Imperio: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems

Lea Schönherr, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa
Ruhr University Bochum, Germany
{lea.schoenherr, steffen.zeiler, thorsten.holz, dorothea.kolossa}@rub.de

Abstract—Automatic speech recognition (ASR) systems are possible to fool via targeted adversarial examples. These can induce the ASR to produce arbitrary transcriptions in response to any type of audio signal, be it speech, environmental sounds, or music. However, in general, those adversarial examples did not work in a real-world setup, where the examples are played over the air but have to be fed into the ASR system directly. In some cases, where the adversarial examples could be successfully played over the air, the attacks require precise information about the room where the attack takes place in order to tailor the adversarial examples to a specific setup and are therefore not transferable to other rooms. Other attacks, which are robust in an over-the-air attack, are either handcrafted examples or human listeners can easily recognize the target transcription, once they have been alerted to its content.

In this paper, we demonstrate the first generic algorithm that produces adversarial examples which remain robust in an over-the-air attack such that the ASR system transcribes the target transcription after actually being replayed. For the proposed algorithm, guessing a rough approximation of the room characteristics is enough and no actual access to the room is required. We use the ASR system Kaldi to demonstrate an end-to-end attack and employ a room-impulse-response (RIR) simulator to harden the adversarial examples against varying room characteristics. Further, the algorithm can also utilize psychoacoustics to hide most of the changes of the original audio signal below the human thresholds of hearing. We show that the adversarial examples work for varying room setups, but also can be tailored to specific room setups. As a result, an attacker can optimize adversarial examples for any kind of target transcription and to arbitrary room setups. Additionally, the adversarial examples remain transferable to varying room conditions with a high probability.

Index Terms—audio adversarial examples, automatic speech recognition, over-the-air attack

I. INTRODUCTION

In restless dreams I walked alone. Narrow streets of cobblestone. 'Neath the halo of a streetlamp. I turned my collar to the cold and damp. When my eyes were stabbed by the flash of a neon light. That split the night. And touched the sound of silence.

Simon & Garfunkel, The Sound of Silence

Adversarial examples have been shown to represent a threat for state-of-the-art automatic speech recognition (ASR) systems. More specifically, several recent works have demonstrated that it is possible to fool different kinds of ASR systems [1]–[4]. Such attacks were demonstrated against both connectionist temporal classification (CTC) loss ASR systems as well as DNN-HMM systems, which are hybrids using a combination of deep neural networks (DNNs) and hidden Markov models (HMMs) in a joint system.

The practical implications and real-world impact of the demonstrated attacks are unclear at the moment. On the one hand, earlier work only feed the audio adversarial directly into ASR system [1]–[3], hence ignoring all the side effects (e.g., reflections or echo) of a real-world, end-to-end attack. On the other hand, some works demonstrated adversarial examples that can be played over-the-air [5], [6], but these proof-of-concept attacks are specifically tailored for one specific, static room setup or are hard to reproduce with a proven success rate in a different environment. We also note that when over-the-air adversarial examples, used against black-box systems, the target transcription is easy to perceive for human listeners, once the intended attack is known.

We argue that adversarial examples for ASR systems can only be considered as a real threat if the targeted recognition is produced even when the signal is played over the air. Due to the high variability of possible room setups and the induced distortions, it is very hard to create robust adversarial examples, which work for a large subset of possible acoustic conditions. Depending on the attack scenario, an attacker can use some prior knowledge, e.g., if the attack is broadcasted via TV commercials, an attacker can try to estimate the actual room geometry and size. For this, the room impulse response (RIR), which describes the transmission of an acoustic signal in a room [7], needs to be taken into account. The transmission can be modeled as a convolution of the original audio signal with room-dependent RIR (see Figure 1 for an illustration), where the RIR depends on various factors. In practice, it is nearly impossible to estimate an exact RIR without having access to the actual room. Therefore, robust adversarial ex-
amples need to take a range of possible RIRs into account to increase the success rate.

The first adversarial audio examples that are imperceptible by humans even if they know the target transcription have been described by Carlini and Wagner [1]. Other approaches [4] have been successful at embedding the changes below the human threshold of hearing, which makes it much harder to detect an adversarial audio file. On the downside, in all of these cases, the attack was not successful when played over-the-air instead of being directly fed into the ASR system.

Other approaches, which did work over the air, have only been tested in a static setup (i.e., fixed position of speaker and microphone with a fixed distance). Yakura’s and Sakuma’s approach can hide the target transcription but requires physical access to the room, which limits their attack to one very specific room setup [8]. Concurrently, Szurley and Kolter published room-dependent robust adversarial examples, which even worked under constraints given by a psychoacoustic model [9]. However, their adversarial examples have only worked in an anechoic chamber (a room designed to absorb reflections). The attack can, therefore, not be compared with a real-world scenario, as the recorded audio signal in an anechoic chamber corresponds to the direct sound with only very minor changes. In other successful over-the-air attacks, human listeners can easily recognize the target transcription once they are alerted to its content [5], [6].

In the visual domain, Athalye and Sutskever presented a real-world adversarial perturbation on a 3D-printed turtle, which is recognized as a rifle from almost every point of view [10]. The algorithm to create this 3D object not only minimizes the distortion for one image, but for all possible projections of a 3D object into a 2D image. We borrow the idea of this approach and transfer it to the audio domain, replacing the projections by convolutions with RIRs and thereby harden the audio adversarial example against the transmission through varying rooms.

More specifically, we introduce in this paper a generic and robust approach to generate over-the-air adversarial examples against ASR systems by utilizing an RIR generator to sample from different room setups. For the simulation, the convolution with the sampled RIR is added to the DNN as an additional layer, which enables us to update the original audio signal directly under the constraints given by the simulated RIR. For this, the RIRs are drawn out of a distribution of room setups to simulate the over-the-air attack. The algorithm is repeated until the target transcription is recognized or a maximum number of iterations is reached. Using this approach, the adversarial examples are hardened to remain robust in real over-the-air attacks across various room setups. We also show an improvement that is based on psychoacoustic hiding [11], by including hearing thresholds in the backpropagation, as proposed by Schönherr et al [3].

We have implemented the proposed algorithm to attack the DNN-HMM ASR system Kaldi [12] under varying room conditions. The attack works in both cases, with and without psychoacoustic hiding. In either case, we can produce successful robust adversarial examples. With the generic approach that we have implemented, it is possible to induce an arbitrary target transcription and the attacker does not need physical access to the room where the attack takes place. We show that for a successful attack, a rough estimation of the room geometry and audio decay time is enough and the adversarial examples still continue to work if the real setup actually differs from the approximation.

In summary, we make the following contributions in this paper:

- **Over-the-air attack.** We propose an approach to generate robust over-the-air adversarial examples for DNN-HMM-based ASR systems. The attack uses a DNN convolution layer to simulate the effect of RIRs, which allows us to back-propagate gradients directly to the raw audio signal.
- **Psychoacoustics.** We show that the attack can be combined with psychoacoustics for hiding transcriptions in arbitrary audio files.
- **Performance Analysis.** We measure the accuracy of the adversarial attack and the degree of perturbation of the audio signal, both with and without hearing thresholds.

A demonstration of our attack is available online at [http://ota-adversarial-examples.selfip.org](http://ota-adversarial-examples.selfip.org) where we present several adversarial audio files, which have been successful when played over-the-air.

## II. BACKGROUND

In the following, we provide an overview of the ASR system that was used in the attack and describe the general approach to calculate audio adversarial examples. Furthermore, we discuss how room simulation can be performed with the help of RIRs and briefly introduce the necessary background from psychoacoustics to understand the rest of this paper.

### A. Automatic Speech Recognition

We chose the open-source toolkit Kaldi [12] as our ASR system, also employing the extension to create adversarial examples that is provided by Schönherr et al. [3].

The DNN-HMM-based ASR system can be divided into three parts: the feature extraction, which transforms the raw input data into representative features (so-called pseudo-posteriors), the DNN as the acoustic model of the system, and the decoding step, which returns the recognized transcription.

In Schönherr et al.’s approach, The feature extraction is integrated into the DNN, which enables us to change the raw audio file directly when creating the adversarial example. The approach is shown in Figure 2.

### B. Audio Adversarial Examples

The ASR system can be described as a function

\[
\hat{y} = \arg \max_y P(\hat{y}|x) = f(x),
\]

mapping some audio signal \( x \) to its corresponding, most likely transcription \( y \). By modifying the original input

\[
x' = x + \delta, \quad \text{such that} \quad f(x) \neq f(x'),
\]

as proposed by Schönherr et al. [3].
an adversarial example can be obtained. Here, \( \delta \) can also be restricted, e.g., via hearing thresholds. In this work, only targeted attacks are considered, where the target transcription \( y' = f(x') \) is defined. The optimization can, therefore, be described as

\[
    x' = \arg \max_{\hat{x}} P(y'|\hat{x}).
\]

To calculate over-the-air-robust adversarial examples, we used the implementation of Schönherr et al. \cite{3}. The system can be divided into three steps: (i) To get the best possible starting point, forced alignment is used to find the optimal pseudo-posterior matrix (representation of the output of the DNN, before the decoding step) for the given audio file and target transcription. (ii) By integration the feature extraction into the DNN, the audio data can be updated directly via gradient descent and (iii) with hearing thresholds, which depend on the original audio file, the added noise is limited to time-frequency ranges, where it is not (or only barely) perceptible by humans.

In the following, we extend this system in order to harden the adversarial examples against the effect of the audio signal transmission through air, which is modeled via the application of an RIR.

C. Room Impulse Response

When the signal is transmitted through a room, as visualized in Figures 1 and 2, the recorded signal can be computed by convolving the room’s impulse response \( h \) with the original audio signal \( x \) according to

\[
x_h = x \ast h.
\]

Here, the convolution operator \( \ast \) is a shorthand notation for

\[
x_h(n) = \sum_{m=n-M+1}^{n} x(m) \cdot h(n-m)
\]

with \( n = 0, \ldots, N - 1 \),

where \( N \) is the length of the audio signal, \( M \) the length of the RIR \( h \), and all \( x(n) \) with \( n < 0 \) are assumed to be zero.

In general, the RIR \( h \) depends on the size of the room, on the positions of the source and the receiver, and on other room characteristics such as the sound reflection properties of the walls, ceiling, floor, and of any furniture or other contents of the room. Therefore, the audio signal that is received by the ASR system is never identical to the original audio and an exact RIR is very hard to predict. We will describe a possible solution after the next section.

D. Psychoacoustics

Psychoacoustics has been shown to yield an appropriate measure of (in-)audibility for the calculation of audio adversarial examples \cite{3, 4}. Psychoacoustic hearing thresholds describe how the dependencies between frequencies and across time lead to masking effects in human perception \cite{11}. Probably the best-known example for an application of these effects is found in MP3 compression \cite{15}, where the compression algorithm applies a set of empirical hearing thresholds to the input signal. The original input signal can be transformed into a smaller but lossy representation by encoding the signal, dependent on hearing thresholds. More specifically, the imperceptible bands are encoded with a very low resolution in comparison to the perceptible bands.

In psychoacoustic adversarial examples, inspired by these methods, the psychoacoustic hearing thresholds are used to limit the changes in the audio signal to time-frequency-areas, where the noise is not, or barely, perceptible by humans.

III. OVER-THE-AIR ADVERSARIAL EXAMPLES

We extend the optimization algorithm of Schönherr et al. \cite{3} to produce robust audio adversarial examples, which still function as intended even after transmission from a loudspeaker to a microphone in a real room environment. For this purpose, we simulate different RIRs and use those in an iterative algorithm to harden the adversarial examples against the signal modifications that are incurred during playback and recording.

A. Adversary Model

Throughout the rest of this paper, we assume the following adversary model. First, we assume a white-box attack, where the adversary knows the attacked ASR system with all of its model parameters. This requirement is in line with prior work on this topic \cite{3}. Using this knowledge, the attacker generates audio samples containing malicious perturbations before the actual attack takes place, i.e., the attacker exploits the ASR system to create an audio file that produces the desired recognition result. Additionally, we assume that the trained ASR system, including the DNN, remains unchanged over time. Finally, we assume that the adversarial examples are played over the air via loudspeakers. Note that we only consider targeted attacks, where the target transcription is predefined (i.e., the adversary chooses the target sentence).

B. Room Impulse Response Simulator

For the RIR simulation, we use the AudioLabs implementation of the image method from Allen and Berkley \cite{7}. The simulator takes the room dimensions, the audio decay time \( T_{60} \), and the source and receiver position as its input and calculates the corresponding RIR for the given parameters.
Fig. 3. The room simulation model. We assume a probability distribution over all possible rooms by defining relevant simulation parameters like e.g. the room geometry, $T_{60}$ time due to numerical problems, source, and receiver positions as random variables. To optimize our over-the-air adversarial examples we sample from this distribution to get a variety of possible RIRs.

For the simulation, we use a cuboid-shaped room, which can be described via its length, width and height, $b = [b_x, b_y, b_z]$. In addition to this, the simulation also requires the three-dimensional source position $s = [s_x, s_y, s_z]$, receiver position $r = [r_x, r_y, r_z]$, and the audio decay time $T_{60}$, which results in 10 freely selectable parameters. All parameters are also sketched in Figure 3.

In order to sample random RIRs, we assume $b$, the audio decay time $T_{60}$, $s$, and $r$ to be random variables. We draw each of these 10 values from a uniform distribution between a minimum and a maximum allowed value. For the room size and for $T_{60}$, the minimum and the maximum values can be chosen arbitrarily. After those parameters are drawn, the ranges for source and receiver positions are drawn next, in a range from zero to the dimensions of the virtual room, to make sure that the source and the receiver are located inside the room. To simplify the notation, in the following, we use $b, s, r$, which additionally prepend a convolutional layer to the DNN. This layer simulates the convolution with the RIR $h$ to model the transmission through the room. We need this convolution to be integrated as a layer in the DNN in order to enable the algorithm to apply gradient descent to the time-domain audio signal (before playback), similar to the integration of the feature extraction into the DNN in prior work [3].

The RIR $h$ can, therefore, be considered as a sample of the distribution $H_b$. For some combinations of parameters, it is not possible to calculate an RIR with the chosen $T_{60}$. In those cases, we sample a new $T_{60}$ from the same distribution.

C. Robust Audio Adversarial Examples

In contrast to approaches that feed adversarial examples directly into the ASR system [1], [3], we include the presence of changing room characteristics, in the form of RIRs, in the optimization problem. This hardened the adversarial examples to remain robust in an over-the-air attack.

For the calculation of an adversarial example, we extend the optimization criterion given in (3) via

$$x' = \text{arg max}_x \mathbb{E}_{h \sim H_b} [P(y' | \hat{x}_h)].$$

This approach is borrowed from the Expectation Over Transformation (EOT) approach in the visual domain, where it is used to consider different two- and three-dimensional transformations, which has led to successful real-world adversarial examples [10]. In our case, instead of visual transformations, we use the convolution with RIRs, drawn from $H_b$, to maximize the expectation over varying RIRs as shown in Equation (6).

For the implementation, we use a DNN that already has been augmented to include the feature extraction, and we additionally prepend a convolutional layer to the DNN. This layer simulates the convolution with the RIR $h$ to model the transmission through the room. We need this convolution to be integrated as a layer in the DNN in order to enable the algorithm to apply gradient descent to the time-domain audio signal (before playback), similar to the integration of the feature extraction into the DNN in prior work [3].

The leftmost part, the RIR simulation layer, is only used for the calculation of adversarial examples and removed during testing, as the actual (either simulated or physical) RIR will then act during the (simulate or real) transmission over the air.

An overview of the proposed DNN is given in Figure 5 where the first part ('Convolution') describes the convolution with the RIR $h$ and the center and right part ('Feature extraction' and 'DNN') show the feature extraction and the acoustic model DNN, which is used to obtain the pseudo-posteriors for the decoding stage.

The inclusion of the convolution as a layer in the DNN requires the layer to be differentiable. Using [5], the derivative can be written as

$$\frac{\partial x_h(n)}{\partial x(m)} = h(n - m) \quad \forall n, m.$$
Feature waveform depends on the distribution and recording conditions. Therefore, the adversarial example gradient descent iterations. This simulates different rooms defined in (6), by sampling a new RIR and the recorded audio signals must be used to calculate adversarial examples, we used a measured RIR simulated RIR and to update the time domain audio signal created before the maximum number of iterations is reached, most iterations. However, if a robust adversarial example is created before the maximum number of iterations is reached, the algorithm does not need to continue. To efficiently calculate adversarial examples, we used a measured RIR to simulate the over-the-air scenario during the calculation to verify whether the example has already achieved over-the-air robustness. The RIR was only used for the verification and it was not applied during the gradient descent step, to not adjust the adversarial examples to the real RIR.

The entire approach is given in Algorithm 1. As can be seen, the psychoacoustic hearing thresholds are also used during the gradient descent to limit the modifications of the signal to those time-frequency ranges, where they are (mostly) imperceptible. \(D \cdot N_0\) describes the augmented DNN (feature extraction and acoustic DNN) in Figure 2 without the RIR simulation since, for the algorithm, this is replaced by the simulated RIR \(h_{test}\).

### IV. Experimental Evaluation

We measured the performance of the algorithm for simulated and real over-the-air attacks, both for unrestricted adversarial examples and for adversarial examples restricted by the psychoacoustic hearing thresholds.

#### A. Metrics

We used the following standard measures to assess the quality of the robust adversarial examples.

1. **Word Error Rate (WER):** To measure performance, we used the word error rate (WER) with respect to the target transcription. The standard metric for this purpose, the Levenshtein distance \([14]\) \(\mathcal{L}\), is used here. It counts the number of deleted \(D\), inserted \(I\), and substituted \(S\) words, and it is divided by the total number of words \(N\) to obtain

\[
WER = 100 \cdot \frac{\mathcal{L}}{N} = 100 \cdot \frac{D + I + S}{N}.
\]

For a real attack, an adversarial example can only be counted as a success if a WER of 0% is achieved. Therefore, we also determined the number of adversarial examples that had been decoded without any errors.
2) **Segmental Signal-to-Noise Ratio:** The segmental signal-to-noise ratio (SNRseg) measures the amount of noise \( \sigma \) added to the original signal \( x \) and is computed via

\[
\text{SNRseg}(\text{dB}) = 10 \frac{1}{K} \sum_{k=0}^{K-1} \log_{10} \left( \frac{\sum_{t=T_k+1}^{T_k+T-1} x^2(t)}{\sum_{t=T_k}^{T_k+T-1} \sigma^2(t)} \right)
\]

(13)

where \( T \) is the segment length and \( K \) the number of segments. Thus, the higher the SNRseg, the less noise was added.

In contrast to the signal-to-noise ratio (SNR), the SNRseg \(^{15}\) is performed frame-wise and gives a better assessment of an audio signal if the original and the added noise are aligned \(^{15}\).

We use a window length of 16 ms which corresponds to \( K = 256 \) at a sampling frequency of 16 kHz.

### B. Evaluation

For the evaluation, we used 100 audio samples, all containing eight seconds of music. For this, we compared different lengths \( M \) for the RIR (512 and 1024 samples) and evaluated the effect of the use of hearing thresholds and of different room configurations. For the hearing thresholds, we employed the most promising version of prior attacks \(^{3}\), which allows the modifications to the signal to exceed the exact hearing thresholds by 20 dB. For the further parameters, we set \( G = 2000 \) and \( Q = 10 \). All experiments were performed on a machine with two Intel Xeon E5-2670 v3 CPUs and 128 GB of DDR4 memory.

### C. Simulated Over-the-Air Attack

The results in Table 1 are calculated with Algorithm 1. The WER is measured for a simulated over-the-air attack using a measured RIR. For this, we used two different distributions \( H \). The approximate dimensions of the real room that was used in the evaluation, are \( b_{\text{real}} \approx [8 \text{ m}, 7 \text{ m}, 2.8 \text{ m}] \) with \( T_{60,\text{real}} \approx 0.3 \text{ ms} \). For the two sampling distributions of the simulated rooms, \( H_{\theta_1} \) and \( H_{\theta_2} \), the parameters are shown in Table 1; \( H_{\theta_1} \) simulates a tighter approximation of the actual room dimensions and \( H_{\theta_2} \) a broader one. We use these different versions to evaluate how tight the room needs to be approximated for a successful over-the-air attack.

Note that even if the WER seems to be high, for an attacker one successful adversarial example with 0% WER would be enough, which is indeed possible, also for real over-the-air examples, which will be shown in the next section.

---

**TABLE I**

**Table of Room Dimensions for the Sampling Distributions \( H_{\theta_1} \) and \( H_{\theta_2} \).**

| \( H_{\theta_1} \) | \( H_{\theta_2} \) |
|------------------|------------------|
| \( b_x \) | \( 6.0 \text{ m} \) | \( 7.0 \text{ m} \) |
| \( b_y \) | \( 5.0 \text{ m} \) | \( 6.0 \text{ m} \) |
| \( b_z \) | \( 3.0 \text{ m} \) | \( 2.5 \text{ m} \) |
| \( T_{60} \) | \( 0.2 \text{ s} \) | \( 0.2 \text{ s} \) |

**TABLE II**

**Table of WER and SNRseg for Simulated Over-the-Air Attacks for Different Room Configurations \( (H_{\theta_1} \text{ and } H_{\theta_2}) \).** \( M \) is the length of the simulated RIRs, results with and without psychoacoustic masking are shown for \( G = 2000 \).

| \( M \) | \( WER \) | \( \text{SNRseg in dB} \) | \( WER \) | \( \text{SNRseg in dB} \) |
|--------|--------|-----------------|--------|-----------------|
| 512    | \( H_{\theta_1} \) w/o thresh. | 66.7% | 12.97±3.91 | \( H_{\theta_2} \) w/o thresh. | 66.7% | 13.28±4.10 |
|        | w/ thresh. | 82.0% | 15.72±4.25 |                    | 76.7% | 16.10±4.01 |
| 1024   | \( H_{\theta_1} \) w/o thresh. | 68.7% | 12.89±4.30 | \( H_{\theta_2} \) w/o thresh. | 66.7% | 13.28±4.35 |
|        | w/ thresh. | 82.3% | 16.03±4.35 |                    | 79.0% | 16.21±4.28 |

The SNRseg in Table II is calculated after applying the same measured RIR to both the original signal and the adversarial examples. We chose this approach since this is also the signal that is perceived by human listeners if the adversarial examples are played over the air.

The SNRseg appears fairly low in general, which indicates that on average more noise has to be added in comparison to adversarial examples which are not hardened to work over the air. However, the adversarial examples that were restricted by the hearing thresholds have a better SNRseg. Also, in the successful cases of real over-the-air adversarial examples discussed in the next section, those examples, which were restricted via the hearing thresholds, show a better SNRseg. Additionally, as SNRseg measures any added noise, and not only the perceptible noise components, the perceptible noise is even lower than SNRseg would suggest for the versions where hearing thresholds are used.

In Figure 7, the WER is plotted as a function of the number of iterations \( G \). Even though the WER may further improve in some cases, we limited the maximum number of iterations to 2000 due to computational reasons. Also, in most
of the cases, after 1000 iterations, the WER did not decrease nearly as rapidly as for the first 1000 iterations.

D. Real End-to-End Over-the-Air Attack

The results for real over-the-air attacks are shown in Table III. These were obtained using the same adversarial examples that were evaluated for the simulated playback in Table I but after these were replayed in the actual lab room setup described above, with $s_{real} = [3.9 \, m, 3.4 \, m, 1.2 \, m]$ and $r_{real} = [1.4 \, m, 1.8 \, m, 1.2 \, m]$. A 360 degree panorama shot of the lab room setup is shown in Figure 6 with the position of the microphone marked by the green dashed circle and the position of the loudspeaker by the solid red circle.

Some of the 100 considered audio samples, especially without the use of hearing thresholds, clipped too much. As it would not be possible to replay those examples, we removed the adversarial examples for the actual over-the-air attack. Each of the remaining adversarial examples was played back five times. We were also possible to transcribe adversarial examples with 0% WER. Depending on the setup, the measured success rate was up to 3.7% of the utterances for the cases with hearing thresholds ($H_{\theta_1}$ and $M = 1024$) and up to 1.8% of the utterances without hearing thresholds ($H_{\theta_2}$ and $M = 512$).

The comparison of the WERs in Table II and Table III shows that the simulated attack gives a sufficient prediction of the expected WER. The largest difference of the WER between the simulated and the real case is 6.2% ($H_{\theta_1}$ and $M = 512$ with hearing thresholds), and on average, the difference is only 3.2%. Also, for the adversarial examples with 0% WER in the simulated attacks the real attacks are more likely to be successful. This shows, that the simulation is a reliable predictor for how successful an attack would be in the real world. Some of those examples are shown at [http://ota-adversarial-examples.selfip.org](http://ota-adversarial-examples.selfip.org).

E. Varying Room Setups

To measure the robustness for cases where the room characteristics differ from the simulated room setup—which is also used to calculate the robust adversarial examples—we used the best versions with and without hearing thresholds and replayed those in different real room setups. For these experiments, we used the setup with $M = 1024$, which led to the best WER in the real over-the-air attack in Table III. The results are shown in Table IV.

The experiments in Room I and Room II are performed in the same room as the experiments shown in Table III but with a modified $T_{60}$ time for Room I, and a changed receiver and source position for Room II, with $s_{roomI} = [1.9 \, m, 4.6 \, m, 1.2 \, m]$ and $r_{roomI} = [5.2 \, m, 3.4 \, m, 1.2 \, m]$. The results for Room II should therefore not significantly change since the algorithm already varies the position on the source but with a modified $T_{60}$ time for Room I, and a changed receiver and source position for Room II, with $s_{roomII} = [1.9 \, m, 4.6 \, m, 1.2 \, m]$ and $r_{roomII} = [5.2 \, m, 3.4 \, m, 1.2 \, m]$. The results for Room II should therefore not significantly change since the algorithm already varies the position on the source and the receiver. This is indeed the case for $\theta_1$, but for $\theta_2$ the WER increases. Nevertheless, the results are still reasonable and the total number of successful trials is still similar.

In general, the results remain equal for all cases and as expected, for $\theta_1$, the results are more robust against varying conditions as this is the less specifically tailored version.

| $H_{\theta_1}$ | $M = 512$ WER | $M = 1024$ WER |
|----------------|----------------|----------------|
| w/o thresh.    | 71.1%          | 72.7%          |
| w/ thresh.     | 88.2%          | 82.4%          |
| $H_{\theta_2}$ | $M = 512$ WER  | $M = 1024$ WER |
| w/o thresh.    | 70.7%          | 66.3%          |
| w/ thresh.     | 81.0%          | 81.0%          |

| Room I | $M = 512$ WER | $M = 1024$ WER |
|--------|----------------|----------------|
| $H_{\theta_1}$ | w/o thresh. | 78.0% | 72.0% |
| w/ thresh.     | 89.7%          | 82.5%          |
| $H_{\theta_2}$ | w/o thresh. | 80.0% | 74.6% |
| w/ thresh.     | 87.2%          | 87.8%          |
V. Related Work

Adversarial attacks on ASR systems focus either on hiding a target transcription [5], [6] or on obfuscating the original transcription [17]. Almost all previous works on attacks against ASR systems did not focus on real-world attacks [5], [18], [19] or were only successful for simulated over-the-air attacks [4].

A. Audio Adversarial Example

Carlini et al. have shown that targeted attacks against HMM-only ASR systems are possible [5]. They use an inverse feature extraction to create adversarial audio samples. However, the resulting audio samples are not intelligible by humans in most of the cases and may be considered as noise, but may make thoughtful listeners suspicious.

A different approach was shown by Vaidya et al. [19], where the authors changed an input signal to fit the target transcription by considering the features instead of the output of the DNN. Nevertheless, the results show high distortions of the audio signal and can easily be detected by a human.

An approach to overcome this limitation was proposed by Zhang et al. They have shown that an adversary can hide a transcription by utilizing non-linearities of microphones to modulate the baseband audio signals with ultrasound above 20 kHz [18]. The main downside of the attack is the fact that the information of the necessary features needs to be retrieved from the audio signal, recorded with the specific microphone, which is costly in practice. Song and Mittael [20] and Roy et al. [21] introduced similar ultrasound-based attacks that are not adversarial examples, but rather interact with the ASR system in a frequency range inaudible to humans.

Carlini and Wagner published a work in which they introduce a general targeted attack on ASR systems using CTC-loss [1]. The attack is based on a gradient-descent-based minimization [22] (as used in previous image classification adversarial attacks), but the adversarial examples are fed directly into the recognizer. CommanderSong [2] is also evaluated against Kaldi and uses backpropagation to find an adversarial example. However, the very limited over-the-air attack highly depends on the speakers and recording devices as the attack parameters have to be adjusted especially for these components. Yakura and Sakuma published a technical report, which describes an algorithm to create over-the-air robust adversarial examples, but with the limitation that it is necessary to have physical access to the room where the attack takes place [8]. Also, they did not evaluate their room-dependent results for varying room conditions. Concurrently, Szuley and Kolter also published a work on room-dependent robust adversarial examples, which worked under constraints given by a psychoacoustic model [9]. However, their adversarial examples have only worked in an anechoic chamber, a room that is designed specifically to eliminate the effect of an RIR. The attack can, therefore, not be compared with a real-world scenario as the audio signal is limited almost completely to the direct sound.

B. Psychoacoustic Hiding

Schönherr et al. published an approach where psychoacoustic modeling, borrowed from the MP3 compression algorithm, was used to re-shape the perturbations of the adversarial examples in such a way as to hide the changes below the human hearing thresholds [3]. However, the adversarial examples that are created in that work need to be fed into the recognizer directly, hence no end-to-end attack in an over-the-air setting was possible.

Simultaneously, Abdullah et al. showed a black-box attack in which psychoacoustics is used to empirically calculate adversarial examples [6]. Their approach focuses on over-the-air attacks, but in many cases, humans can perceive the hidden message once they are alerted to its content.

As an extension of Carlini’s and Wagner’s attack [1], Qin et al. introduced the first implementation of RIR-independent adversarial examples [4]. Unfortunately, their approach only worked in a simulated environment and not for real over-the-air attacks, but the authors also utilize psychoacoustics to limit the perturbations.

Our approach is the first targeted attack that focuses on RIR-independent robust adversarial examples and we demonstrate how to generate adversarial examples which appear to be mostly unaffected by the environment, as ascertained by verifying their success in a broad range of room characteristics. We utilize psychoacoustics to limit the perturbations of the audio signal to remain under, or at least close to, the human thresholds of hearing, and we show that the examples remain robust to playback over the air.

C. Robust Adversarial Examples in the Visual Domain

In the visual domain, Evtimov et al. showed one of the first real-world adversarial attacks [23]. They created and printed stickers, which can be used to obfuscate traffic signs. For humans, the stickers are visible. However, they seem very inconspicuous and could possibly fool autonomous cars.

Athalye and Sutskever presented another real-world adversarial perturbation on a 3D-printed turtle, which is recognized as a rifle from almost every point of view [10]. The algorithm to create this 3D object not only minimizes the distortion for one image, but for all possible projections of a 3D object into a 2D image.

In contrast to the visual domain, audio adversarial examples are time-dependent and need to be considered as time series signals, whereas images do not change over time, which makes the calculation of adversarial audio signals algorithmically more challenging. Our approach is capable of successfully performing such an attack.

VI. Discussion

The above results show that the SNRseg for the proposed attack is lower in comparison to adversarial attacks that are not hardened to work over the air. But while the mean SNRseg is lower, we have been able to successfully create adversarial examples which have SNRsegs beyond 14 dB for a real over-the-air attack.
For an attacker, one successful adversarial example is enough. Therefore, even if the WER are high, we have shown that it is possible to create adversarial examples, which remain robust after being replayed (WER of 0%), with and without restrictions via hearing thresholds.

In general, the results show a trade-off between the WER and the SNRseg: if no hearing thresholds are used, the WER, in general, is significantly better in comparison to the versions with hearing thresholds. However, on the other hand, also if the WER is better in cases where no hearing thresholds are used, we have shown that it is indeed possible to calculate over-the-air-robust adversarial examples with hearing thresholds. Those adversarial examples contain less perceptible noise and are, therefore, less likely to be detected by human listeners. As another advantage, for the adversarial examples with hearing thresholds, fewer examples had to be discarded due to artifacts.

The experiments in Table III show that the WER is the better if the actual room geometry and the audio decay time $T_{00}$ is known. Specifically, the adversarial examples with $\theta_2$, which uses a smaller range for the parameter distributions, have a better WER. Therefore, the attack algorithm apparently tailors the adversarial examples to the real room parameters.

On the other hand, the adversarial examples that are computed for a specific room setup will likely be less successful, if the properties of the room have changed. This can be observed in Table IV where we changed the room for the over-the-air attack. As a result, the setup with $\theta_2$ was more successful in Table III but is outperformed by the $\theta_1$ setup in Table IV in most of the cases.

The computation of robust adversarial examples is costly in comparison to approaches, which are not tailored to work over the air. This is mainly due to the fact that more iterations are required because of the changing RIRs. However, by choosing appropriate hyper-parameters, the run-time can be reduced, e.g., by reducing the length $M$ of the RIRs. The length $M$ barely affects the WER or the SNRseg, even when the results with $M = 1024$ require a longer computation time.

Plotting the WER as a function of the number of iterations shows that the WER might even decrease further after 2000 iterations. However, Figure 7 also indicates that due to the long calculation time, it is not efficient to run for more iterations. Instead one should try more audio files to obtain a specific target transcription.

In a black-box scenario, the attacker has no access to the ASR system. However, even for this more challenging attack, it has been shown that it is possible to calculate adversarial examples, but with the caveat that humans can perceive the hidden transcription if they get aware of it. The proposed approach is not easy to apply to black-box adversarial examples. However, it is feasible to use a similar approach in combination with a parameter-stealing attack [24]–[28]. Once the attacker is able to rebuild her own system, which reassembles the black-box system, the proposed algorithm can be used with that system as well.

To prevent such an attack, an ASR system needs either some kind of detection mechanism, or it needs a recognition that is robust to adversarial examples. The detection of adversarial examples for known attacks might be feasible. However, it is not guaranteed that the detection will also work for new approaches. Therefore, in the long term, it makes sense to build the ASR system in such a way as to be adversarial-example-robust, e.g., by mimicking the human perception of speech similar to images encoded in jpeg format [29]. One step in this direction can be to focus the ASR to only those signal components that are perceptible to the human listener as well, similar to the MP3 encoding. However, even in that case, it will still be possible to create adversarial examples, with the limitation that the perturbations are moved to human perceptible areas of the audio.

Therefore, not only the input data should be considered, but also the ASR, e.g., the DNN, itself. In the visual domain, a first adversarial-example-robust recognizer has been proposed and evaluated for MNIST data [30]. The approach synthesizes each possible class of the output in respect to the input data and then decides dependent on the synthesized versions. This makes it very hard (or even impossible) to embed adversarial distortions, as the input data will be replaced by a general representation of the class.

VII. CONCLUSION

In this paper, we have demonstrated that ASR systems are vulnerable against adversarial examples which are replayed over the air. To this end, we have shown that it is possible to calculate adversarial examples with a generic algorithm. Compared to prior work on this topic, which used a fixed setup only, our approach takes the characteristics of the room and the position of the microphone and the loudspeaker into account. By simulating varying RIRs during the calculation of adversarial examples, we can create robust adversarial examples, which can be played over the air. The examples can be tailored to specific rooms, but also work, if a more general setup is used or the room situation does change. To substantiate our approach, we performed end-to-end attacks against Kaldi, which uses a state-of-the-art DNN-HMM system and presented the results of empirical attacks for different room configurations. The algorithm can be used with and without hearing thresholds, which limit the perturbations to be less perceptible by humans. In both cases, we have shown that it is possible to calculate robust adversarial examples.

Future work should investigate possible countermeasures such as using only the perceptible areas of the audio or use the DNN recognition itself, to be hardened against adversarial examples.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) within the framework of the Excellence Strategy of the Federal Government and the States EXC 2092 CASA – 390781972.
