Dompteur: Taming Audio Adversarial Examples

Thorsten Eisenhofer, Lea Schönherr, and Joel Frank, *Ruhr University Bochum;*
Lars Speckemeier, *University College London;* Dorothea Kolossa and
Thorsten Holz, *Ruhr University Bochum*

This paper is included in the Proceedings of the 30th USENIX Security Symposium.

August 11–13, 2021

978-1-939133-24-3

DOMPTEUR: Taming Audio Adversarial Examples

Thorsten Eisenhofer
Ruhr University Bochum

Lea Schönherr
Ruhr University Bochum

Joel Frank
Ruhr University Bochum

Lars Speckemeier
University College London

Dorothea Kolossa
Ruhr University Bochum

Thorsten Holz
Ruhr University Bochum

Abstract

Adversarial examples seem to be inevitable. These specifically crafted inputs allow attackers to arbitrarily manipulate machine learning systems. Even worse, they often seem harmless to human observers. In our digital society, this poses a significant threat. For example, *Automatic Speech Recognition* (ASR) systems, which serve as hands-free interfaces to many kinds of systems, can be attacked with inputs incomprehensible for human listeners. The research community has unsuccessfully tried several approaches to tackle this problem.

In this paper we propose a different perspective: We accept the presence of adversarial examples against ASR systems, but we require them to be *perceivable* by human listeners. By applying the principles of *psychoacoustics*, we can remove semantically irrelevant information from the ASR input and train a model that resembles human perception more closely. We implement our idea in a tool named DOMPTEUR[1] and demonstrate that our augmented system, in contrast to an unmodified baseline, successfully focuses on perceptible ranges of the input signal. This change forces adversarial examples into the audible range, while using minimal computational overhead and preserving benign performance. To evaluate our approach, we construct an *adaptive attacker* that actively tries to avoid our augmentations and demonstrate that adversarial examples from this attacker remain clearly perceivable. Finally, we substantiate our claims by performing a hearing test with crowd-sourced human listeners.

1 Introduction

The advent of deep learning has changed our digital society. Starting from simple recommendation techniques [1] or image recognition applications [2], machine-learning systems have evolved to solve and play games on par with humans [3–6], to predict protein structures [7], identify faces [8], or recognize speech at the level of human listeners [9]. These systems are now virtually ubiquitous and are being granted access to

---

[1]The French word for *tamer*

critical and sensitive parts of our daily lives. They serve as our personal assistants [10], unlock our smart homes' doors [11], or drive our autonomous cars [12].

Given these circumstances, the discovery of *adversarial examples* [13] has had a shattering impact. These specifically crafted inputs can completely mislead machine learning-based systems. Mainly studied for image recognition [13], in this work, we study how adversarial examples can affect *Automatic Speech Recognition* (ASR) systems. Preliminary research has already transferred adversarial attacks to the audio domain [14–19]. The most advanced attacks start from a harmless input signal and change the model's prediction towards a target transcription while simultaneously *hiding* their malicious intent in the inaudible audio spectrum.

To address such attacks, the research community has developed various defense mechanisms [20–25]. All of the proposed defenses—in the ever-lasting cat-and-mouse game between attackers and defenders—have subsequently been broken [26]. Recently, Shamir et al. [27] even demonstrated that, given certain constraints, we can expect to always find adversarial examples for our models.

Considering these circumstances, we ask the following research question: *When we accept that adversarial examples exist, what else can we do?* We propose a paradigm shift: Instead of preventing *all* adversarial examples, we accept the presence of *some*, but we want them to be audibly changed.

To achieve this shift, we take inspiration from the machine learning community, which sheds a different light on adversarial examples: Illyas et al. [28] interpret the presence of adversarial examples as a disconnection between human expectations and the reality of a mathematical function trained to minimize an objective. We tend to think that machine learning models must learn meaningful features, e.g., a cat has paws. However, this is a human's perspective on what makes a cat a cat. Machine learning systems instead use *any* available feature they can incorporate in their decision process. Consequently, Illyas et al. demonstrate that image classifiers utilize so-called *brittle features*, which are highly predictive, yet not recognizable by humans.

Recognizing this mismatch between human expectations and the inner workings of machine learning systems, we propose a novel design principle for ASR system inspired by the human auditory system. Our approach is based on two key insights: (i) the human voice frequency is limited to the band ranges of approximately $300 - 5000 Hz$ [29], while ASR systems are typically trained on 16kHz signals, which range from $0 - 8000 Hz$, and (ii) audio signal can carry information, inaudible to humans [15]. Given these insights, we modify the ASR system by restricting its access to frequencies and applying psychoacoustic modeling to remove *inaudible* ranges. The effects are twofold: The ASR system can learn a better approximation of the human perception during training (i.e., discarding unnecessary information), while simultaneously, adversaries are forced to place any adversarial perturbation into audible ranges.

We implement these principles in a prototype we call DOMPTEUR. In a series of experiments, we demonstrate that our prototype more closely models the human auditory system. More specifically, we successfully show that our ASR system, in contrast to an unmodified baseline, focuses on perceptible ranges of the audio signal. Following Carlini et al. [30], we depart from the lab settings predominantly studied in prior work: We assume a white-box attacker with real-world capabilities, i.e., we grant them full knowledge of the system and they can introduce an unbounded amount of perturbations. Even under these conditions, we are able to force the attacker to produce adversarial examples with an average of 24.33 dB of added perturbations while remaining accurate for benign inputs. Additionally, we conduct a large scale user study with 355 participants. The study confirms that the adversarial examples constructed for DOMPTEUR are easily distinguishable from benign audio samples and adversarial examples constructed for the baseline system.

In summary, we make the following key contributions:

- **Constructing an Augmented ASR.** We utilize our key insights to bring ASR systems in better alignment with human expectations and demonstrate that traditional ASR systems indeed utilize non-audible signals that are not recognizable by humans.

- **Evaluation Against Adaptive Attacker.** We construct a realistic scenario where the attacker can adapt to the augmented system. We show that we successfully force the attacker into the audible range, causing an average of 24.33 dB added noise to the adversarial examples. We could not find adversarial examples when applying very aggressive filtering; however, this causes a drop in the benign performance.

- **User Study.** To study the auditory quality of adversarial examples, we perform a user study with an extensive crowd-sourced listening test. Our results demonstrate

that the adversarial examples against our system are significantly more perceptible by humans.

To support further research in this area, we open-source our prototype implementation, our pre-trained models, and audio samples online at github.com/rub-syssec/dompteur.
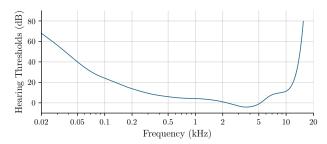
2   Technical Background

In the following, we discuss the background necessary to understand our augmentation of the ASR system. For this purpose, we briefly introduce the fundamental concepts of ASRs and give an overview of adversarial examples. Since our approach fundamentally relies on psychoacoustic modeling, we also explain masking effects in human perception.

Speech Recognition   ASR constitutes the computational core of today's voice interfaces. Given an audio signal, the task of an ASR system is to transcribe any spoken content automatically. For this purpose, traditionally, purely statistical models were used. They now have been replaced by modern systems based on deep learning methods [31–33], often in the form of hybrid neural/statistical models [34].
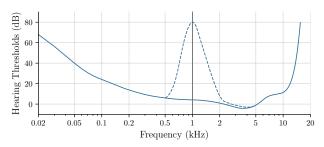
In this paper, we consider the open-source toolkit KALDI [35] as an example of such a modern hybrid system. Its high performance on many benchmark tasks has led to its broad use throughout the research community as well as in commercial products like e. g., Amazon's Alexa [36–38].

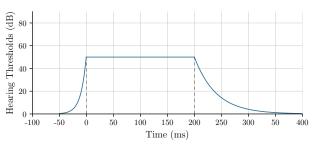KALDI, and similar DNN/HMM hybrid systems can generally be described as three-stage systems:

1. *Feature Extraction.* For the feature extraction, a framewise *discrete Fourier transform* (DFT) is performed on the raw audio data to retrieve a frequency representation of the input signal. The input features of the *Deep Neural Networks* (DNN) are often given by the log-scaled magnitudes of the DFT-transformed signal.

2. *Acoustic Model DNN.* The DNN acts as the *acoustic model* of the ASR system. It calculates the probabilities for each of the distinct speech sounds (called *phones*) of its trained language being present in each time frame from its DFT input features. Alternatively, it may compute probabilities, not of phones, but of so-called *clustered tri-phones* or, more generally, of data-driven units termed *senones*.

3. *Decoding.* The output matrix of the DNN is used together with an *hidden Markov model* (HMM)-based language model to find the most likely sequence of words, i. e., the most probable transcription. For this purpose, a dynamic programming algorithm, e.g., Viterbi decoding, is used to search the best path through the underlying HMM. The language model describes the probabilities of word sequences, and the acoustic model output gives the probability of being in each HMM state at each time.

(a) Absolute Hearing Thresholds

(b) Frequency Masking

(c) Temporal Masking

Figure 1: **Psychoacoustic allows to describe limitations of the human auditory system.** Figure 1a shows the average human hearing threshold in quiet. Figure 1b shows an example of masking, illustrating how a loud tone at 1kHz shifts the hearing thresholds of nearby frequencies and Figure 1c shows how the recovery time of the auditory system after processing a loud signal leads to temporal masking.

Psychoacoustic Modeling Recent attacks against ASR systems exploit intrinsics of the human auditory system to make adversarial examples less conspicuous [17, 39–41]. Specifically, these attacks utilize limitations of human perception to hide modifications of the input audio signal within inaudible ranges. We use the same effects for our approach to *remove* inaudible components from the input:

- *Absolute Hearing Threshold.* Human listeners can only perceive sounds in a limited frequency range, which diminishes with age. Moreover, for each frequency, the sound pressure is important to determine whether the signal component is in the audible range for humans. Mea-

suring the *hearing thresholds*, i. e., the necessary sound pressures for each frequency to be audible in otherwise quiet environments, one can determine the so-called *absolute hearing threshold* as depicted in Figure 1a. Generally speaking, everything above the *absolute hearing thresholds* is perceptible in principle by humans, which is not the case for the area under the curve. As can be seen, much more energy is required for a signal to be perceived at the lower and higher frequencies. Note that the described thresholds only hold for cases where no other sound is present.

- *Frequency Masking.* The presence of another sound— a so-called *masking tone*—can change the described *hearing thresholds* to cover a larger area. This *masking effect* of the masking tone depends on its sound pressure and frequency. Figure 1b shows an example of a 1 kHz masking tone, with its induced changes of the *hearing thresholds* indicated by the dashed line.

- *Temporal Masking.* Like frequency masking, temporal masking is also caused by other sounds, but these sounds have the same frequency as the masked tone and are close to it in the time domain, as shown in Figure 1c. Its root cause lies in the fact that the auditory system needs a certain amount of time, in the range of a few hundreds of milliseconds, to recover after processing a higher-energy sound event to be able to perceive a new, less energetic sound. Interestingly, this effect does not only occur at the end of a sound but also, although much less distinct, at the beginning of a sound. This seeming causal contradiction can be explained by the processing of the sound in the human auditory system.

Adversarial Examples Since the seminal papers by Szegedy et al. [13] and Biggio et al. [42], a field of research has formed around adversarial examples. The basic idea is simple: An attacker starts with a valid input to a machine learning system. Then, they add small perturbations to that input with the ultimate goal of changing the resulting prediction (or in our case, the transcription of the ASR).

More formally, given a machine learning model f and an input-prediction pair $\langle x, y \rangle$, where $f(x) = y$, we want to find a small perturbation δ s.t.:

$$x' = x + \delta \quad \wedge \quad f(x') \neq f(x).$$

In this paper, we consider a stronger type of attack, a targeted one. This has two reasons: the first is that an untargeted attack in the audio domain is fairly easy to achieve. The second is that a targeted attack provides a far more appealing (and thus, far more threatening) real-life use case for adversarial examples. More formally, the attacker wants to perturb an input phrase x (i.e., an audio signal) with a transcription y (e.g., "Play the Beatles") in such a way that the ASR transcribes

an attacker-chosen transcription y' (e.g., "Unlock the front door"). This can be achieved by computing an adversarial example x' based on a small adversarial perturbation δ s.t.:

$$x' = x + \delta \quad \wedge \quad ASR(x') = y' \quad \wedge \quad y \neq y'. \quad (1)$$

There exist a multitude of techniques for creating such adversarial examples. We use the method introduced by Schönherr et al. [17] for our evaluation in Section 4. The method can be divided into three parts: In a first step, attackers choose a fixed output matrix of the DNN to maximize the probability of obtaining their desired transcription y'. As introduced before, this matrix is used in the ASR system's decoding step to obtain the final transcription. They then utilize gradient descent to perturb a starting input x (i.e., an audio signal feed into the DNN), to obtain a new input x', which produces the desired matrix. This approach is generally chosen in white-box attacks [16, 18]. Note that we omit the feature extraction part of the ASR; however, Schönherr et al. have shown that this part can be integrated into the gradient step itself [17]. A third (optional) step is to utilize psychoacoustic hearing thresholds to restrict the added perturbations to inaudible frequency ranges. More technical details can be found in the original publication [17].

3 Modeling the Human Auditory System

We now motivate and explain our design to better align the ASR system with human perception. Our approach is based on the fact that the human auditory system only uses a subset of the information contained in an audio signal to form an understanding of its content. In contrast, ASR systems are not limited to specific input ranges and utilize every available signal – even those *inaudible* for the human auditory system. Consequently, an attacker can easily hide changes within those ranges. Intuitively, the smaller the overlap between these two worlds, the harder it becomes for an attacker to add malicious perturbations that are inaudible to a human listener. This is akin to reducing the attack surface in traditional systems security.

To tackle these issues, we leverage the following two design principles in our approach:

(i) *Removing inaudible parts:* As discussed in Section 2, audio signals typically carry information imperceptible to human listeners. Thus, before passing the input to the network, we utilize psychoacoustic modeling to remove these parts.

(ii) *Restricting frequency access:* The human voice frequency range is limited to a band of approximately $300 - 5000\,Hz$ [29]. Thus, we implement a band-pass filter between the feature extraction and model stage (cf. Section 2) to restrict the acoustic model to the appropriate frequencies.

3.1 Implementation

In the following, we present an overview of the implementation of our proposed augmentations. We extend the state-of-the-art ASR toolkit KALDI with our augmentations to build a prototype implementation called DOMPTEUR. Note that our proposed methods are universal and can be applied to *any* ASR system.

Psychoacoustic Filtering Based on the psychoacoustic model of MPEG-1 [43], we use psychoacoustic hearing thresholds to remove parts of the audio that are not perceivable to humans. These thresholds define how dependencies between certain frequencies can mask, i.e., make inaudible, other parts of an audio signal. Intuitively, these parts of the signal should not contribute any information to the recognizer. They do, however, provide space for an attacker to hide adversarial noise.

We compare the absolute values of the complex valued *short-time Fourier transform* (STFT) representation of the audio signal $\mathbf{S}$ with the hearing thresholds $\mathbf{H}$ and define a mask via

$$\mathbf{M}(n,k) = \begin{cases} 0 & \text{if } \mathbf{S}(n,k) \leq \mathbf{H}(n,k) + \Phi, \\ 1 & \text{else} \end{cases}, \quad (2)$$

with $n = 0, \ldots, N-1$ and $k = 0, \ldots, K-1$. We use the parameter Φ to control the effect of the hearing thresholds. For $\Phi = 0$, we use the original hearing threshold, for higher values we use a more aggressive filtering, and for smaller values we retain more from the original signal. We explore this in detail in Section 4. We then multiply all values of the signal $\mathbf{S}$ with the mask $\mathbf{M}$

$$\mathbf{T} = \mathbf{S} \odot \mathbf{M}, \quad (3)$$

to obtain the filtered signal $\mathbf{T}$.

Band-Pass Filter High and low frequencies are not part of human speech and do not contribute significant information. Yet, they can again provide space for an attacker to hide adversarial noise. For this reason, we remove low and high frequencies of the audio signal in the frequency domain. We apply a band-pass filter after the feature extraction of the system by discarding those frequencies that are smaller or larger than certain thresholds (the so-called cut-off frequencies). Formally, the filtering can be described via

$$\mathbf{T}(n,k) = 0 \quad \forall f_{\max} < k < f_{\min}, \quad (4)$$

where $f_{\max}$ and $f_{\min}$ describe the lower and the upper cut-off frequencies of the band-pass.

3.2 Attacker Model

While some of our augmentations improve the ASR system's overall performance, we are specifically interested in its performance against adversarial perturbations. To achieve any meaningful results, we believe the attacker needs to have *complete* control over the input. Following guidelines recently established by Carlini et al. [30], we embark from theoretical attack vectors towards the definition of a realistic threat model, capturing real-world capabilities of attackers.

The key underlying insight is that the amount of perturbations caused by a real-world attack cannot be limited. This is easy to see: in the worst case, the attacker can always force the target output by replacing the input with the corresponding audio command. Note that this, in turn, implies that we cannot completely prevent adversarial attacks *without* also restricting benign inputs.

We can also not rely on obfuscation. Previous works have successfully shown so-called parameter-stealing attacks, which build an approximation of a black-box system [44–48]. Since an attacker has full control over this approximated model, they can utilize powerful white-box attacks against it, which transfer to the black-box model.

In summary, we use the following attacker model:

- *Attacker Knowledge:* Following Kerckhoffs' principle [49], we consider a *white-box* scenario, where the attacker has complete knowledge of the system, including all model parameters, training data, etc.

- *Attacker Goals:* To maximize practical impact, we assume a targeted attack, i. e., the attacker attempts to perturb a given input x to fool a speech recognition system into outputting a false, *attacker-controlled* target transcription y' based on Equation (1).

- *Attacker Capabilities:* The attacker is granted complete control over the input, and we explicitly do not restrict them in any way on how δ should be crafted. Note, however, that δ is commonly minimized during computation according to some distance metric. For example, by measuring the *perceived* noise, an attacker might try to minimize the conspicuousness of their attack [17].

We choose this attacker model with the following in mind: We aim to limit the attacker, not in the amount of applied perturbations, but rather confine the nature of perturbations themselves. In particular, we want adversarial perturbations to be clearly perceptible by humans and, thus, strongly perturb the initial input such that the added noise becomes audible for a human listener. In this case, an attack—although still viable—significantly loses its malicious impact in practice.

4 Evaluation

With the help of the following experiments, we empirically verify and assess our proposed approach according to the following three main aspects:

(i) *Benign Performance.* The augmentation of the system should impair the performance on benign input as little as possible. We analyze different parameter combinations for the psychoacoustics and our band-pass filter to show that our augmented model retains its practical use.

(ii) *Adaptive Attacker.* To analyze the efficacy of the augmented system, we construct and assess its robustness against adversarial examples generated by a strong attacker with white-box access to the system. This attacker is aware of our augmentations and *actively* factors them into the optimization.

(iii) *Listening Test.* Finally, we verify the success of our method by a crowd-sourced user study. We conduct a listening test, investigating the quality (i.e., the inconspicuousness) of the adversarial examples computed from the adaptive attacker against the augmented ASR system.

All experiments were performed on a server running Ubuntu 18.04, with 128 GB RAM, an Intel Xeon Gold 6130 CPU, and four Nvidia GeForce RTX 2080 Ti. For our experiments, we use KALDI in version 5.3 and train the system with the default settings from the *Wall Street Journal* (WSJ) training recipe.

4.1 Metrics

To assess the quality of adversarial examples both in terms of efficacy and inconspicuousness, we use two standard measures.

Word Error Rate (WER) The *Word Error Rate* (WER) is computed based on the Levenshtein distance [50], which describes the *edit distance* between the reference transcription and the ASR output (i.e., the minimum number of edits required to transform the output text of the ASR system into the correct text).

We compute the Levenshtein distance $\mathcal{L}$ as the sum over all substituted words S, inserted words I, and deleted words D:

$$\text{WER} = 100 \cdot \frac{\mathcal{L}}{N} = 100 \cdot \frac{S+D+I}{N},$$

where N is the total number of words of the reference text. The smaller the WER, the fewer errors were made by the ASR system.

To evaluate the efficacy of adversarial examples, we measure the WER between the adversarial target transcription and

the output of the ASR system. Thus, a *successful adversarial example* has a WER of 0 %, i. e., fully matching the desired target description y'. Note that the WER can also reach values above 100 %, e. g., when many words are inserted. This can especially happen with unsuccessful adversarial examples, where mostly the original text is transcribed, which leads to many insertions.

Segmental Signal-to-Noise Ratio (SNRseg) The WER can only measure the success of an adversarial example in fooling an ASR system. For a real attack, we are also interested in the (in-)conspicuousness of adversarial examples, i. e., the level of the added perturbations. For this purpose, we quantify the changes that an attacker applies to the audio signal. Specifically, we use the *Signal-to-Noise Ratio* (SNR) to measure the added perturbations. More precisely, we compute the *Segmental Signal-to-Noise Ratio* (SNRseg) [51, 52], a more accurate measure of distortion than the SNR, when signals are aligned [52].

Given the original audio signal $x(t)$ and the adversarial perturbations $\sigma(t)$ defined over the sample index t, the SNRseg can be computed via

$$\text{SNRseg(dB)} = \frac{10}{K} \sum_{k=0}^{K-1} \log_{10} \frac{\sum_{t=Tk}^{Tk+T-1} x^2(t)}{\sum_{t=Tk}^{Tk+T-1} \sigma^2(t)},$$

with T being the number of samples in a segment and K the total number of segments. For our experiments, we set the segment length to 16 ms, which corresponds to $T = 256$ samples for a 16 kHz sampling rate.

The *higher* the SNRseg, the *less* noise has been added to the audio signal. Hence, an adversarial example is considered less conspicuous for higher SNRseg values. Note that we use the SNRseg ratio only as an approximation for the perceived noise. We perform a listening test with humans for a realistic assessment and show that the results of the listening test correlate with the reported SNRseg (cf. Section 4.4).

4.2 Benign Performance

Our goal is to preserve accuracy on benign inputs (i. e., non-malicious, unaltered speech) while simultaneously impeding an attacker as much as possible. To retain that accuracy as much as possible, the parameters of the band-pass, and the psychoacoustic filter need to be carefully adjusted. Note that adversarial robustness is generally correlated with a loss in accuracy for image classification models [53]. Accordingly, we assume that higher adversarial robustness likely incurs a trade-off on benign input performance.

All models in this section are trained with the default settings for the *Wall Street Journal* (WSJ) training recipe of the KALDI toolkit [35]. The corresponding WSJ-based speech corpus [54] contains approximately 81 hours of training data and consists of uttered sentences from the Wall Street Journal.
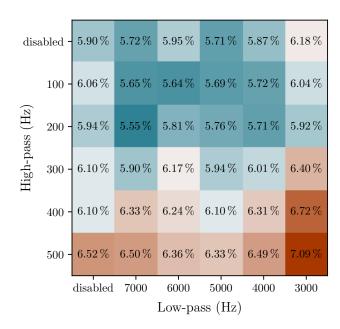

Figure 2: **Word Error Rate (WER) for different band-pass filters.** For each filter, we train three models and report the best accuracy in terms of WER (the lower, the better).

We train three models for each configuration and report the WER on the test set for the model with the best performance. For the test set, we use the `eval92` subset consisting of 333 utterances with a combined length of approximately 42 minutes.

Band-Pass Filtering The band-pass filter limits the signal's frequency range by removing frequencies below and above certain thresholds. Our goal is to remove parts of the audio that are not used by the human voice. We treat these values as classical hyperparameters and select the best performing combination by grid searching over different cut-off frequencies; for each combination, we train a model from scratch, using the training procedure outlined above. The results are depicted in Figure 2. If we narrow the filtered band (i. e., remove more information), the WER gradually increases and, therefore, worsens the recognizer's accuracy. However, for many cases, even when removing a significant fraction of the signal, the augmented system either achieves comparable results or even surpasses the baseline (WER 5.90 %). In the best case, we reach an improvement by 0.35 % percentage points to a WER of 5.55 % (200 Hz-7000 Hz). This serves as evidence that the unmodified input contains signals that are not needed for transcription. In Section 4.3.3, we further confirm this insight by analyzing models with narrower bands. We hypothesize that incorporating a band-pass filter might generally improve the performance of ASR systems but note that further research on this is needed.

For the remaining experiments, if not indicated otherwise, we use the 200-7000 Hz band-pass.

Table 1: **Recognition rate of the ASR system on benign input.** We report the performance of an unmodified KALDI system as well as two variants hardened by our approach. For our model, the scaling factor ϕ is set to 0 and the band-pass filter configured with 200-7000Hz. Note, when feeding standard input to DOMPTEUR, we disable its psychoacoustic filtering capabilities.

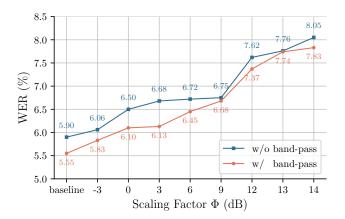|	KALDI	DOMPTEUR	
		w/o band-pass	w/ band-pass
Standard Input	WER 5.90 %	WER 6.20 %	WER 6.33 %
Processed Input	WER 8.74 %	WER 6.50 %	WER 6.10 %

Figure 3: **Recognition rate for psychoacoustic filtering.** For each ϕ we train a model both with and without band-pass filter (200-7000Hz) and report the best accuracy from three repetitions. A negative scaling factor partially retains inaudible ranges. Note that the benefits of the band-pass filter are retrained, even when we incorporate psychoacoustic filtering.

Psychoacoustic Filtering The band-pass filter allows us to remove high- and low-frequency parts of the signal; however, the attacker can still hide within this band in inaudible ranges. Therefore, we use psychoacoustic filtering as described in Section 3.1 to remove these parts in the signal. We evaluate different settings for Φ from Equation (2) – by increasing Φ, we artificially increase the hearing thresholds, resulting in more aggressive filtering. We plot the results in Figure 3 for both psychoacoustic filtering and a baseline WER, with and without band-pass, respectively. The WER increases with increasing Φ, i.e., the performance drops if more of the signal is removed, independent of the band-pass filter.

When we use no band-pass filter, the WER increases from 5.90 % (baseline) to 6.50 % for $\Phi = 0$ dB, which is equivalent to removing everything below the human hearing thresholds. When we use more aggressive filtering—which results in better adversarial robustness (cf. Section 4.3)—the WER increases up to 8.05 % for $\Phi = 14$ dB. Note that the benefits of
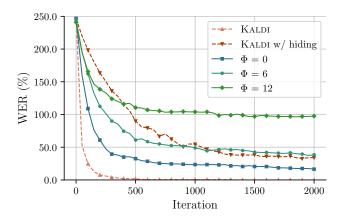

Figure 4: **Progress of attack for computing adversarial examples.** We run the attack against multiple instances of DOMPTEUR with different values of Φ and a 200Hz-7000Hz band-pass filter. The baseline refers to the attack from Schönherr et al. [17] against an unaltered instance of KALDI. For each attack report the Word Error Rate (WER) for the first 2000 iterations.

the band-pass filter remain even in the presence of psychoacoustic filtering and results in improving the WER to 6.10 % ($\Phi = 0$ dB) and 7.83 % ($\Phi = 14$ dB). We take this as another sign that a band-pass filter might generally be applicable to ASR systems.

Cross-Model Benign Accuracy Finally, we want to evaluate if DOMPTEUR indeed only uses relevant information. To test this hypothesis, we compare three different models. One completely unaugmented model (i.e., an unmodified version of KALDI), one model trained with psychoacoustics filtering, and one model trained with both psychoacoustics filtering and a band-pass filter. We feed these models two types of inputs: (i) *standard inputs*, i.e., inputs directly lifted from the WSJ training set, and (ii) *processed inputs*, these inputs are processed by our psychoacoustic filtering. If our intuitive understanding is correct and DOMPTEUR does indeed learn a better model of the human auditory system, it should retain a low WER even when presented with non-filtered input. Thus, the model has learned to *ignore* unnecessary parts of the input. The results are shown in Table 1 and match our hypothesis: DOMPTEUR's performance only drops slightly (6.10 % → 6.33 %) when presented with unfiltered input or does even improve if the band-pass is disabled (6.50 % → 6.20 %). KALDI, on the other hand, heavily relies on this information when transcribing audio, increasing its WER by 2.84 percentage point (5.90 % → 8.74 %). Thus, the results further substantiate our intuition that we filter only irrelevant information with our approach.

Table 2: **Number of successful Adversarial Examples (AEs) and Segmental Signal-to-Noise (SNRseg) ratio for the experiments with the adaptive attacker.** We report the numbers for all computed adversarial examples against the augmented models as well as our two baselines (with and without psychoacoustic hiding). As the success rate and SNRseg depend on the learning rate, we combine these in the last row. For this, we select the best (i.e., least noisy) AE for each utterance among the four learning rates. For the SNRseg, we only consider successful AEs. The higher the SNSseg, the *less* noise (i. e., adversarial perturbation) is present in the audio signal. Negative values indicate that the energy of the noise exceeds the energy in the original signal.

Learning Rate	Metric	KALDI		DOMPTEUR						
		baseline w/o hiding	baseline w/ hiding	$\Phi = 0$	$\Phi = 3$	$\Phi = 6$	$\Phi = 9$	$\Phi = 12$	$\Phi = 13$	$\Phi = 14$
0.05	AEs	50/50	17/50	31/50	28/50	10/50	4/50	0/50	0/50	0/50
	SNR	5.80/ 14.44	13.48/ 18.50	6.03/10.63	3.61/ 8.31	1.21/5.53	1.50/ 3.23	—	—	—
0.01	AEs	50/50	28/50	38/50	34/50	22/50	10/50	0/50	0/50	0/50
	SNR	2.15/ 10.59	9.36/ 15.81	3.74/ 9.53	0.47/ 6.41	-0.68/3.60	-1.31/ 1.10	—	—	—
0.5	AEs	49/50	23/50	48/50	44/50	42/50	20/50	1/50	1/50	0/50
	SNR	-8.54/ -0.02	1.08/ 8.63	-3.78/ 3.24	-6.51/ 0.11	-7.74/-1.47	-8.69/-3.35	-13.56/-13.56	-15.69/-15.69	—
1	AEs	50/50	16/50	49/50	50/50	43/50	23/50	1/50	1/50	0/50
	SNR	-13.68/ -5.03	-1.81/ 4.50	-7.44/-0.29	-10.50/-3.00	-10.99/-4.34	-11.98/-6.37	-17.69/-17.69	-11.73/-11.73	—
Best AEs	**AEs**	**50/50**	**37/50**	**50/50**	**50/50**	**46/50**	**27/50**	**2/50**	**2/50**	**0/50**
	SNR	**5.80/ 14.44**	**8.71/ 18.50**	**3.36/10.63**	**0.85/ 8.31**	**-4.71/5.53**	**-7.14/ 3.23**	**-15.62/-13.56**	**-13.71/-11.73**	**—**

AEs: Successful adversarial examples; *SNR*: SNRseg/SNRseg$_{max}$ in dB

4.3 Adaptive Attacker

We now want to evaluate how robust DOMPTEUR is against adversarial examples. We construct a strong attacker with complete knowledge about the system and, in particular, our modifications. Ultimately, this allows us to create successfully adversarial examples. However, as inaudible ranges are removed, the attacker is now forced into human-perceptible ranges, and, consequently, the attack loses much of its malicious impact. We provide further support for this claim in Section 4.4 by performing a user study to measure the perceived quality of adversarial examples computed with this attack.

Attack. We base our evaluation on the attack by Schönherr et al. [17], which presented a strong attack that works with KALDI. Recent results show that it is crucial to design adaptive attacks as simple as possible while simultaneously resolving any obstacles for the optimization [55]. To design such an attacker against DOMPTEUR, we need to adjust the attack to consider the augmentations in the optimization. Therefore, we extend the baseline attack against KALDI to include both the band-pass and psychoacoustic filter into the computation. This allows the attacker to compute gradients for the entire model in a white-box fashion.

More specifically, we extend the gradient descent step s.t. (i) the band-pass filter and (ii) the psychoacoustic filter component back-propagates the gradient respectively.

(i) *Band-Pass Filter.* For the band-pass filter we remove those frequencies that are smaller and larger than the cut-off frequencies of the band-pass filter. This is also applied to the gradients of the back propagated gradient to ignore changes that will fall into ranges of the removed signal

$$\nabla_{\mathbf{T}(n,k)} = 0 \quad \forall f_{\max} < k < f_{\min}. \tag{5}$$

(ii) *Psychoacoustic Filter.* The same principle is used for the psychoacoustic filtering, where we use the mask $\mathbf{M}$ to zero out components of the signal that the network will not process

$$\nabla_{\mathbf{S}} = \nabla_{\mathbf{T}} \odot \mathbf{M}. \tag{6}$$

Experimental Setup. We evaluate the attack against different versions of DOMPTEUR. Each model uses a $200 - 7000\,Hz$ band-pass filter, and we vary the degrees of the psychoacoustic filtering ($\Phi \in \{0, 3, 6, 9, 12, 13, 14\}$). We compare the results against two baselines to evaluate the inconspicuousness of the created adversarial examples. First, we run the attack of Schönherr et al. without psychoacoustic hiding against an unaltered version KALDI. Second, we re-enable psychoacoustic hiding and run the original attack against KALDI, to generate state-of-the-art inaudible adversarial examples. As a sanity check, we also ran the original attack (i. e., with psychoacoustic hiding) against DOMPTEUR. As expected, this attack did not create any adversarial examples since we filter the explicit ranges the attacker wants to utilize.

As a target for all configurations, we select 50 utterances with an approximate length of 5s from the WSJ speech corpus test set eval92. The exact subset can be found in appendix A. We use the same target sentence *send secret financial report* for all samples.

(a) Unmodified Signal

(b) Adversarial Example against KALDI

(c) Adversarial Example against DOMPTEUR ($\Phi = 12$)
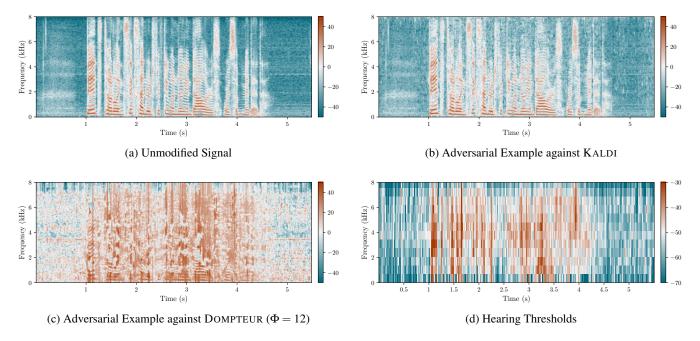

(d) Hearing Thresholds

Figure 5: **Spectrograms of adversarial examples.** Figure 5a shows the unmodified signal, Figure 5b depicts the baseline with an adversarial example computed against KALDI with psychoacoustic hiding, Figure 5c an adversarial example computed with the adaptive attack against DOMPTEUR, and Figure 5d shows the computed hearing thresholds for the adversarial example.

These parameters are chosen such that an attacker needs to introduce ~4.8 phones per second into the target audio, which Schönherr et al. suggests as both effective and efficiently possible [17]. Furthermore, we picked the utterances and target sentence to be *easy* for an attacker in order to decouple the influence on our analysis. Specifically, for these targets the baseline has a very high success rate and low SNRseg (cf. Table 2). Note that the attack is capable of introducing arbitrary target sentences (up to a certain length). In Section 4.3.2, we further analyze the influence of the phone rate, and in particular, the influence of the target utterance and sentence on the SNRseg. We compute adversarial examples for different learning rates and a maximum of 2000 iterations. This number is sufficient for the attack to converge, as shown in Figure 4, where the WER is plotted as a function of the number of iterations.

Results. The main results are summarized in Table 2. We report the average SNRseg over all adversarial examples, the best (SNRseg$_{max}$), and the number of successful adversarial examples created.

We evaluate the attack using different learning rates (0.05, 0.10, 0.5, and 1). In our experiments, we observed that while small learning rates generally produce less noisy adversarial examples, they simultaneously get more stuck in local optima. Thus, to simulate an attacker that would run an extensive search and uses the best result we also report the intersection of successful adversarial examples over all learning rates. If

success rate is the primary goal, we recommend a higher learning rate.

By increasing Φ, we can successfully force the attacker into audible ranges while also decreasing the attack's success rate. When using very aggressive filtering ($\Phi = 14$), we can prevent the creation of adversarial examples completely, albeit with a hit on the benign WER ($5.55\% \rightarrow 7.83\%$). Note, however, that we only examined 50 samples of the test corpus, and other samples might still produce valid adversarial examples. We see that adversarial examples for the augmented systems are more distorted for all configurations compared to the baselines. When using $\Phi \geq 12$, we force a negative SNRseg for all learning rates. For these adversarial examples, the noise (i. e., adversarial perturbations) energy exceeds the energy of the signal. With respect to the baselines, the noise energy increases on average by 21.42 dB (without psychoacoustic hiding) and 24.33 dB (with hiding enabled). This means there is, on average, ten times more energy in the adversarial perturbations than in the original audio signal. A graphical illustration can be found in Figure 5, where we plot the power spectra of different adversarial examples compared to the original signal.

4.3.1 Non-speech Audio Content

The task of an ASR system is to transcribe audio files with spoken content. An attacker, however, might pick other content, i.e., music or ambient noise, to obfuscate his hidden commands. Thus, we additionally evaluated adversarial ex-

Table 3: **Number of successful Adversarial Examples (AEs) and mean Segmental Signal-to-Noise (SNRseg) ratio for non-speech audio content.** For each AE, we selected the least noisiest example, from running the attack with learning rates ($\{0.05, 0.1, 0.5, 1.\}$). For the SNRseg we only consider successful AEs and report the difference to the baseline (KALDI). We highlight the highest loss in bold.

	Birds			Music		
	AEs	SNRseg (dB)	Loss	AEs	SNRseg (dB)	Loss
KALDI						
w/o hiding	50/50	11.83		45/50	23.26	
w/ hiding	5/50	17.76	(+5.93)	3/50	28.06	(+4.80)
DOMPTEUR						
$\Phi = 0$	50/50	9.58	(-2.25)	50/50	26.35	(+3.09)
$\Phi = 6$	31/50	-2.15	(-13.98)	45/50	16.03	(-7.23)
$\Phi = 12$	5/50	-12.25	(**-24.08**)	3/50	1.94	(**-21.32**)

Table 4: **Attack for different cut-off frequencies of the band-pass filter.** We report the number of successful adversarial examples (AEs) and the mean Segmental Signal-to-Noise (SNRseg) ratio. For the SNRseg we only consider successful AEs.

Band-pass	300Hz-7000Hz	300Hz-5000Hz	300Hz-3000Hz	500Hz-7000Hz	500Hz-5000Hz	500Hz-3000Hz
AEs	18/20	18/20	11/20	20/20	17/20	12/20
SNRseg	7.82	7.55	7.27	8.45	7.90	7.39
WER	5.90 %	5.94 %	6.40 %	6.50 %	6.33 %	7.09 %

amples based on audio files containing music and bird sounds. The results are presented in Table 3.

We can repeat our observations from the previous experiment. When we utilize a more aggressive filter, we observe that the perturbation energy of adversarial examples increases with respect to the baselines by up to 24.08 dB (birds) and 21.32 dB (music). Equally, the attack's general success decreases to 5/50 (birds) and 3/50 (music) successful adversarial examples.

Note that the SNRseg for music samples are in general higher than that of speech and bird files as these samples have a more dynamic range of signal energy. Hence, potentially added adversarial perturbations have a smaller impact on the calculation of the SNRseg. The absolute amount of added perturbations, however, is similar to that of other content. Thus, when listening to the created adversarial examples[2] the samples are similarly distorted. This is further confirmed in Section 4.4 with our listening test.

4.3.2 Target Phone Rate

The success of the attack depends on the ratio between the length of the audio file and the length of the target text, which we refer to as the *target phone rate*. This rate describes how
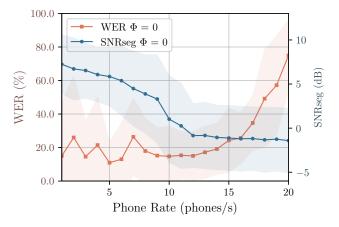
---

[2] rub-syssec.github.io/dompteur

Figure 6: **Word Error Rate (WER) and Segmental Signal-to-Noise (SNRseg) ratio for different phone rates.** We report the mean and std. deviation for adversarial examples computed for targets with varying length.

many phones an attacker can hide within one second of audio content.

In our experiments, we used the default ratios recommended by Schönherr et al. However, a better rate might exist for our setting. Therefore, to evaluate the effect of the target phone rate, we sample target texts of varying lengths from the WSJ corpus and compute adversarial examples for different target phone rates. We pick phone rates ranging from 1 to 20 and run 20 attacks for each of them for at most 1000 iterations, resulting in 400 attacks.

The results in Figure 6 show that, in general, with increasing phone rates, the SNRseg decreases and stagnates for target phone rate beyond 12. This is expected as the attacker tries to hide more phones and, consequently, needs to change the signal more drastically. Thus, we conclude that the default settings are adequate for our setting.

4.3.3 Band-Pass Cut-off Frequencies

So far, we only considered a relatively wide band-pass filter (200-7000 Hz). We also want to investigate other cut-off frequencies. Thus, we disable the psychoacoustic filtering and compute adversarial examples for different models examined in Section 4.2. We run the attack for each band-pass model with 20 speech samples for at most 1000 iterations.

The results are reported in Table 4. We observe that the energy amount of adversarial perturbation remains relatively constant for different filters, which is expected since the attacker has complete knowledge of the system. As we narrow the frequency band, the attacker adopts and puts more perturbation within these bands.

Apart from the SNRseg, we also observe a decrease in the attack success, especially for small high cut-off frequencies, with only 11/20 (300-3000 Hz) and 12/20 (500-3000 Hz) successful adversarial examples.

Table 5: **Regression results for perceived sound quality predicted by different audio stimuli.** The dependent variable is the quality score assigned to each audio stimulus. We trained three different models, one for each data set (speech/music/bird). Each model consists of two steps, with the first step entering the audio stimulus as a predictor and the second step entering type of device as a covariate. All models include the control variables gender, age, and language. All regressions use ordinary least squares. Cluster adjusted standard errors are indicated in parentheses. The R^2 values indicate the percentage of the variance of the perceived sound quality explained by the respective audio stimuli.

	Speech		Music		Bird	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
Audio stimulus	-.905**	-.905**	-.871**	-.871**	-.830**	-.830**
	(.131)	(.131)	(.166)	(.166)	(.171)	(.171)
Device		.030**		.008		.045**
		(.473)		(.597)		(.615)
Controls	Included		Included		Included	
Obs.	4259		4259		4259	
R^2	.820	.821	.760	.761	.690	.692

P-value < 0.05 = *, *P*-value < 0.01 = **

4.4 Listening Tests

Our goal is to make an adversarial attack noticeable by forcing modification to an audio signal into perceptible ranges. We have used the SNRseg as a proxy of the perceived audio quality of generated adversarial examples. However, this value can only give a rough approximation, and we are in general more interested in the judgment of human listeners. Specifically, we are interested to quantify if and to what extent malicious perturbations are audible to human listeners.

Therefore, we have conducted a *Multiple Stimuli with Hidden Reference and Anchor* (MUSHRA) test [56], a commonly used test to assess the quality of audio stimuli. This test allows us to get a ranking of the perceived quality of adversarial examples in comparison to an unmodified reference signal. Based on this measure, we can derive whether a participant 1) could detect any difference between an adversarial example and a clean signal (i.e., whether perturbations are audible) and, 2) obtain a subjective estimate on the amount of perceived perturbations (i.e., poorly rated samples are perceived more noisy).

Study Design In a MUSHRA test, the participants are presented with a set of differently processed audio files, the *audio stimuli*. They are asked to rate the quality of these stimuli on a scale from 0 (bad) to 100 (excellent). To judge whether the participants are able to distinguish between different audio conditions, a MUSHRA test includes two additional stimuli: (i) an unaltered version of the original signal (the so-called *reference*) and (ii) a worst-case version of the signal, which

is created by artificial degrading the original stimulus (the so-called *anchor*). In an ideal setting, the reference should be rated best, the anchor worst.

We want to rank the perceived quality of adversarial examples computed against DOMPTEUR and KALDI. For DOMPTEUR, we select three different versions: each model uses a $200 - 7000\,Hz$ band-pass filter, and we vary the degree of the psychoacoustic filtering ($\Phi \in \{0, 6, 12\}$). For KALDI, we calculate adversarial examples against the unaltered system with psychoacoustic hiding enabled (cf. Section 2) to compare against state-of-the-art adversarial examples.

As the reference, we use the original utterance, on which the adversarial examples are based. To be a valid comparison, we require the anchor to sound *similar*, yet noisier than the adversarial examples. Otherwise, it could be trivially identified and would not serve as a valid comparison.

Thus, we construct the anchor as follows: For a given set, we scale and sum the noise of each of the three adversarial examples and add this sum to the original stimulus, such that 1) each noise signal contributes the same amount of energy and 2) the SNRseg of the anchor is at least 6dB lower than the SNRseg of any of the adversarial examples in the set.

We have prepared a MUSHRA test with six test sets based on three different audio types: two speech sample sets, two music sample sets, and two sample sets with bird sounds.

These sets were selected among the sets of successful adversarial examples against all four models. For each set, we picked the samples whose adversarial examples produced the highest SNR (i. e., the "cleanest") for the strongest version of DOMPTEUR ($\Phi = 12$). The target text remained the same for all adversarial examples, and in all cases, the attacks were successful within 2000 iterations.

Results To test our assumptions in the field, we have conducted a large-scale experimental study. The G*Power 3 analysis [57] identified that a sample size of 324 was needed to detect a high effect size of $\eta^2 = .50$ with sufficient power $(1 - \beta > .80)$ for the main effect of univariate analyses of variance (UNIANOVA) among six experimental conditions and a significance level of $\alpha = .05$.

We used Amazon MTurk to recruit 355 participants ($\mu_{age} = 41.61$ years, $\sigma_{age} = 10.96$; 56.60% female). Participants were only allowed to use a computer and no mobile device. However, they were free to use headphones or speakers as long as they indicated what type of listening device was used. To filter individuals who did not meet the technical requirements needed, or did not understand or follow the instructions, we used a control question to exclude all participants who failed to distinguish the anchor from the reference correctly.

In the main part of the experiment, participants were presented with six different audio sets (2 of each: speech/bird/music), each of which contained six audio stimuli varying in sound quality. After listening to each sound, they were asked to rank the individual stimulus by its perceived

sound quality. After completing of the tasks, participants answered demographic questions, were debriefed (MTurk default), and compensated with 3.00 USD. The participant required on average approximately 20 minutes to finish the test.

In a first step, we first use an UNIANOVA to examine whether there is a significant difference between the six audio stimuli and the perceived sound quality. Our analysis reveals a significant main effect of the audio stimulus on the perceived sound quality, $F(5, 12780) = 8335.610$, $p < .001$, $\eta^2 = .765$. With an alpha level of $> 1\%$ for our p-value and an effect size of $\eta^2 > .5$, our result shows a high experimental significance [58]. Thus, we can conclude that DOMPTEUR indeed forces adversarial perturbations into the perceptible acoustic range of human listeners.

To examine whether the effect remains stable across different audio samples and listening devices, we further conducted multiple regression analyses. We entered the audio stimuli as our main predictors (first step) and the type of device (second step) as covariates for each analysis. Our results remain stable across all audio types. The highest predictive power was found in the *speech* sets, where 82.1% of the variance is explained by our regression model, followed by *music* (76.1%) and *bird* sets (69.2%) (see Table 5 for details). Moreover, we found a small yet significant positive coefficient for the type of device used across all audio types. This finding suggests that headphone users generally indicate higher quality rankings, potentially due to better sound perceptions. The results with listening device *speaker* are presented in Figure 7. Importantly, all results remain stable across the control variables of age, gender, and first language.

In conclusion, the results strongly support our hypothesis that DOMPTEUR forces the attacker into the audible range, making the attack clearly noticeable for human listeners.

5 Related Work

In this section, we summarize research related to our work, surveying recent attacks and countermeasures.

Audio Adversarial Examples Carlini and Wagner [59] introduced targeted audio adversarial examples for ASR systems. For the attack, they assume a white-box attacker and use an optimization-based method to construct general adversarial examples for arbitrary target phrases against the ASR system DEEPSPEECH [32].

Similarly, Schönherr et al. [17] and Yuan et al. [16] have proposed an attack against the KALDI [35] toolkit. Both assume a white-box attacker and also use optimization-based methods to find adversarial examples. Furthermore, the attack from Schönherr et al. [17] can optionally compute adversarial examples that are especially unobtrusive for human listeners.

Alzantot et al. [60] proposed a black-box attack, which does not require knowledge about the model. For this, the authors have used a genetic algorithm to create their adversarial
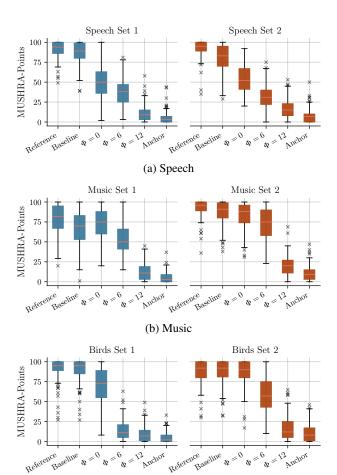

(a) Speech

(b) Music

(c) Birds

Figure 7: **Ratings of participants with listening device *speaker*.** In the user study, we tested six audio samples, divided into two samples each of spoken content, music and bird twittering.

examples for a keyword spotting system. Khare et al. [61] proposed a black-box attack based on evolutionary optimization, and also Taori et al. [62] presented a similar approach in their paper.

Recently, Chen et al. [63] and Schönherr et al. [18] published works where they can calculate over-the-air attacks, where adversarial examples are optimized such that these remain viable if played via a loudspeaker by considering room characteristics.

Aghakhani et al. [64] presented another line of attack, namely a poisoning attack against ASR systems. In contrast to adversarial examples, these are attacks against the training set of a machine learning system, with the target to manipulate the training data s.t a model that is trained with the poisoned data set misclassifies specific inputs.

Abdullah et al. [19] provides a detailed overview of existing attacks in their systemization of knowledge on attacks against speech systems.

Countermeasures There is a long line of research about countermeasures against adversarial examples in general and especially in the image domain (e. g., [23–25]), but most of the proposed defenses were shown to be broken once an attacker is aware of the employed mechanism. In fact, due to the difficulty to create robust adversarial example defenses, Carlini et al. proposed guidelines for the evaluation of adversarial robustness. They list all important properties of a successful countermeasure against adversarial examples [30]. Compared to the image domain, defenses against audio adversarial examples remained relatively unnoticed so far. For the audio domain, only a few works have investigated possible countermeasures. Moreover, these tend to focus on specific attacks and not adaptive attackers.

Ma et al. [65] describe how the correlation of audio and video streams can be used to detect adversarial examples for an audiovisual speech recognition task. However, all of these simple approaches—while reasonable in principle—are specifically trained for a defined set of attacks, and hence an attacker can easily leverage that knowledge as demonstrated repeatedly in the image domain [25].

Zeng et al. [66] proposed an approach inspired by multi-version programming. Therefore, the authors combine the output of multiple ASR systems and calculate a *similarity score* between the transcriptions. If these differ too much, the input is assumed to be an adversarial example. The security of this approach relies on the property that current audio adversarial examples do not transfer between systems — an assumption that has been already shown to be wrong in the image domain [45].

Yang et al. [67], also utilize specific properties of the audio domain and uses the temporal dependency of the input signal. For this, they compare the transcription of the whole utterance with a segment-wise transcription of the utterance. In the case of a benign example, both transcriptions should be the same, which will not be the case for an adversarial example. This proved effective against static attacks, and the authors also construct and discussed various adaptive attacks but these were later shown to be insufficient [55].

Besides approaches that aim to harden models against adversarial examples, there is a line of research that focuses on detecting adversarial examples: Liu and Ditzler [68] utilizing quantization error of the activations of the neural network, which appear to be different for adversarial and benign audio examples. Däubener et al. [69] trained neural networks capable of uncertainty quantification to train a classifier on different uncertainty measures to detect adversarial examples as outliers. Even if they trained their classifier on benign examples only, it will most likely not work for any kind of attack, especially those aware of the detection mechanism.

In contrast, our approach does not rely on detection by augmenting the entire system to become more resilient against adversarial examples. The basic principle of this has been discussed as a defense mechanism in the image domain with

JPEG compression [70, 71] as well as in the audio domain by Carlini and Wagner [59], Rajaratnam et al. [72], Andronic et al. [73], and Olivier et al. [74]. These approaches, however, were only used as a pre-processing step to remove semantically irrelevant parts from the input and thereby destroy adversarial perturbations added by (static) attackers. In contrast, we aim to train an ASR system that uses the same information set as the human auditory systems. Consequently, adversarial examples computed against this system are also restricted to this set, and an attack cannot be hidden in inaudible ranges. Similar to the referenced approaches, we rely on psychoacoustics and baseband filtering. However, we do not solely employ this as a pre-processing step but train a new system with our augmentation data (i.e., removing imperceptible information from the training set). This allows us to not simply destroy adversarial perturbations but rather confine the available attack surface.

6 Discussion

We have shown how we can augment an ASR system by utilizing psychoacoustics in conjunction with a band-pass filter to effectively remove semantically irrelevant information from audio signals. This allows us to train a hardened system that is more aligned with human perception.

Model Hardening Our results from Section 4.2 suggest that the hardened models primarily utilize information available within audible ranges. Specifically, we observe that models trained on the unmodified data set appear to use *any* available signals and utilize information *both* from audible and non-audible ranges. This is reflected in the accuracy drop when presented with psychoacoustically filtered input (where only audible ranges are available). In contrast, the augmented model performs comparably well on both types of input. Hence, the model focuses on the perceivable audible ranges and *ignores* the rest.

Robustness of the System We demonstrated how we can create a more realistic attacker, which actively factors in the augmentations during the calculation of adversarial examples. In this case, however, the attack is forced into the audible range. This makes the attack significant more perceptible — resulting in an average SNRseg drop of up to 24.33 dB for speech samples. These results also transfer to other types of audio content (i.e., music and birds tweeting) and are further confirmed by the listening test conducted in Section 4.4. In summary, the results of these experiments show that an attack is clearly perceivable. Further, we find that the adversarial examples, calculated with the adaptive attack, are easily distinguishable from benign audio files by humans.

Implementation Choices In general, our augmentations can be implemented in the form of low-cost pre-processing steps with no noteworthy performance overhead. Only the model needs to be retrained from scratch. However, the cost of this could—in theory—be partially alleviated by transfer learning. We leave this question as an interesting direction for future research.

Robustness-Performance Tradeoff The results of the adaptive attack (cf. Table 2) show that a larger margin Φ leads to stronger robustness. Specifically, for $\Phi = 14$, the attacker was unable to find *any* successful adversarial example in our experiments. However, this incurs an expected robustness-performance trade-off as previous research indicates that adversarial robustness is generally correlated with a loss in accuracy [53].

In the case of our strong white-box attacker, we recommend a margin $\Phi \geq 12$, which result in a degraded system performance by at least 1.82 percentage points in terms of the benign WER. In this case, though, we already granted the attacker many concessions: full access to the model with all parameters, ideal playback (i.e., adversarial examples are fed directly into the recognizer and are not played over-the-air), and an easy target. We chose to study our attacker in this setting as this poses the strongest class of attacks and allows us to gain meaningful insights.

In contrast to white-box attacks, black-box attack don't have direct access to the gradient and for example rely on surrogate models [75] or generative algorithms [76] to construct adversarial examples. Therefore, adversarial examples from these attacks are typically more conspicuous and can even introduce semantic changes such that humans can perceive the hidden transcription if they are made aware of it [75]. Considering our augmentations, we expect that current black-box attacks are able to construct valid adversarial examples against DOMPTEUR. However, we expect these to be significantly more noisy (in comparison to the adaptive attacker) as DOMPTEUR forces modifications to the signal into audible ranges regardless of the underlying attack strategy. Especially in a realistic over-the-air setting, we suspect much higher distortions since the attacker is much more constrained. In such a setting, a smaller Φ might also already suffice. We leave this as an interesting research direction for future work.

Improvement of the Attack The adaptive attack presented in Section 4.3 can successfully compute adversarial examples, except for very aggressive filtering. While Figure 4 clearly shows that the attack has converged, we were still unable to find working adversarial examples. However, other target/input utterance combinations may still exist, for which the attack works and novel attack strategies should be studied.

Forcing Semantics into Adversarial Examples We have shown how we can force adversarial audio attacks into the audible range. This makes them clearly perceivable. Ultimately, the goal is to push adversarial examples towards the perceptual boundary between original and adversarial message. Intuitively, adversarial examples should require such extensive modification that a *human listener* will perceive the target transcription, i. e., that the adversarial perturbation carries *semantic* meaning. We view our work as a first successful step into that direction and leave the exploration of this strategy as an interesting question for future work.

7 Conclusion

In this work, we proposed a broadly applicable design principle for ASR systems that enables them to resemble the human auditory system more closely. To demonstrate the principle, we implemented a prototype of our approach in a tool called DOMPTEUR. More specifically, we augment KALDI using psychoacoustic filtering in conjunction with a band-pass filter. In several experiments, we demonstrate that our method renders our system more robust against adversarial examples, while retaining a high accuracy on benign audio input.

We have argued that an attacker can find adversarial examples for any kind of countermeasure, particularly if we assume the attack to have full white-box access to the system. Specifically, we have calculated adversarial examples for DOMPTEUR via an adaptive attack, which leverages the full knowledge of the proposed countermeasures. Although this attack is successful in computing adversarial examples, we show that the attack becomes much less effective. More importantly, we find that adversarial examples are of poor quality, as demonstrated by the SNRseg and our listening test.

In summary, we have taken the first steps towards bridging the gap between human expectations and the reality of ASR systems—hence taming adversarial attacks to a certain extent by robbing them of their stealth abilities.

References

[1] Michael J Pazzani and Daniel Billsus. Content-Based Recommendation Systems. In *The Adaptive Web*. Springer, 2007.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level Control through Deep Reinforcement Learning. *nature*, 2015.

[4] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *nature*, 2016.

[5] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with Large Scale Deep Reinforcement Learning. *Computing Research Repository (CoRR)*, abs/1912.06680, 2019.

[6] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster Level in StarCraft II using Multi-Agent Reinforcement Learning. *nature*, 2019.

[7] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved Protein Structure Prediction using Potentials from Deep Learning. *nature*, 2020.

[8] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving Human Parity in Conversational Speech Recognition. *Computing Research Repository (CoRR)*, abs/1610.05256, 2016.

[10] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. Conversational AI: The Science Behind the Alexa Prize. In *Alexa Prize*, 2017.

[11] Lauren Goode. Amazon's Alexa will now lock your door for you (if you have a 'smart' lock). https://www.theverge.com/circuitbreaker/2016/7/28/123 05678/amazon-alexa-works-with-august-smart-lock-door-WiFi-bridge. Accessed: 2021-06-02.

[12] Stephen Shankland. Meet Tesla's self-driving car computer and its two AI brains. https://www.cnet.com/news/meet-tesla-self-driving-car-computer-and-its-two-ai-brains/. Accessed: 2021-06-02.

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[14] Liwei Song and Prateek Mittal. POSTER: Inaudible Voice Commands. In *ACM Conference on Computer and Communications Security (CCS)*, 2017.

[15] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. DolphinAttack: Inaudible Voice Commands. In *ACM Conference on Computer and Communications Security (CCS)*, 2017.

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

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

[18] Lea Schönherr, Thorsten Eisenhofer, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Imperio: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems. In *Annual Computer Security Applications Conference (ACSAC)*, 2020.

[19] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. In *IEEE Symposium on Security and Privacy (S&P)*, 2020.

[20] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2015.

[21] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

[23] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On Detecting Adversarial Perturbations. In *International Conference on Learning Representations (ICLR)*, 2017.

[24] Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting Adversarial Samples from Artifacts. *Computing Research Repository (CoRR)*, abs/1703.00410, 2017.

[25] Nicholas Carlini and David Wagner. Adversarial Examples are Not Easily Detected: Bypassing Ten Detection Methods. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2017.

[26] Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. Motivating the Rules of the Game for Adversarial Example Research. *Computing Research Repository (CoRR)*, abs/1807.06732, 2018.

[27] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A Simple Explanation for the Existence of Adversarial Examples with Small Hamming. *Computing Research Repository (CoRR)*, abs/1901.10861, 2019.

[28] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[29] Brian B. Monson, Eric J. Hunter, Andrew J. Lotto, and Brad H. Story. The Perceptual Significance of High-frequency Energy in the Human Voice. *Frontiers in Psychology*, 2014.

[30] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On Evaluating Adversarial Robustness. *Computing Research Repository (CoRR)*, abs/1902.06705, 2019.

[31] Herve A. Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Press, 1994.

[32] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep Speech: Scaling Up End-to-End Speech Recognition. *Computing Research Repository (CoRR)*, abs/1412.5567, 2014.

[33] Alex Graves and Navdeep Jaitly. Towards End-to-End Speech Recognition with Recurrent Neural Networks. In *International Conference on Machine Learning (ICML)*, 2014.

[34] Jian Kang, Wei-Qiang Zhang, Wei-Wei Liu, Jia Liu, and Michael T. Johnson. Advanced Recurrent Network-Based Hybrid Acoustic Models for Low Resource Speech Recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018.

[35] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi Speech Recognition Toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[36] Jun Du, Yan-Hui Tu, Lei Sun, Feng Ma, Hai-Kun Wang, Jia Pan, Cong Liu, Jing-Dong Chen, and Chin-Hui Lee. The USTC-iFlytek System for CHiME-4 Challenge. In *ISCA Workshop on Speech Processing in Everyday Environments (CHiME)*, 2016.

[37] Naoyuki Kanda, Rintaro Ikeshita, Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, Xiaofei Wang, Vimal Manohar, Nelson Enrique Yalta Soplin, Matthew Maciejewski, Szu-Jui Chen, et al. The Hitachi/JHU CHiME-5 System: Advances in Speech Recognition for Everyday Home Environments Using Multiple Microphone Arrays. In *ISCA Workshop on Speech Processing in Everyday Environments (CHiME)*, 2018.

[38] Ivan Medennikov, Ivan Sorokin, Aleksei Romanenko, Dmitry Popov, Yuri Khokhlov, Tatiana Prisyach, Nikolay Malkovskii, Vladimir Bataev, Sergei Astapov, Maxim Korenevsky, and Alexander Zatvornitskiy. The STC System for the CHiME 2018 Challenge. In *ISCA Workshop on Speech Processing in Everyday Environments (CHiME)*, 2018.

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

[40] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R. B. Butler, and Joseph Wilson. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. In *Symposium on Network and Distributed System Security (NDSS)*, 2019.

[41] Joseph Szurley and J. Zico Kolter. Perceptual Based Adversarial Audio Attacks. *Computing Research Repository (CoRR)*, abs/1906.06355, 2019.

[42] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2013.

[43] ISO Central Secretary. Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to 1.5 Mbits/s – Part3: Audio. Standard 11172-3, International Organization for Standardization, 1993.

[44] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *International Conference on Machine Learning (ICML)*, 2018.

[45] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples. *Computing Research Repository (CoRR)*, abs/1605.07277, 2016.

[46] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium*, 2016.

[47] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks Against Machine Learning. In *ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, 2017.

[48] Binghui Wang and Neil Zhenqiang Gong. Stealing Hyperparameters in Machine Learning. In *IEEE Symposium on Security and Privacy (S&P)*, 2018.

[49] Auguste Kerckhoffs. La Cryptographic Militaire. *Journal des Sciences Militaires*, 1883.

[50] Gonzalo Navarro. A Guided Tour to Approximate String Matching. *ACM Computing Surveys (CSUR)*, 2001.

[51] Stephen Voranl and Connie Sholl. Perception-Based Objective Estimators of Speech. In *IEEE Workshop on Speech Coding for Telecommunications: Speech Coding for Interoperable Global Colmmunications (SCFT)*, 1995.

[52] Wonho Yang. *Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based on Audible Distortion and Cognition Model*. PhD thesis, Temple University, 1999.

[53] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May be at Odds with Accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

[54] Douglas B. Paul and Janet M. Baker. The Design for the Wall Street Journal-Based CSR Corpus. In *Workshop on Speech and Natural Language*, 1992.

[55] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[56] Nadja Schinkel-Bielefeld, Netaya Lotze, and Frederik Nagel. Audio Quality Evaluation by Experienced and Inexperienced Listeners. In *International Congress on Acoustics (ICA)*, 2013.

[57] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical Power Analyses using G* Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods*, 2009.

[58] John T.E. Richardson. Eta Squared and Partial Eta Squared as Measures of Effect Size in Educational Research. *Educational Research Review*, 2011.

[59] Nicholas Carlini and David Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *IEEE Deep Learning and Security Workshop (DLS)*, 2018.

[60] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? Adversarial Examples Against Automatic Speech Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[61] Senthil Mani Shreya Khare, Rahul Aralikatte. Adversarial Black-Box Attacks on Automatic Speech Recognition Systems using Multi-Objective Evolutionary Optimization. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[62] Rohan Taori, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. Targeted Adversarial Examples for Black Box Audio Systems. In *IEEE Deep Learning and Security Workshop (DLS)*, 2019.

[63] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting Inaudible Commands Into Over-the-Air Voice Controlled Systems. In *Symposium on Network and Distributed System Security (NDSS)*, 2020.

[64] Hojjat Aghakhani, Thorsten Eisenhofer, Lea Schönherr, Dorothea Kolossa, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. VENOMAVE: Clean-Label Poisoning Against Speech Recognition. *Computing Research Repository (CoRR)*, abs/2010.10682, 2020.

[65] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Detecting Adversarial Attacks On Audio-Visual Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[66] Qiang Zeng, Jianhai Su, Chenglong Fu, Golam Kayas, and Lannan Luo. A Multiversion Programming Inspired Approach to Detecting Audio Adversarial Examples. In *Conference on Dependable Systems and Networks (DSN)*, 2019.

[67] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing Audio Adversarial Examples Using Temporal Dependency. In *International Conference on Learning Representations (ICLR)*, 2019.

[68] Heng Liu and Gregory Ditzler. Detecting Adversarial Audio via Activation Quantization Error. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[69] Sina Däubener, Lea Schönherr, Asja Fischer, and Dorothea Kolossa. Detecting Adversarial Examples for Speech Recognition via Uncertainty Quantification. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[70] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A Study of the Effect of JPG Compression on Adversarial Images. *Computing Research Repository (CoRR)*, abs/1608.00853, 2016.

[71] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Shield: Fast, Practical Defense and vaccination for Deep Learning using JPEG Compression. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

[72] Krishan Rajaratnam, Kunal Shah, and Jugal Kalita. Isolated and Ensemble Audio Preprocessing Methods for Detecting Adversarial Examples against Automatic Speech Recognition. In *Conference on Computational Linguistics and Speech Processing (ROCLING)*, 2018.

[73] Iustina Andronic, Ludwig Kürzinger, Edgar Ricardo Chavez Rosas, Gerhard Rigoll, and Bernhard U Seeber. MP3 Compression to Diminish Adversarial Noise in End-to-End Speech Recognition. In *International Conference on Speech and Computer*, 2020.

[74] Raphael Olivier, Bhiksha Raj, and Muhammad Shah. High-Frequency Adversarial Defense for Speech and Audio. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[75] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices. In *USENIX Security Symposium*, 2020.

[76] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems. In *ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, 2020.

A Targets

Table 6: **Target utterances for the experiments with the adaptive attacker.** For the experiments we select 50 utterances as target with an approximate length of 5s from the WSJ speech corpus test set `eval92`.

Utterance	Length	Utterance	Length	Utterance	Length
440c0407	5.47s	440c040i	5.07s	440c040j	4.08s
441c0409	4.91s	441c040c	5.57s	441c040l	5.26s
441c040m	5.50s	441c040s	4.58s	441c040y	4.08s
442c0402	5.14s	442c040c	5.69s	442c040d	4.80s
442c040h	4.63s	442c040k	5.37s	442c040w	5.21s
443c0402	5.05s	443c040b	4.69s	443c040c	4.73s
443c040d	5.54s	443c040j	4.61s	443c040l	4.23s
443c040p	4.10s	443c040v	4.82s	443c0417	4.55s
444c0407	4.76s	444c0409	5.20s	444c040i	4.98s
444c040n	5.18s	444c040u	4.37s	444c040w	5.52s
444c040z	4.29s	444c0410	4.16s	445c0409	5.43s
445c040j	4.99s	445c040l	5.64s	445c040w	4.92s
445c040x	4.68s	445c0411	4.34s	446c0402	5.59s
446c040b	4.10s	446c040d	5.18s	446c040e	4.94s
446c040f	4.66s	446c040o	5.33s	446c040p	5.04s
446c040s	5.18s	446c040v	4.07s	447c040g	4.64s
447c040p	5.23s	447c040z	4.68s		