to 2.97 on a five-point scale) and judged the same talker less often (76% to 60%), with no detectable change in the share of words understood.**

***Index Terms*—speaker verification, adversarial examples, room impulse response, expectation over transformation, voice privacy, physical adversarial attacks**

I. Introduction

A voice-cloning system needs only a few seconds of speech. A speaker encoder reduces that audio to a fixed-length embedding capturing vocal identity, and a synthesiser then generates arbitrary new utterances in that voice. The defensive response studied here is to add a perturbation to speech *before* it is released, chosen so that a human still hears the original talker while the encoder's embedding moves away from the true identity.

Most evaluations of this defence are digital: perturbations are computed on a waveform file, applied to that file, and evaluated on that file. In the closest protection methods we compare (Table I) the perturbation is optimised without any room response, and two of them report physical tests in which the protection is played through a loudspeaker and recorded. Robustness to the room has been pursued much further in attacks on speech and speaker recognition, with simulated [10], [12] and measured [11] impulse responses. But the threat that motivates the defence is physical—a voice is spoken aloud, crosses a room and is captured by a microphone—and for protection against speaker encoders a basic design question has not, in the works we compare, been tested under matched conditions: *which* channel model should the perturbation be optimised over?

What is and is not new. Every component here is standard: projected gradient descent [15], Expectation over Transformation [9], x-vector-style speaker encoders [14] and impulse responses measured by exponential sine sweep [16]. The contribution is an experimental protocol and what it found.

Contributions.

1) *A protocol* for evaluating voice protection over real channels: four channels measured with consumer devices, leave-one-channel-out training and testing, an independently trained judge never used in optimisation, a decision threshold derived under the evaluation conditions, channel-matched enrolment, comparisons at equal budget and at equal SNR, and repeated optimisation runs (Sections V-C and VI).
2) *A matched comparison of channel models, and a control.* Against simulated channels matched in number, reverberation time and post-processing, train-

Corresponding author: mayank.goswami@ph.iitr.ac.in

ing over measured channels lowers the independent judge's similarity by 0.045 more, in all four held-out rooms and all five runs (Section VII-B). A control shows that about 80% of this advantage comes from a spectral colouring shared by the measured training and test channels, much of it made by our deconvolution, a step the simulated responses do not go through (Section VII-C).

3) *Ablations.* Removing any step of the measurement processing from the training responses weakens the protection; because the test channel keeps every step, this is consistent with matching its processing (Section VII-D).
4) *The operating point.* Through held-out measured channels the protection cuts an independent adversary's identification rate from 98% to 28%, with identity, quality (PESQ) and intelligibility (STOI, ESTOI) reported separately (Section VII-G).
5) *A physical feasibility test.* One protected clip, played and re-recorded in four rooms, lowers the independent judge's similarity in every room but by less than through the channel models; we measure a gap between modelled and real recordings that may explain this, and show that extending EOT to the non-linear stages of the chain helps only slightly (Sections VII-H and VII-I).
6) *Listening tests.* An uncontrolled test with 22 listeners and a controlled test, with its analysis fixed before data collection, with 28 listeners, which shows that the protection is clearly audible (Section VII-K).
7) *Negative results and corrections.* Four attempts to improve transfer that did not help (Section VII-J), and ten methodological errors made during the project, with how each was corrected or accounted for (Section VIII).

Constraints. A course requirement forbade pretrained models. Every network here was trained from scratch and all code was written for this project. The constraint shapes the results and we return to it in Section IX.

II. Related Work

Protection against cloning and recognition. Table I summarises the closest protection methods as their papers describe them. AttackVC [1] and Voice Guard [4] perturb speech so that voice-conversion models fail to reproduce the speaker; Voice Guard keeps its perturbation below a psychoacoustic masking threshold. AntiFake [3] optimises against an ensemble of speaker encoders and steers the embedding toward a dissimilar target. SafeSpeech [6] and SpeechShield [7] target the training or fine-tuning of synthesisers; SpeechShield couples the perturbation to timbre through harmonic-energy and spectral-flatness losses. VoiceBlock [2] and V-Cloak [5] act on speaker recognition directly and in real time. None of these optimises over a room response. SafeSpeech and SpeechShield report physical tests—protective noise played live and recorded at 50 cm, and protected audio played and recorded with laptop and external devices at 30–45 cm—without modelling the room. From the attacker's side, De-AntiFake [8] shows that purification can remove much of a protective perturbation; we do not test purification (Section IX).

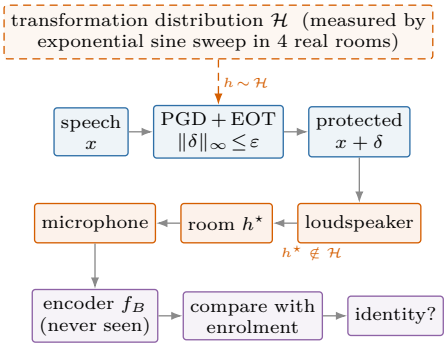


Fig. 1. Threat model. Blue: digital, under the talker's control. Orange: the physical path. Purple: the adversary, whose encoder is never used during optimisation. The dashed box is the transformation distribution; its contents are the subject of this paper. The orange path also contains stages that no impulse response describes—loudspeaker saturation, microphone gain control, clock drift and lossy coding. Section VII-H suggests that these stages and the colouring added by our measurement processing (C10), rather than reverberation, separate the modelled channel from the physical one.

Robustness to the acoustic channel. Expectation over Transformation [9] is the standard device for making adversarial examples survive a physical channel: the gradient is averaged over a distribution of transformations, so the perturbation is optimised for a family of channels rather than one realisation. For speech recognition, Qin et al. [10] optimised over simulated environmental distortions, and Metamorph [11] used 370 *measured* channel impulse responses from four public datasets and demonstrated attacks over the air in an office building, noting that such datasets cannot cover all channel conditions. For speaker recognition, Xie et al. [12] trained universal perturbations over impulse responses simulated for one shoebox room, without a physical test. These attacks try to be *recognised* as someone; protection tries *not* to be recognised, but the channel problem is the same.

Where this paper sits. Measured impulse responses and over-the-air tests are therefore not new in themselves. What we did not find among the protection methods above is a controlled test of whether the channel model used in optimisation matters: measured against simulated responses with the number of channels, the post-processing, the reverberation time, the budget and the SNR matched, evaluated through a held-out measured channel by a judge never used in optimisation. That comparison, a control for the colouring that the measured channels share, its ablations, and a physical feasibility check are the content of this paper.

III. Threat Model and Problem Statement

Actors. A *talker* releases speech. An *adversary* obtains a recording and wishes to decide whether it belongs to

TABLE I

Closest published protection methods, as described in their papers. Room model: whether any room or channel response is used while the perturbation is optimised. Re-recorded: whether the protection was played through a loudspeaker and recorded again. Unseen model: whether the protection was evaluated on a model not used in optimisation. Listeners: whether people judged the protected audio.

Work	Protects against	Room model	Re-recorded	Unseen model	Listeners
AttackVC [1], SLT'21	voice conversion (two zero-shot models)	none	no	proxy-trained models (black-box)	yes (same-speaker judgements)
VoiceBlock [2], NeurIPS'22	speaker recognition, real time	none (clean speech, no augmentation)	no ("over-the-line")	one unseen model	yes (crowdsourced, MUSHRA-style)
AntiFake [3], CCS'23	speech synthesis (encoder ensemble)	none	no	yes, incl. a commercial synthesiser	yes (24; usability and quality)
Voice Guard [4], IJCAI'23	voice conversion	none	no	across two VC models	no (predicted MOS)
V-Cloak [5], USENIX Sec.'23	speaker recognition (anonymisation), real time	none (RIRs named as future work)	no	yes, four further verifiers incl. a commercial one	yes (102)
SafeSpeech [6], USENIX Sec.'25	TTS fine-tuning and zero-shot cloning	none	yes: protective noise played live, recorded at 50 cm	yes, unseen zero-shot models	yes (80)
SpeechShield [7], DSN'26	TTS training	none	yes: laptop and external loudspeakers and microphones, 30–45 cm	across TTS architectures	no (objective metrics)
This work	speaker encoders (identity extraction)	**measured** responses of 3 rooms, the 4th held out; matched simulated controls	yes: one clip in four rooms (feasibility)	yes, an independent encoder never used in optimisation	yes (22, uncontrolled); controlled test (28)

a particular enrolled identity—the step on which any downstream cloning depends.

Adversary knowledge. The adversary is *black-box* with respect to the protection: the encoder used to construct the perturbation (model A) is not the encoder used to evaluate it (model B), which is independently trained with a different architecture, parameter count and seed. White-box numbers against model A are reported for completeness only; an attack evaluated on the network it was optimised against measures nothing about generalisation.

Adversary enrolment. We assume an adversary who knows recordings are reverberant and enrols accordingly. In the computational experiments the enrolment embedding is the normalised mean of the victim's enrolment utterance and that utterance convolved with each *training* channel. In the physical experiment (Section VII-H) it is the mean over actual re-recordings of the victim made in the other rooms. Both are strictly more generous to the adversary than enrolling on pristine digital audio; Section VIII explains why that generosity is necessary.

Objective. Given speech x, find δ with $\|\delta\|_\infty \leq \varepsilon$ such that for an unseen channel and an unseen encoder f_B, the cosine similarity to the victim's enrolment falls below the encoder's decision threshold, while $x+\delta$ remains acceptable to a human.

IV. Speaker Encoders Trained From Scratch

Corpus and split. LibriSpeech `train-clean-100` [13]: 28,539 utterances, 251 speakers, 16 kHz. The split is *speaker-disjoint*—speakers shuffled with a fixed seed and divided 200/51. A clip-level split would let the same voice appear on both sides and would inflate every verification number here.

Front end. A 400-sample (25 ms) Hann window at 160-sample (10 ms) hop, 512-point FFT, 80-band mel filterbank (HTK), log compression. The Hann window is used rather than a rectangular one because its side lobes fall at about −31 dB against −13 dB; rectangular leakage would place energy in mel bands where the signal has none.

Architectures. All three encoders share a pattern: a 2-D convolutional trunk over the log-mel spectrogram, statistics pooling (concatenated mean and standard deviation over time, after [14]), a linear projection to an embedding, and a classification head used only in training. After training the head is discarded and identity is compared by cosine similarity between ℓ_2-normalised embeddings. The three differ deliberately:

- **A** (attacked): four blocks, 3×3 kernels, 192-dim, 0.53M parameters; EER 0.92% on clean held-out speakers.
- **B** (independent judge): five blocks beginning with a 5×5 kernel, an extra time-strided block, 256-dim, 1.65M parameters, different seed. *Never used in any optimisation.*
- **C** (ensemble candidate): five blocks with dilated convolutions, 320-dim, different seed; EER 8.50%.

Training is cross-entropy over the 200 training speakers with Adam.

V. Protection Method

A. Projected gradient descent

We use projected gradient descent [15]. With $e_0 = f(x)/\|f(x)\|$ the embedding of the unprotected signal,

$$\min_{\|\delta\|_\infty \le \varepsilon} \left\langle \frac{f(x+\delta)}{\|f(x+\delta)\|}, e_0 \right\rangle, \tag{1}$$

updated by the sign of the gradient,

$$\delta \leftarrow \Pi_{\|\cdot\|_\infty \le \varepsilon}\big(\delta - \alpha\,\mathrm{sign}(\nabla_\delta L)\big), \qquad \alpha = \varepsilon/10, \tag{2}$$

for $K = 300$ iterations. The sign is not a heuristic: under an ℓ_∞ constraint it is the exact maximiser of the linearised objective, by Hölder duality between ℓ_∞ and ℓ_1.

B. Expectation over transformation

A perturbation optimised for the clean signal is optimised for one point in signal space; a channel moves the signal elsewhere. Following [9] we optimise the expected objective over a distribution of channels,

$$\min_{\|\delta\|_\infty \le \varepsilon} \mathbb{E}_{h\sim\mathcal{H}} \left[\left\langle \frac{f(T_h(x+\delta))}{\|f(T_h(x+\delta))\|}, e_0 \right\rangle \right], \tag{3}$$

estimated with $B = 4$ samples per iteration. The question this paper studies is what T_h should be. We take it first to be convolution with a *measured* impulse response (Section V-C), compare that with simulated responses under matched conditions (Section VII-B), and then, in Section VII-I, extend it toward the full physical chain. After convolution the channel output is rescaled to the peak amplitude of its input, so that channels differ in colour and decay rather than in gain (*output-level normalisation*).

C. Measuring the channels

Excitation. An exponential sine sweep from 20 Hz to 8 kHz over 10 s, with 1 s lead-in and 2 s tail, played from a laptop loudspeaker and captured by a smartphone held fixed. The exponential sweep is preferred to noise or an impulse because its harmonic distortion products appear at negative time in the deconvolved response and can be separated from the linear part [16].

Deconvolution. With Y the recorded and X the reference spectrum,

$$H(f) = \frac{Y(f)\,X^*(f)}{|X(f)|^2 + \lambda(f)}, \tag{4}$$

with λ *frequency dependent*. Some regularisation is essential: below 20 Hz the sweep has almost no energy, so the denominator approaches zero, and below 150–250 Hz a laptop loudspeaker radiates almost nothing, so the recording there contains only room noise; a naive inverse turns both into large spurious low-frequency components (Section VIII, C3). We set $\lambda(f) = 10^{-2}\max_f |X(f)|^2\,[1 + 10^4(1-\beta(f))]$ and multiply H by β, the smooth band-pass weight defined below; so λ is 10^{-2} of the peak power inside 250 Hz–7 kHz and rises toward the band edges to four orders of magnitude more. This choice colours the band (C10). The power of an exponential sweep falls as $1/f$, so at 5 kHz $|X|^2$ is about 300 times below its peak and smaller than λ, and the regularisation attenuates high frequencies; the rise of λ toward the band edges reaches well into the band. Every measured response carries the resulting colouring.

Out-of-band handling. A filter that is zero below 250 Hz does not describe a room—it describes a high-pass filter, and convolving speech with it removes the fundamental and first formant. We therefore blend the measured response in-band with unit gain out-of-band,

$$H_{\text{final}} = \beta(f)\,H_{\text{meas}}(f) + \big(1-\beta(f)\big)\,e^{-j2\pi f t_0}, \tag{5}$$

β a smooth band-pass weight and the second term a unit impulse at the direct-path delay. The measured response is normalised to unit direct-path amplitude *before* the blend so the two terms are commensurate. The responses are kept to 1 s.

Reverberation time. RT_{60} by Schroeder backward integration [17], fitted between -5 and -25 dB and extrapolated ($T_{20} \times 3$), with the decay curve truncated at the noise floor.

VI. Experimental Setup

Rooms. Four channels: two microphone positions in a large room (centre and corner), a normal-sized room, and a stairwell. Within a room the loudspeaker and microphone stay fixed, so one impulse response describes the geometry for every clip recorded there. The same laptop loudspeaker and smartphone are used in every room.

Protocol. Leave-one-out throughout: the perturbation is optimised by EOT over three channels and evaluated on the fourth, measured one. The main evaluation uses 20 held-out speakers, one utterance each, so every condition has $n = 80$ speaker×room trials.

Decision threshold. A threshold is meaningful only under the conditions in which it was measured. We compute the equal-error threshold for each judge under the evaluation conditions—reverberant test audio, channel-matched enrolment—using 30 held-out speakers across the 4 rooms ($n = 120$ genuine and 120 impostor trials per judge):

Judge	EER	τ	genuine / impostor
A (white-box)	18.33%	0.734	0.797 ± 0.089 / 0.651 ± 0.098
B (independent)	16.25%	0.793	0.839 ± 0.056 / 0.700 ± 0.086

These are far from model A's clean-speech threshold of 0.463 at 0.92% EER: reverberation and channel-matched enrolment degrade both verifiers substantially. Using the clean threshold would be a serious error, and we made it before correcting it (C2).

Equal budget and equal SNR. Comparisons are made two ways. At *equal budget* every condition uses the same ε. At *equal SNR*, in each trial, every perturbation is afterwards scaled down to the SNR of the quietest of the ten protected conditions (median 25.4 dB at $\varepsilon = 0.004$ and 31.0 dB at $\varepsilon = 0.002$); scaling down never violates the ε constraint, and no condition can then win by being louder.

TABLE II
MEASURED ACOUSTIC CHANNELS. EARLY ENERGY IS THE FRACTION OF IMPULSE RESPONSE ENERGY ARRIVING WITHIN 50 MS.

Channel	RT_{60} (s)	Early energy	real–sim
Normal room	0.42	97.6%	0.364
Large room, centre	1.05	93.4%	0.323
Large room, corner	1.81	95.4%	0.296
Stairwell	1.58	76.5%	0.389

Repeated runs. The only randomness in the optimisation is which training channels are drawn at each step (and, for one simulated condition, which simulated responses are chosen). Every condition of Sections VII-B, VII-D and VII-G was run five times at $\varepsilon = 0.004$ and three times at $\varepsilon = 0.002$, each run with its own fixed seed.

Statistics. For each condition every speaker×room trial is averaged over runs, and conditions are compared by paired differences over the 80 trials: the mean with a t-based 95% confidence interval (for $\cos_B$, bootstrap intervals agree to within 0.001), paired t and Wilcoxon tests, and Holm correction within each comparison table. The four trials of one speaker share a voice and are not independent, so every comparison is also made at the speaker level: a t-test over the 20 speaker means and a speaker-cluster bootstrap that resamples speakers with all their trials (10,000 resamples, percentile interval). We state every case in which the speaker level disagrees with the trial level. The rule for reading the comparison was fixed before the run: a statement that one condition protects better is made only if the 95% interval of the paired difference excludes zero at both equal budget and equal SNR.

Perceptual measures. SNR of the perturbation, $20\log_{10}$ of the ratio of the standard deviations of speech and perturbation; wideband PESQ [19] for quality; STOI [20] and ESTOI [21] for intelligibility. All three are signal-processing measures, not trained models. Because ε bounds the perturbation in absolute amplitude, the resulting SNR depends on the loudness of the source clip. The 80-trial evaluation uses each utterance at its recorded level; the clip re-recorded in the rooms and the listening-test clips were level-normalised before protection (C7).

VII. RESULTS

A. Channel characterisation

Table II and Fig. 2 characterise the channels. The stairwell is distinguished not by decay time—the large room's corner is longer—but by its direct-to-reverberant ratio: only 76.5% of its energy arrives in the first 50 ms against 93–98% for the rooms. That is the signature of a tall, hard-surfaced space, and it makes the stairwell the most demanding held-out channel.

The *real–sim* column anticipates the physical result. It compares the embedding of a clip convolved with the measured response against the embedding of the same clip physically recorded in that room. Values of 0.30–0.39 say that the measured channel reproduces the direction of the real degradation but not its character. We return to this in Section VII-H.

B. Measured versus simulated channels, matched

Our first comparison of measured and simulated training channels differed in more than the channel model (C8): 200 simulated responses against 3 measured ones, truncated to 0.5 s against 1 s, normalised to their peak rather than to the direct path, without the in-band treatment of Section V-C, and on a single clip. We therefore repeated it with everything but the training channels held fixed. All conditions protect the same 20 speakers, are evaluated through the same held-out measured channel, and use the same ε, K, α and B:

- *Sim., 200 raw, 0.5 s*: the simulated setting of the earlier run—200 responses from OpenSLR-28 [18], all from four virtual large rooms (median RT_{60} 1.34 s), truncated to 0.5 s and peak-normalised.
- *Sim., same 200, processed*: the same responses given every step of the measured responses' processing except the regularised division itself, which a simulated response does not need: band weighting, normalisation, out-of-band blend and 1-s length (the *post-processing*).
- *Sim., 3 rooms, RT_{60}-matched*: for each training room, one simulated response drawn at random from the five whose RT_{60} is closest to that room's, among 3000 scanned across small, medium and large simulated rooms; redrawn in every run and post-processed in the same way.
- *Measured, 3 rooms*: the protocol of this paper.

Table III and Fig. 3 give the result. In the key comparison—same number of channels, same post-processing, matched reverberation time—training over measured channels lowers $\cos_B$ by 0.045 more than training over simulated ones (95% CI 0.036–0.054) at equal budget and by 0.039 (0.031–0.048) at equal SNR. The direction holds in all four held-out rooms (per room 0.039–0.049) and in each of the five runs (0.043–0.047), and at the speaker level ($n = 20$) $p = 4.0 \times 10^{-5}$, with speaker-cluster bootstrap intervals of 0.030–0.062 and 0.024–0.055. At $\varepsilon = 0.002$ the same comparison gives 0.024 (0.017–0.030) and 0.021 (0.015–0.027). Without the speaker of C9, whose equal-SNR trials are unprotected in every condition, the equal-SNR difference at $\varepsilon = 0.004$ is 0.041 (0.032–0.050). Under the rule fixed in advance, the statement that measured channels protect better than these simulated ones stands; Section VII-C shows where most of the difference comes from.

Two other rows qualify it. First, a larger, differently composed simulated set—200 responses from four large virtual rooms, post-processed like ours—protects better than three RT_{60}-matched ones ($\Delta = 0.023$); size and composition are confounded in this comparison. The margin of measured channels over that set is smaller, 0.022 (0.014–0.030) at equal budget and 0.017 (0.009–0.024) at equal SNR, so it closes about half of the key gap at $\varepsilon = 0.004$.

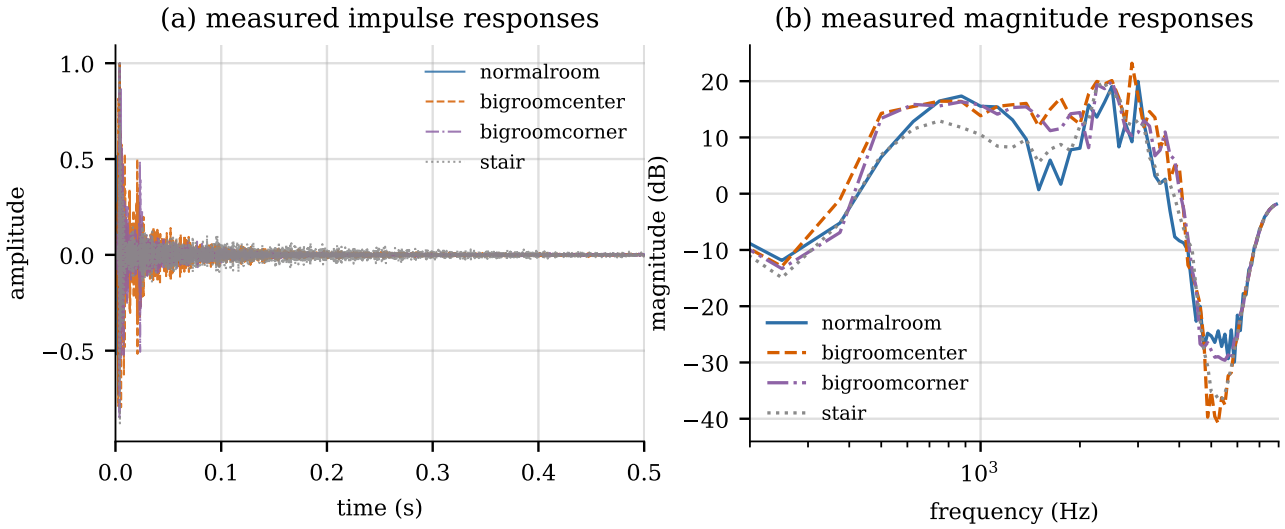


Fig. 2. The four measured channels. (a) impulse responses; (b) magnitude responses. The curves share a common gross shape: a fall below about 500 Hz and a deep dip between 5 and 6 kHz. Our processing produces both by itself: applied to a perfectly flat channel it returns −30 dB at 350 Hz and −37 dB at 5.5 kHz relative to 1 kHz (C10). The loudspeaker and microphone common to all four measurements may add to the shape; the rooms differ in the fine structure and in the decay of (a). Every held-out test channel carries the same shape (Section VII-C).

TABLE III

Matched comparison and ablations. Mean $\cos_B$ (independent judge, lower is better) and the share of trials in which the attacker succeeds ($\cos_B > \tau_B$, %), over 20 held-out speakers × 4 held-out measured channels ($n = 80$), averaged over 5 runs at $\varepsilon = 0.004$ and 3 at $\varepsilon = 0.002$. Across runs the SD is at most 0.0012 for $\cos_B$ and 2.6 points for the success rate. Equal SNR: in each trial every perturbation scaled to the quietest of the ten (median 25.4 dB at $\varepsilon = 0.004$, 31.0 dB at $\varepsilon = 0.002$). Ablations change only the three training channels; testing is always through the fully processed held-out channel. †For one speaker the PGD perturbation stayed zero (C9). This inflates PGD's SNR, PESQ and ESTOI (without that speaker, at $\varepsilon = 0.004$: 24.7 dB, 2.22, 0.86) and leaves that speaker's four trials unprotected in every condition at equal SNR.

	$\varepsilon = 0.004$							$\varepsilon = 0.002$			
	equal budget					equal SNR		equal budget		equal SNR	
Training channels	$\cos_B$	succ.	SNR	PESQ	ESTOI	$\cos_B$	succ.	$\cos_B$	succ.	$\cos_B$	succ.
Unprotected	0.875	97.5	–	–	–	0.875	97.5	0.875	97.5	0.875	97.5
PGD, no channel†	0.813	65.0	31.2	2.34	0.87	0.813	65.0	0.837	81.3	0.837	81.3
Sim., 200 raw, 0.5 s (earlier run)	0.781	50.5	24.3	2.13	0.86	0.787	52.3	0.811	73.3	0.815	74.2
Sim., same 200, processed	0.769	37.8	24.2	2.08	0.86	0.776	42.5	0.805	67.5	0.810	72.5
Sim., 3 rooms, RT_{60}-matched	0.792	53.8	24.2	2.14	0.87	0.799	58.3	0.820	79.6	0.825	79.6
Measured, 3 rooms	**0.747**	**27.8**	23.9	2.07	0.86	**0.760**	**34.3**	**0.797**	**60.4**	**0.804**	**63.3**
Ablations of the measured-response processing											
no freq.-dep. regularisation	0.776	42.8	23.9	2.09	0.87	0.786	49.0	0.813	70.4	0.819	75.0
no out-of-band blend	0.757	42.0	23.7	2.10	0.86	0.771	47.0	0.809	68.3	0.816	71.7
no normalisation before blend	0.787	57.0	23.8	2.13	0.86	0.797	61.8	0.819	74.6	0.826	77.5
no output-level normalisation	0.755	29.0	24.1	2.07	0.86	0.764	34.5	0.803	62.9	0.809	66.7
no processing at all	0.799	62.3	24.0	2.11	0.86	0.806	65.5	0.827	77.9	0.832	80.0

At $\varepsilon = 0.002$ the margin shrinks to 0.008 and 0.006 (lower limit 0.0003), so two thirds or more of the gap is closed, and neither is significant at the speaker level ($p = 0.15$ and 0.28; speaker-cluster intervals from −0.002 to 0.018 and from −0.003 to 0.016). Second, post-processing helps simulated responses too: post-processing the earlier 200 in the same way improves them by 0.013 (0.009–0.017).

What this comparison does not show. Every measured response, and every held-out test channel, passed through the same laptop loudspeaker, smartphone microphone and regularised deconvolution; the simulated responses share only the post-processing (Fig. 2). Leave-one-room-out is therefore neither leave-one-device-out nor free of the deconvolution, and part of the measured advantage may come from what the test channel shares with the training channels rather than from the rooms themselves. Section VII-C measures how much.

C. A control for the colouring shared by all measured responses

The measured responses share a common spectral shape that the simulated ones lack, and the held-out test channel has it too (Fig. 2). To measure how much of the advantage of Section VII-B it explains, we estimated this shape for each held-out room from the three training responses

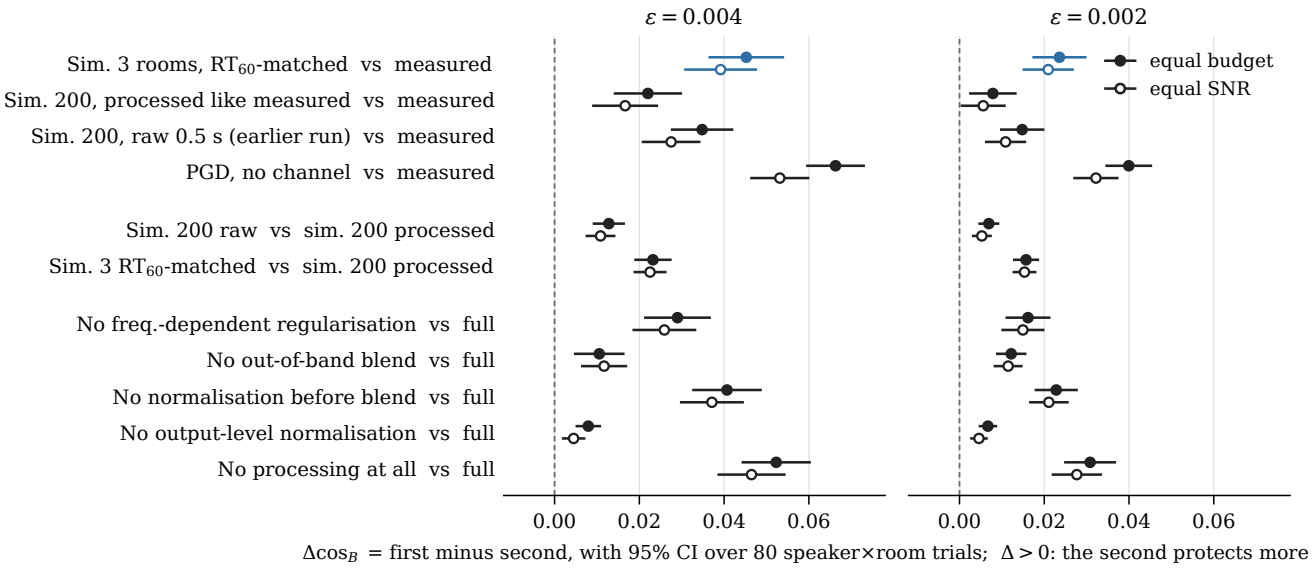


Fig. 3. Paired differences in the independent judge's similarity, with 95% confidence intervals over the 80 speaker×room trials (runs averaged). Filled: equal budget; open: equal SNR. A positive value means the second condition of the row protects more. Blue: the key comparison, measured against simulated channels with the number of channels, post-processing and RT_{60} matched. Every interval excludes zero; the closest, measured against 200 processed simulated responses at $\varepsilon = 0.002$, equal SNR (lower limit 0.0003), is not significant at the speaker level ($p = 0.28$), and neither are the same comparison at equal budget ($p = 0.15$) and the out-of-band blend at $\varepsilon = 0.004$, equal budget ($p = 0.08$). Equal-SNR values include the four trials of the speaker in C9, unprotected in every condition.

TABLE IV
CONTROL FOR THE SHARED COLOURING: A SEPARATE RUN WITH THE SAME 20 SPEAKERS, HELD-OUT CHANNELS AND RANDOM DRAWS AS TABLE III, THREE RUNS PER ε, EQUAL BUDGET. MEAN $\cos_B$, ATTACKER SUCCESS (%), AND Δ, THE PAIRED DIFFERENCE IN $\cos_B$ FROM THE MEASURED CONDITION ($\Delta > 0$: MEASURED PROTECTS MORE). ROWS "+ COLOURING": THE COLOURING SHARED BY THE THREE TRAINING RESPONSES ADDED TO EACH SIMULATED RESPONSE. AS IN TABLE III, Δ OF THE PROCESSED 200 AT $\varepsilon = 0.002$ IS NOT SIGNIFICANT AT THE SPEAKER LEVEL.

	$\varepsilon = 0.004$			$\varepsilon = 0.002$		
Training channels	$\cos_B$	succ.	Δ	$\cos_B$	succ.	Δ
Measured, 3 rooms	0.747	27.9	–	0.797	59.6	–
Sim., 3, RT_{60}-matched	0.792	54.6	0.045	0.820	79.6	0.023
+ colouring	0.754	34.6	0.007	0.801	65.8	0.005
Sim., 200, processed	0.769	37.1	0.022	0.805	67.1	0.008
+ colouring	0.723	24.2	−0.023	0.780	51.3	−0.017

only—the average of their 1/3-octave-smoothed magnitude responses, scaled to keep the in-band power, left at unit gain outside 250 Hz–7 kHz, and realised as a 4096-tap minimum-phase filter—and added it to each simulated response. The run repeated the measured and simulated conditions with the same random draws; they reproduced the first run to within 0.002 in $\cos_B$ on average (largest difference 0.019). At equal SNR every perturbation of a trial was scaled to the quietest of the five conditions of this run. The rule for reading the result was fixed before the run, as in Section VII-B.

Table IV gives the result. With the colouring added, three RT_{60}-matched simulated responses come within 0.007 of the measured ones at $\varepsilon = 0.004$, equal budget (95% CI 0.004–0.010; speaker-cluster bootstrap 0.003–0.011), and within 0.008 (0.005–0.010) at equal SNR. The colouring removes 84% of the gap at equal budget (speaker-cluster bootstrap 75–91%) and 81% (72–89%) at equal SNR, and 80% and 78% at $\varepsilon = 0.002$ (61–92% and 55–91%). What remains excludes zero at both budgets, in both evaluations and at the speaker level ($p = 0.0008$ to 0.010), so under the rule the measured responses still protect better than these simulated ones, but by less than a quarter of the original margin. The larger simulated set gains more: with the colouring, the 200 simulated responses lower $\cos_B$ by 0.023 more than the measured ones (0.018–0.029; speaker-cluster 0.014–0.033) at equal budget and by 0.021 at equal SNR (0.017 and 0.016 at $\varepsilon = 0.002$), in every held-out room and run. Their attacker success is lower too, 24.2% against 27.9%, but not significantly (95% CI of the difference −8.8 to +1.3 points, $p = 0.14$), and their PESQ is lower by 0.07 (2.00 against 2.07, $p < 10^{-17}$; 0.11 at $\varepsilon = 0.002$).

The colouring was meant to stand for the laptop loudspeaker and smartphone microphone that every measurement shares, but its dominant features come from our processing. Applied to a perfectly flat channel, the processing of Section V-C returns −30 dB at 350 Hz and −37 dB at 5.5 kHz relative to 1 kHz (C10), where the measured responses have their deepest features (Fig. 2). The held-out test channel, processed the same way, carries this colouring too. Most of the advantage of measured training channels in this protocol therefore comes from matching a colouring of the test channel, much of it an artefact of our deconvolution, rather than from describing the rooms better. What remains (0.007) may come from the rooms, the devices or parts of the processing that a

magnitude-only filter does not capture; these data cannot tell. The physical test of Section VII-H is the only test of the protection against an encoder in this paper that does not pass through this processing.

D. Ablations of the measurement processing

Each ablation rebuilds the three training responses from the original sweep recordings with one processing step (Sections V-B and V-C) switched off. With every step on, the rebuild reproduces the stored responses exactly (cosine similarity 1.000). Testing is always through the fully processed held-out channel. Every step helps, and every ablation interval excludes zero at both budgets and both evaluations (Fig. 3). Normalising before the blend ($\Delta = 0.041$ at $\varepsilon = 0.004$, equal budget) and frequency-dependent regularisation (0.029) matter most; the out-of-band blend (0.011) and output-level normalisation (0.008, and 0.004 at equal SNR) least, and the blend's effect is not significant at the speaker level at $\varepsilon = 0.004$, equal budget ($p = 0.08$; speaker-cluster interval from -0.001 to 0.021). With no processing at all, measured responses protect clearly less than either set of 200 simulated responses ($\Delta = 0.030$ and 0.017 at $\varepsilon = 0.004$, equal budget; speaker level $p < 0.01$) and no better than three RT_{60}-matched ones ($\Delta = 0.007$, 95% CI 0.002–0.012, not significant at the speaker level, $p = 0.13$). The advantage belongs to measured responses processed as described above, not to measurement alone. Because the held-out test channel keeps every step, these ablations cannot tell whether a step helps because it makes the training responses more accurate or because it makes them more like the test channel; Section VII-C and C10 point to the second.

E. Variation across runs

Across the five runs at $\varepsilon = 0.004$ and three at $\varepsilon = 0.002$, the SD of a condition's mean $\cos_B$ is at most 0.0012 and of its success rate at most 2.6 points (largest ranges 0.003 and 6.3 points). Every comparison in Fig. 3 has the same sign in every run. The variation that matters is between speakers and rooms, and the confidence intervals above are computed over it.

Identity, quality and intelligibility move largely separately. Across the channel-trained conditions at $\varepsilon = 0.004$, PESQ lies between 2.07 and 2.14 and ESTOI between 0.86 and 0.87 (Table III), while $\cos_B$ ranges from 0.747 to 0.799. The quality differences are small but not zero: conditions that hide identity better tend to score slightly lower on PESQ (correlation 0.81 across the nine conditions), and measured channels score 0.06 lower than RT_{60}-matched simulated ones (speaker level $p < 0.001$). PGD without a channel scores higher partly because its perturbation stayed zero for one speaker (C9); without that speaker its SNR is 24.7 dB and its PESQ 2.22.

F. Budget and perceptual cost

Similarity falls monotonically with budget, but audio quality falls faster: by $\varepsilon = 0.02$ the SNR is 2.2 dB and PESQ has reached the floor of the scale. A protection that destroys the recording is not a protection. The useful landmark is not the threshold but the *impostor mean* of 0.700: crossing τ_B at $\varepsilon = 0.0025$ is an 0.011 margin and lies inside the measurement noise, whereas at $\varepsilon = 0.004$ the protected clip scores 0.703, the impostor mean itself. We adopt $\varepsilon = 0.004$ as the operating point. This clip was quieter than average, so its SNRs are pessimistic; over the 20 speakers at their recorded levels the same ε gives 23.9 dB (Table VI).

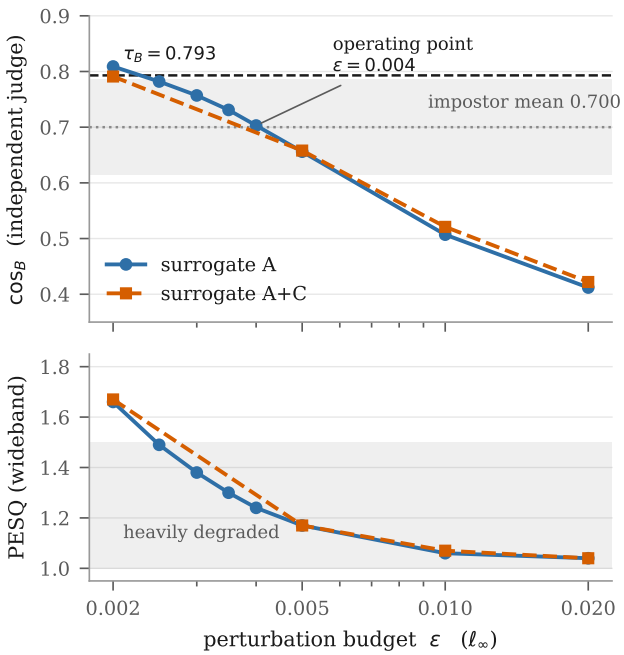


Fig. 4. Protection–quality trade-off for a single clip, mean over four held-out measured channels. The shaded band in the upper panel is the impostor distribution (0.700 ± 0.086): a protected clip entering it is indistinguishable from a different speaker. A second surrogate encoder (A+C) buys nothing.

TABLE V
TRADE-OFF FOR A SINGLE CLIP, MEAN OVER FOUR HELD-OUT CHANNELS. $\tau_B = 0.793$, IMPOSTOR MEAN 0.700.

Surrogate	ε	SNR (dB)	PESQ	$\cos_B$	$\cos_A$
–	0	–	4.64	0.919	0.878
A	0.0025	19.0	1.49	0.782	0.534
A	0.0030	17.4	1.38	0.757	0.503
A	0.0035	16.2	1.30	0.731	0.470
A	0.0040	15.1	1.24	0.703	0.444
A	0.0050	13.3	1.17	0.656	0.386
A	0.0100	7.8	1.06	0.507	0.234
A	0.0200	2.2	1.04	0.412	0.153
A+C	0.0050	13.3	1.17	0.658	0.424
A+C	0.0100	7.7	1.07	0.521	0.291

G. The operating point through held-out measured channels

Table VI and Fig. 5 give the result when the measured channel is applied computationally. The paired difference in $\cos_B$ is 0.128 ± 0.014, $t(79) = 18.84$, $p = 5.9 \times 10^{-31}$, $d_z = 2.11$; at the speaker level $t(19) = 10.08$, $p =$

TABLE VI

OPERATING POINT $\varepsilon = 0.004$, $n = 80$ (20 HELD-OUT SPEAKERS × 4 HELD-OUT CHANNELS), CHANNEL APPLIED COMPUTATIONALLY, MEAN OF 5 RUNS. ±: 95% CONFIDENCE INTERVAL; FOR THE SUCCESS RATE, SD ACROSS RUNS.

	Unprotected	Protected
Attacker succeeds ($\cos_B > \tau_B$)	97.5%	$\mathbf{27.8 \pm 1.4}$%
$\cos_B$ (independent judge)	0.875 ± 0.008	$\mathbf{0.747 \pm 0.017}$
$\cos_A$ (white-box)	0.850 ± 0.010	0.528 ± 0.023
SNR	–	23.9 dB
PESQ (quality)	4.64	2.07
STOI / ESTOI (intelligibility)	–	0.93 / 0.86

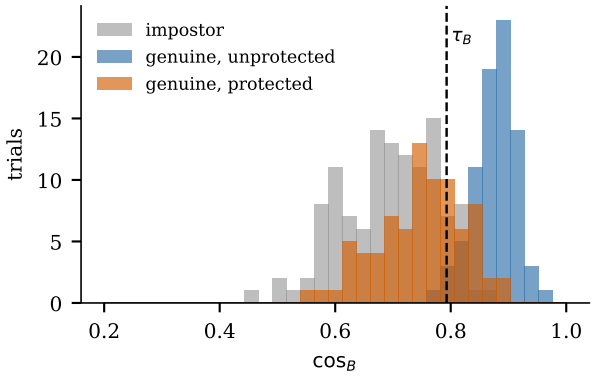


Fig. 5. Score distributions under the independent judge, measured channel models, from the original single-run evaluation (mean protected $\cos_B$ 0.747, the same as the five-run mean of Table VI). The protected distribution moves down into the impostor distribution; the residual overlap above τ_B is the 28% of trials in which the adversary still succeeds.

4.6×10^{-9}, and all 20 speakers move down. The protected mean of 0.747 lies below $\tau_B = 0.793$, and nearer the impostor mean (0.700) than the genuine mean (0.839) measured under the same conditions: on average, to an encoder it never saw, the protected talker scores more like a different speaker than like themselves.

We state plainly what this is not. It is not complete protection—the adversary still succeeds in over a quarter of trials. Nor is the perturbation inaudible: PESQ 2.07 predicts audible degradation, and we report quality and intelligibility as separate outcomes rather than folding them into one claim. And the verifier being defeated has a 16.25% EER under these conditions, which is weak; a stronger verifier would be harder to defeat.

H. Physical feasibility test: one clip, four rooms

This test asks whether the effect survives a real air gap at all, not how large it is in general: one talker, one clip, one loudspeaker and one microphone, four rooms. Each room gives one outcome, so "100% to 75%" means that the adversary succeeded in all four rooms before protection and in three of the four after it. It shows feasibility under these conditions and does not show generalisation across speakers, sentences, devices or verifiers.

Within those limits (Table VII, Fig. 6), the direction survives the air gap: protection lowers the independent

TABLE VII

PHYSICAL FEASIBILITY TEST WITH ONE PROTECTED CLIP. THE CLIP WAS PLAYED THROUGH A LAPTOP LOUDSPEAKER, CROSSED EACH ROOM, AND WAS RE-RECORDED ON A SMARTPHONE. JUDGE B IS THE INDEPENDENT ENCODER; ENROLMENT IS THE MEAN OVER ACTUAL RE-RECORDINGS OF THE CLEAN CLIP IN THE OTHER THREE ROOMS. $\tau_B = 0.793$. EACH ROOM IS ONE OUTCOME.

Room	clean	protected	drop
Normal room	0.892	0.853	0.039
Large room, centre	0.937	0.860	0.077
Large room, corner	0.899	0.786	0.113
Stairwell	0.897	0.840	0.057
mean	**0.906**	**0.835**	**0.071**
Attacker succeeds	4 of 4 rooms	3 of 4 rooms	

TABLE VIII

PERTURBATIONS EVALUATED THROUGH THE FULL SIMULATED CHAIN (ROOM, SATURATION, GAIN CONTROL, DRIFT, NOISE AND A REAL AAC ROUND-TRIP AT 64 KBIT/S). JUDGE B, $\varepsilon = 0.004$, SNR 23.6–24.0 DB.

Held-out room	none	room-only EOT	full-chain EOT
Normal room	0.856	0.707	**0.675**
Large room, centre	0.968	0.836	**0.825**
Large room, corner	0.883	0.863	**0.854**
Stairwell	0.951	0.881	**0.856**
mean	0.915	0.822	**0.802**

judge's similarity in every room, by 0.039 to 0.113. The size is smaller: the mean drop is 0.071, against 0.128 through the channel models. The two are not like-for-like—one clip with same-utterance enrolment against 80 trials of other clips—but the difference is consistent with the analysis below. The single room in which the adversary failed, the large room's corner, did so by a margin of 0.007—noise rather than a result. The white-box judge shows the same pattern more strongly ($0.878 \rightarrow 0.717$), which is expected and is not evidence of generalisation.

Why. We can locate the discrepancy quantitatively. Enrolling the adversary on clips convolved with our measured impulse responses, rather than on actual re-recordings, gives a similarity of only 0.336 to a genuine re-recording of the same clip in the same room—while two genuine re-recordings of that clip match at 0.906. The measured channel and the physical channel are, from the encoder's point of view, almost different signals. Part of the difference may be the colouring that our processing adds to every measured response (C10), which a real recording does not have. Reverberation is modelled; loudspeaker saturation, microphone automatic gain control, sample-clock mismatch and lossy coding are not, and these are plausible causes of losing a low-amplitude, fine-structured perturbation.

I. Extending the channel model

Section VII-H points to the non-linear stages, so we addressed them directly: the transformation T_h was extended from convolution alone to the full chain—random gain, tanh saturation standing for loudspeaker compression, a

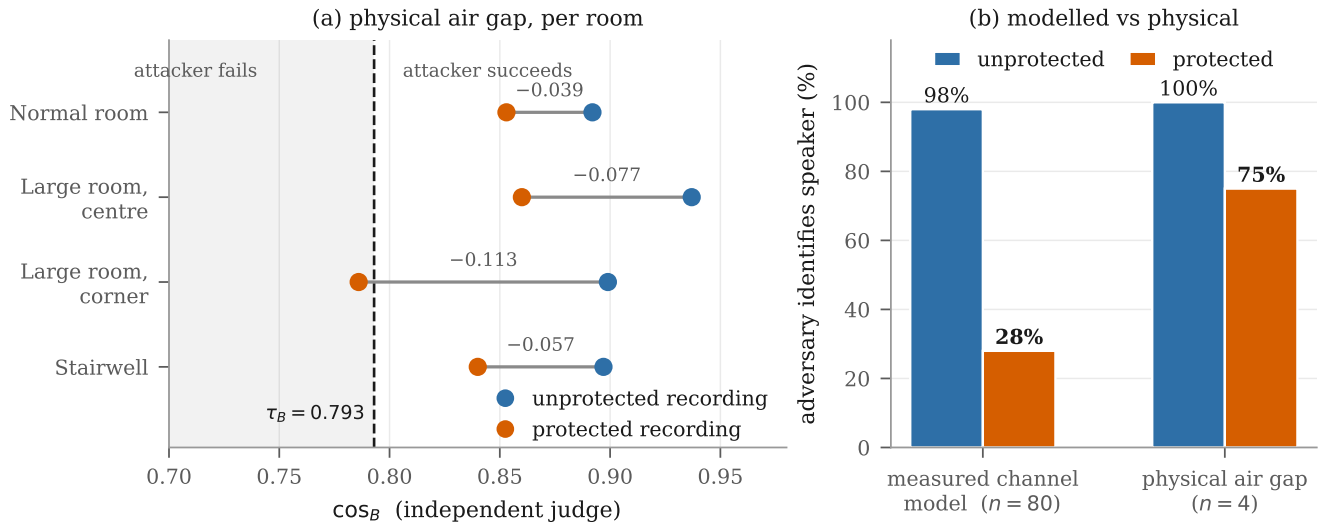


Fig. 6. The physical feasibility test. (a) Each room's unprotected and protected re-recording under the independent judge; the number on each connector is the drop achieved by protection. Three of four protected recordings remain on the "attacker succeeds" side of the threshold, and the fourth clears it by 0.007. (b) The same method evaluated two ways: through the measured channel models (original single-run evaluation, 80 trials) and through the physical loudspeaker–room–microphone path (one clip, four rooms); the two are not like-for-like. The difference between them is the subject of Sections VII-H and VII-I.

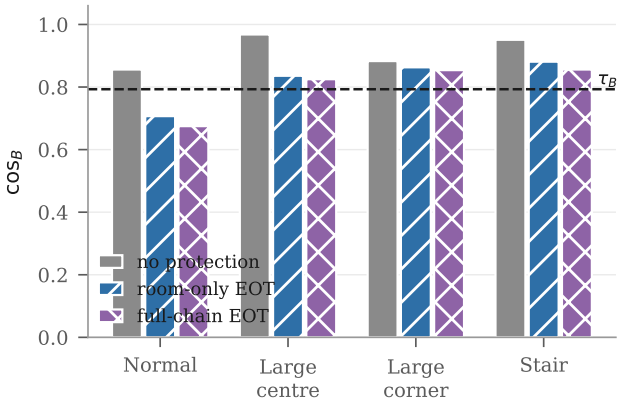


Fig. 7. Extending the transformation distribution from the room alone to the full chain improves protection in all four channels, and by far too little. Evaluation includes a genuine AAC round-trip at 64 kbit/s.

soft dynamic-range compressor standing for microphone gain control, a ±0.1% resampling standing for clock drift, additive noise, and a differentiable spectral-masking proxy for perceptual coding. Evaluation used the genuine AAC encoder at 64 kbit/s rather than the proxy.

The result (Table VIII and Fig. 7) is consistent and insufficient. Full-chain training improves on room-only training in *all four* channels, by 0.009 to 0.032, mean 0.020. The direction is consistent with the diagnosis; the magnitude does not change any outcome, since closing the gap to Table VI would require roughly 0.15.

J. Negative results on transferability

Four standard improvements to black-box transfer were attempted and none helped at matched SNR.

Surrogate ensembling. Optimising against models A and C jointly gave no consistent gain over A alone. We attribute this to insufficient diversity: C differs in depth, dilation and embedding size but consumes the *same* 80-band log-mel front end, so the ensemble largely teaches the perturbation to exploit that feature extractor.

Targeted objective. Steering the embedding toward a deliberately dissimilar held-out speaker (cosine −0.262 to the source), as in [3], did not improve $\cos_B$.

Gradient momentum. Momentum of 0.9 on the normalised gradient, after [25], did not improve transfer.

Psychoacoustic masking. Shaping the per-sample budget by a smoothed speech envelope raised SNR by about 6 dB at equal ε, but at *equal SNR* performed no better. The masking buys quietness, not efficiency; comparing the two at equal ε would have produced a spurious conclusion (C6).

K. Listening tests

PESQ, STOI and ESTOI are models of perception, not measurements of it, so we also asked people, in two tests.

Test 1 (uncontrolled, $n = 22$). A web page presented nine trials built from the same 3-s utterance as the physical test. Five were forced-choice "same person or different?" pairs: the clean against the protected clip, both digital; the clean against the protected clip, both re-recorded in the normal room; and three controls pairing the talker with someone else—twice with a speaker matched in pitch (against the clean and against the protected clip) and once with a clearly different speaker, as an attention check. Four trials asked for a quality rating from 1 (bad) to 5 (excellent), always in the order clean, clean re-recorded, protected re-recorded, protected. The five pairs were shuffled for each listener, and which clip of a pair came first

TABLE IX
LISTENING TEST 1, $n = 22$ LISTENERS WHO COMPLETED ALL NINE TRIALS ON THEIR OWN LOUDSPEAKERS. TOP: SHARE ANSWERING "SAME PERSON", WITH WILSON 95% INTERVALS; THE TRUE ANSWER IS IN PARENTHESES. BOTTOM: MEAN OPINION SCORE (1–5) ± ONE STANDARD DEVIATION.

Pair	"Same"	95% CI
Clean vs protected, digital (same)	86%	67–95%
Clean vs protected, re-recorded (same)	95%	78–99%
Control 1: clean vs pitch-matched speaker (diff.)	73%	52–87%
Control 2: protected vs dissimilar speaker (diff.)	9%	3–28%
Control 3: protected vs pitch-matched speaker (diff.)	59%	39–77%
Clip	**MOS**	
Clean, digital	3.77 ± 0.87	
Protected, digital	3.59 ± 1.14	
Clean, re-recorded	2.64 ± 0.85	
Protected, re-recorded	2.27 ± 0.77	

was randomised. Listeners used their own laptop or phone loudspeakers at a volume they were asked to keep fixed. Twenty-two untrained volunteers completed all nine trials; incomplete sessions are excluded.[1]

Table IX gives the result. Listeners judged the protected clip to be the same person as the clean one in 86% of digital pairs and 95% of re-recorded pairs. The controls qualify this. A different speaker matched in pitch was also judged "the same person" 73% of the time, so these listeners tell similar voices apart poorly, and the protected pairs were not significantly more often judged "same" than that control (exact McNemar $p = 0.45$ for the digital pair, $p = 0.06$ for the re-recorded pair). The test therefore shows that the protection did not make the talker sound like someone else; with this panel it cannot show more. The attention check behaved as intended: 2 of the 22 listeners judged the dissimilar speaker "same", and excluding them changes no conclusion (90% and 95% "same" for the two protection pairs).

On quality, the protected clip was rated 0.18 points below the clean clip (Wilcoxon signed-rank $p = 0.51$), and after re-recording 0.36 points below the clean re-recording ($p = 0.07$); re-recording itself cost more than a point. PESQ predicts clearly audible degradation (2.07 against 4.64), whereas listeners on consumer loudspeakers rated the digital difference as small. This does not show that the perturbation is inaudible: the clean clip itself averaged only 3.77, which reflects the playback conditions, and an untrained panel of 22 on uncontrolled devices has little power to detect a small difference.

Test 2 (controlled, analysis fixed in advance). Test 1 used one talker, one sentence and no check of the playback device, so we designed a second test whose analysis was fixed before any data were collected. Its stimuli come from 12 held-out LibriSpeech speakers (6 female, 6 male), none used to train any encoder, each level-normalised and protected at $\varepsilon = 0.004$ by EOT over all four measured channels (SNR 23.2–24.3 dB, mean 23.9; PESQ mean 2.08; ESTOI mean 0.88). Each listener first sets a comfortable volume and then completes a headphone screen of three trials in which one of three tones is played in antiphase between the ears [22]; through loudspeakers the antiphase tone cancels and the quietest tone is misjudged. The screen is recorded, not enforced: listeners who fail it continue and are analysed in a sensitivity analysis. Each listener then hears 4 of the 12 speakers (2 female, 2 male; one of three fixed groups, chosen at random), so all 12 are covered across listeners, in three tasks:

- *Intelligibility*: type 4 sentences (2 clean, 2 protected; two counterbalanced lists), each played at most twice; scored as the share of words correct, with one-letter misspellings of words of five or more letters accepted.
- *Identity*: 14 same-or-different pairs, never the same words: for each speaker, clean against clean, clean against protected, and clean against a different speaker of the same gender with the closest median pitch; plus 2 attention pairs of a man and a woman.
- *Quality*: 9 clips rated from 1 to 5 (4 clean, 4 protected, 1 deliberately degraded anchor).

Question order and the order within pairs are random for every listener. The unit of analysis is the listener. The primary outcomes are paired differences, protected minus clean: P1 the share of "same person" answers on true same-speaker pairs, P2 the share of words typed correctly, and P3 the mean quality rating. Each is tested with a Wilcoxon signed-rank test, reported with a bootstrap 95% interval over listeners, and corrected across the three by Holm's method [23]. Equivalence is tested by two one-sided tests [24] with margins fixed in advance (P1 ±0.10, P2 ±0.05, P3 ±0.30), and the word "equivalent" is used only if that test passes. Runs made by the researchers to check the page are not data. Sessions are excluded if unfinished, if they are a repeat of the same name, or if both attention pairs were answered "same". Sensitivity analyses restrict to listeners who passed all three headphone trials, to self-reported headphone users, and to listeners without a reported hearing difficulty.

TABLE X
LISTENING TEST 2 (ANALYSIS FIXED IN ADVANCE; UNIT = LISTENER; $n = 28$). DIFFERENCE = PROTECTED MINUS CLEAN, WITH BOOTSTRAP 95% INTERVAL OVER LISTENERS; p_{TOST} TESTS EQUIVALENCE WITHIN THE MARGINS FIXED IN ADVANCE (±0.10, ±0.05, ±0.30).

Outcome	Clean	Prot.	Diff. [95% CI]	p_{Holm}	p_{TOST}
P1 "same" (true pairs)	0.76	0.60	−0.16 [−0.27, −0.06]	0.021	0.86
P2 words correct	0.64	0.59	−0.05 [−0.14, 0.04]	0.30	0.49
P3 quality (MOS)	4.15	2.97	−1.18 [−1.48, −0.87]	3×10^{-5}	1.00

Table X gives the result for the 28 listeners who had finished by 30 October; none met an exclusion rule, and all answered both attention pairs correctly. Protection

[1]The accompanying video also counts three incomplete sessions, one of them the first author's own test run. This moves each figure by at most three percentage points or 0.06 MOS (there, for example, 88% and 96% for the two protection pairs and 70% for control 1).

lowered the quality rating by 1.18 points (95% CI 0.87–1.48; $p_{\text{Holm}} = 3 \times 10^{-5}$) and the share of "same person" answers on true same-speaker pairs from 76% to 60% (-0.16, 0.06–0.27; $p_{\text{Holm}} = 0.021$), while pairs of different speakers were judged "same" only 10% of the time. The share of words typed correctly fell from 64% to 59%, a difference that is not significant ($p_{\text{Holm}} = 0.30$) but whose interval (-0.14 to 0.04) is too wide to show equivalence. No outcome is equivalent within the margins fixed in advance. The 12 listeners who passed all three headphone trials and the 16 who reported using headphones or earphones show the same pattern (quality -1.40 and -1.31; "same" -0.21 and -0.20). The protection at $\varepsilon = 0.004$ is therefore clearly audible, and it makes listeners less sure that the talker is the same person; this agrees with PESQ and qualifies the result of test 1.

VIII. Methodological Corrections

We document errors made during the project, because several produced plausible and flattering numbers, and because the corrections are the part of this work most likely to help others attempting the same measurement.

C1. An evaluation in which the room did the work. Our first air-gap result was striking: re-recorded protected speech at about -0.3 cosine against about $+0.3$ for re-recorded clean speech. But the *clean* re-recording already scored below threshold. The room alone was defeating the encoder and the perturbation was receiving credit. The cause was enrolling the adversary on pristine digital audio, which no realistic adversary would do. Every number in this paper uses channel-matched enrolment.

C2. A borrowed decision threshold. Scores were initially interpreted against $\tau = 0.463$, model A's equal-error threshold on *clean* audio. That is neither model B's threshold nor valid under reverberation. Measured properly, the thresholds are 0.734 and 0.793.

C3. Low-frequency blow-up in the deconvolution. The first impulse responses gave RT_{60} estimates of 0.98, 2.15 and 5.52 s with more than 99% of energy arriving after 100 ms—visibly broadband noise, with magnitude rising to +30 dB below 100 Hz. The cause was dividing room noise by a loudspeaker output that is essentially zero below 150 Hz. Frequency-dependent regularisation restored physically plausible responses.

C4. A band-limited filter is not a room. Having band-limited to 250 Hz–7 kHz, convolving speech with the result destroyed speaker identity far more thoroughly than the real room did ($\cos_A$ -0.02 against $+0.22$), because the filter deleted the fundamental and first formant. The pass-through blend fixed this.

C5. A scaling bug in that blend. Our first implementation added a *unit-amplitude* impulse to an unnormalised measured response. The impulse dominated and, contributing only out-of-band, produced a band-*stop* filter that removed the mid band where identity lives. The symptom was $RT_{60} = 0.00$ s. Normalising before blending resolved it.

C6. An SNR confound in a method comparison. Masking raises SNR at fixed ε, so comparing masked and unmasked methods at equal ε compares a quieter perturbation with a louder one. Re-running at matched SNR reversed the apparent conclusion.

C7. ε is absolute; SNR is relative. A quiet source receives a proportionally louder perturbation. Our first operating-point clip sat at 15.1 dB SNR where the population mean at the same ε was 23.9 dB. The clip re-recorded in the rooms and used in listening test 1 was therefore regenerated from a level-normalised source, which places it at that mean.

C8. An unmatched comparison of channel models. Our first comparison of measured and simulated training channels, on one clip, found measured channels better in each of the three held-out rooms it tested. It differed in more than the channel model: the number of channels (3 against 200), their length, their normalisation, the in-band treatment and the source of the simulated rooms. Section VII-B replaces it with a matched comparison on 20 speakers with repeated runs.

C9. A stationary starting point. PGD without a channel starts from $\delta = 0$, where its objective, the cosine to the clean embedding, is at its maximum of 1 and the gradient vanishes; the attack moves only because floating-point noise gives the gradient a sign. For one of the 20 speakers the computed gradient was exactly zero, so its perturbation stayed zero at both budgets. We found this after the run, while checking the paper's numbers against the raw results. It inflates PGD's mean SNR (31.2 instead of 24.7 dB at $\varepsilon = 0.004$), and because equal-SNR scaling follows the quietest perturbation of each trial, it leaves that speaker unprotected in every condition at equal SNR. We report the planned analysis with that speaker included and give the values without it alongside; no conclusion changes. The EOT conditions are not affected, because the channel moves the signal away from the maximum. A random start would avoid the problem.

C10. A colouring made by the deconvolution. We chose λ to stop the low-frequency blow-up of C3 and did not check what it does inside the band. Applied to a perfectly flat channel—a recording identical to the reference sweep—our processing does not return a flat response: relative to 1 kHz, its 1/3-octave level is -30 dB at 350 Hz, -2.8 dB at 2 kHz and -37 dB at 5.5 kHz. The power of the exponential sweep falls as $1/f$, 296 times below its peak at 5 kHz, so the in-band λ exceeds $|X|^2$ at high frequencies, and the rise of λ toward the band edges reaches well into the band (Section V-C). Every measured response carries this colouring, and so does every held-out test channel; the simulated responses, which need no deconvolution, do not. It accounts for much of the shared colouring of Section VII-C, and so for much of the advantage of measured over simulated training channels, and it is consistent with the ablations (Section VII-D): training over processed responses prepares the protection for the colouring of the processed test channel. The physical test (Section VII-H) does not pass through this processing. We

found the problem after the control of Section VII-C, while checking what its filter contained. A regularisation that follows the spectrum of the sweep, or the inverse filter of [16], would avoid it; repeating the comparison with the same recordings deconvolved that way is future work.

Resolved discrepancy. Earlier we measured PESQ 1.66 at $\varepsilon = 0.002$ for measured-channel EOT, against about 2.62 in an earlier simulated-channel run at comparable SNR, and could not say whether this was a real perceptual difference. In the matched run, over the same 80 trials, PESQ at $\varepsilon = 0.002$ is 2.75 for measured and 2.72 for the earlier simulated setting, at mean SNRs of 29.5 and 29.9 dB: the earlier gap was not caused by the channel model.

Practical measurement notes. Smartphone recorder applications may apply automatic gain control, which breaks the linear time-invariance the measurement presumes; recordings must be checked for level pumping along the sweep. Clipping must be checked for, because deconvolution cannot undo it. And partially-synchronised cloud files decode as truncated audio and produce silently wrong impulse responses.

IX. Limitations

Processing, devices and rooms are confounded. Every measured response and every held-out test channel passed through the same deconvolution, laptop loudspeaker and smartphone microphone. A colouring shared by all of them explains about 80% of the advantage of measured over simulated training channels (Section VII-C), and much of that colouring is made by the deconvolution (C10). The comparison should be repeated with responses deconvolved without this artefact, and with several devices. Until then, the channel-model results of this paper describe protection through our processed responses; of the tests against an encoder, only the physical test (Section VII-H) is free of the processing.

One simulator. All simulated responses come from one corpus of image-method rooms [18]. A different simulator may narrow or widen the gap.

No adaptive attacker. We do not test purification [8], denoising, or an encoder retrained on protected audio; any of these may remove much of the protection.

No end-to-end cloning system is evaluated. We measure identity extraction by speaker encoders, the component a cloning pipeline depends on, but we do not show that a synthesiser fails to reproduce the voice. The no-pretrained-model constraint makes this infeasible. Passing or failing a verification threshold is not the same as producing a usable fake, and we make no claim about the relationship between the two.

Small encoders, limited data. Our encoders have 0.5–1.7M parameters trained on 100 hours. Methods achieving strong black-box transfer ensemble several large pretrained encoders trained on thousands of hours. Our limited transfer is more likely a consequence of surrogate weakness and homogeneity than of the method.

Weak verifiers under test conditions. Both judges degrade to 16–18% EER under reverberation with channel-matched enrolment.

Dependent trials. The 80 trials are 20 speakers in 4 rooms. We therefore report speaker-level tests alongside the trial-level intervals; the key comparison holds at both levels, and the exceptions are stated where they occur.

The physical test is a feasibility check. It uses one talker's clip, one loudspeaker and one microphone in four rooms, and enrols on re-recordings of the *same* utterance used at test time, which inflates both columns of Table VII; the gap between them is the meaningful quantity. Consumer hardware restricts the measurement band to 250 Hz–7 kHz and may contribute much of the non-linearity we did not model.

Listening tests. Test 1 is small and uncontrolled: 22 untrained listeners, one utterance, their own loudspeakers, no hearing screening. Test 2 is controlled for the playback device only by a headphone screen, runs on listeners' own equipment (12 of 28 passed all three headphone trials), and 28 listeners are too few to show equivalence within the margins.

X. Conclusion

We asked whether the channel model used to optimise a voice protection matters, and built a protocol to answer it: four channels measured with consumer devices, leave-one-channel-out evaluation, an independently trained judge, a threshold derived under the test conditions, and comparisons at equal budget and equal SNR over repeated runs.

Under this protocol, protection trained over measured channels cut an unseen encoder's identification rate from 98% to 28% through held-out measured channels, at 23.9 dB SNR, with PESQ 2.07 and ESTOI 0.86. Against simulated channels matched in number, reverberation time and post-processing, measured channels protected better in every held-out room and every run. A control then showed that about 80% of that advantage came from a spectral colouring shared by the measured training and test channels, much of it made by our own deconvolution; with that colouring added, a larger simulated set lowered the judge's similarity more than the measured channels did. In this protocol, then, the advantage of measured channels is mostly a match to a colouring of the test channel, and gives little evidence that measured responses describe real rooms better. Repeating the comparison with responses free of this artefact is the next experiment.

In a physical feasibility test with one clip in four rooms, the only test against an encoder here that does not pass through our processing, the protection kept its direction but lost part of its size. The modelled channels differ from the physical ones both in the colouring of C10 and in the non-linear stages of the chain, and modelling those stages helped only slightly. For anyone evaluating voice protection, the practical lessons are to pass a perfectly flat channel through any measurement processing before trusting the responses it returns, to match conditions before comparing channel models, to report identity, quality

and intelligibility separately, and, since in our one-clip test the modelled channel overstated the protection, to check any modelled channel—simulated or measured—against real recordings.

Ethics Statement

No real person's voice was cloned or impersonated. All speech is drawn from a public research corpus of consenting speakers, or recorded by the first author. The method studied is defensive and provides no capability for impersonation. Listening-test participants were volunteers who were told how their answers would be used and gave consent before starting. In test 2 no voice was recorded, and each participant's name and answers were encrypted so that only the researchers could read them.

Reproducibility

All models were trained from scratch and all code was written for this project. Splits are speaker-disjoint with a fixed seed. The matched comparison, the colouring control, the ablations, the repeated runs and the operating point (Sections VII-B–VII-E and VII-G) fix the seed of every random choice and store every trial in two files (13,440 and 5,280 rows). Every result of these sections, of Tables III, IV and VI and of C9 that comes from these runs is computed from the two files by one script, `paper_numbers.py`, with the conventions of Section VI; the same script runs the flat-channel check of C10. The earlier runs (trade-off, physical test, full chain) did not seed the random choice of rooms inside the optimisation, so a re-run reproduces them up to that randomness. Measured impulse responses, recordings, numerical results and the script are retained.

XI. Video

https://youtu.be/BhY64VnKid0

References

[1] C.-y. Huang, Y. Y. Lin, H.-y. Lee and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 552–559.

[2] 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 *NeurIPS*, 2022.

[3] Z. Yu, S. Zhai and N. Zhang, "AntiFake: Using adversarial audio to prevent unauthorized speech synthesis," in *ACM CCS*, 2023.

[4] J. Li, D. Ye, L. Tang, C. Chen and S. Hu, "Voice Guard: Protecting voice privacy with strong and imperceptible adversarial perturbation in the time domain," in *IJCAI*, 2023.

[5] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang and W. Xu, "V-Cloak: Intelligibility-, naturalness- & timbre-preserving real-time voice anonymization," in *USENIX Security*, 2023.

[6] Z. Zhang *et al.*, "SafeSpeech: Robust and universal voice protection against malicious speech synthesis," in *USENIX Security*, 2025.

[7] J. Liu *et al.*, "SpeechShield: Latency-efficient and robust timbre-aware voice protection against speech synthesis deepfake attacks," in *IEEE/IFIP DSN*, 2026.

[8] W. Fan, K. Chen, C. Liu, W. Zhang and N. Yu, "De-AntiFake: Rethinking the protective perturbations against voice cloning attacks," in *ICML*, 2025.

[9] A. Athalye, L. Engstrom, A. Ilyas and K. Kwok, "Synthesizing robust adversarial examples," in *ICML*, 2018.

[10] Y. Qin *et al.*, "Imperceptible, robust and targeted adversarial examples for automatic speech recognition," in *ICML*, 2019.

[11] T. Chen, L. Shangguan, Z. Li and K. Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *NDSS*, 2020.

[12] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP*, 2020.

[13] V. Panayotov *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[14] D. Snyder *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.

[15] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[16] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES Convention*, 2000.

[17] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, 1965.

[18] 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.

[19] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ)," 2001.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[21] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[22] K. J. P. Woods, M. H. Siegel, J. Traer and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, pp. 2064–2072, 2017.

[23] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[24] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social Psychological and Personality Science*, vol. 8, no. 4, pp. 355–362, 2017.

[25] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *CVPR*, 2018.