Beyond L_p clipping: Equalization-based Psychoacoustic Attacks against ASRs

Hadi Abdullah hadi10102@ufl.edu
Muhammad Sajidur Rahman rahmanm@ufl.edu
Christian Peeters cpeeters@ufl.edu
Cassidy Gibson c.gibson@ufl.edu
Washington Garcia w.garcia@ufl.edu
Vincent Bindschaedler vbindsch@cise.ufl.edu
Thomas Shrimpton teshrim@ufl.edu
Patrick Traynor traynor@ufl.edu
University of Florida

Abstract

Automatic Speech Recognition (ASR) systems convert speech into text and can be placed into two broad categories: traditional and fully end-to-end. Both types have been shown to be vulnerable to adversarial audio examples that sound benign to the human ear but force the ASR to produce malicious transcriptions. Of these attacks, only the "psychoacoustic" attacks can create examples with relatively imperceptible perturbations, as they leverage the knowledge of the human auditory system. Unfortunately, existing psychoacoustic attacks can only be applied against traditional models, and are obsolete against the newer, fully end-to-end ASRs. In this paper, we propose an equalization-based psychoacoustic attack that can exploit both traditional and fully end-to-end ASRs. We successfully demonstrate our attack against real-world ASRs that include DeepSpeech and Wav2Letter. Moreover, we employ a user study to verify that our method creates low audible distortion. Specifically, 80 of the 100 participants voted in favor of *all* our attack audio samples as less noisier than the existing state-of-the-art attack. Through this, we demonstrate both types of existing ASR pipelines can be exploited with minimum degradation to attack audio quality.

1. Introduction

ASRs have become increasingly popular as they can enable users to seamlessly interface with their devices. This has greatly increased accessibility by improving authentication and communication with users and their electronic devices. However, ASRs are vulnerable to adversarial audio samples Schönherr et al. (2019); Qin et al. (2019); Carlini et al. (2016); Carlini and Wagner (2018); Abdoli et al. (2019); Kreuk et al. (2018); Cissé et al. (2017). These are inputs that sound benign to human listeners but force ASRs to produce malicious transcriptions. This allows attackers to force home assistants to execute arbitrary commands Abdullah et al. (2019a); Yuan et al. (2018) or evade surveillance systems Abdullah et al. (2019b). ASRs can be categorized into two broad types: traditional Hannun et al. (2014); Povey et al. (2011); Lamere et al. (2003) and fully end-to-end Sainath et al. (2015); Jaitly and Hinton (2011); Palaz et al. (2013); Tüske et al. (2014); Hoshen et al. (2015); Zeghidour et al. (2018a,b); Fu et al. (2018). While traditional ASRs use signal processing algorithms for feature extraction (e.g., STFT), the newer end-to-end models use additional trainable layers that *learn* to extract the correct features. This divergence from a signal processing based feature extraction has made most of the existing attacks obsolete Qin et al. (2019); Schönherr et al. (2019) to the newer end-to-end ASRs. Unfortunately, the remaining attacks Carlini and Wagner (2018); Abdullah et al. (2020) that do work against the newer models produce low quality adversarial audio samples. Even though, past researchers have used the L_p clipping to control the attack audio quality Abdullah et al. (2020), we show in this paper that such strategies are inherently flawed. They fail to consider psychoacoustics (i.e., the functioning of the human ear), resulting in noisy audio.

We overcome these limitations and present a single unified attack that not only works against both ASR types (traditional and end-to-end) but also creates high quality adversarial audio. Our equalization-based attack uses the masking thresholds generated via the psychoacoustic model to control the quality of the perturbations. Constraining via the masking thresholds ensures that resulting perturbations will be largely imperceptible to the human ear. We evaluated our attack against both traditional and fully end-to-end ASRs. Specifically, we attacked DeepSpeech (traditional) and Wav2letter (fully end-to-end) and achieved success rates of 100% and 76%, respectively. Upon further investigation, we observed that these success rates are comparable to existing optimization attacks Carlini and Wagner (2018) on such architectures. We then ran a user study to determine whether our attack produces less low audible distortion than the existing state-of-the-art attack due to Carlini and Wagner (CW) Carlini and Wagner (2018), which can also exploit end-to-end models. However, in contrast to our work, their attack does not capitalize on psychoacoustics to minimize audible distortion. Our user study showed that 80 of the 100 participants found *all* of our attack audio samples to be less noisy when compared to those of the CW attack. Our work demonstrates that, while improvements to ASRs architectures have made many existing attacks obsolete, they remain vulnerable to attacks.

2. Related Work

The process of generating adversarial audio samples has generally followed one of two approaches. The first is to model the ASR as an optimization function. Since the ASR learns an optimal mapping between audio and intended output, attacking the optimization directly is a viable approach when given knowledge of the target ASR internals Bispham et al. (2018). With such knowledge, an adversary can directly optimize an adversarial objective function that maximizes the error of the ASR model's objective function Carlini et al. (2016). This is practical in a variety of settings, including voice assistants Carlini et al. (2016); Yuan et al. (2018), speech-to-text systems Cissé et al. (2017); Schönherr et al. (2019); Carlini and Wagner (2018), music content analysis Kereliuk et al. (2015), and speaker-verification systems Kreuk et al. (2018). The second is the signal processing approach, which assumes that certain physical properties of audio are modeled by the ASR. Often, the model of these physical properties is imperfect, or does not faithfully re-create signals that are likely to be encountered in the wild Vaidya et al. (2015); Zhang et al. (2017); Du et al. (2019). An attacker can also leverage properties of the audio pipeline (e.g., FFT windows, MFCC parameters) to generate adversarial audio directly for a desired psychoacoustic effect Abdullah et al. (2019a).

Attacks that leverage knowledge of human psychoacoustics may produce higher quality adversarial audio. These "psychoacoustic" attacks use the information about the human ear to craft samples with low audible distortion. Thus, to a human, the resulting adversarial audio samples sound similar to the original audio. These attacks a) assume knowledge of the model internals, and b) leverage the underlying psychoacoustic properties that control the modeling of audio. The knowledge required for such attacks can vary. Researchers have proposed certain white-box approaches Qin et al. (2019); Schönherr et al. (2019) which assume that the ASR is using a particular signal processing algorithm (e.g., STFT) is being used for feature extraction. Our work relaxes this requirement and simply assumes the victim model is a fully end-to-end ASR model, which need not contain any such signal processing algorithm for feature extraction.

3. Background

3.1. Short-Time Fourier Transform

The Short-Time Fourier Transform (STFT) provides the frequency information of a given signal over time (i.e., a spectrogram). This is accomplished via the sum of a sequence of Fourier transforms of overlapping windowed blocks of the time-domain signal. In this paper, we denote the application of an STFT over an audio sample $\boldsymbol{x}$ as: $S(\boldsymbol{x}) = \text{STFT}_{m,n}\{\boldsymbol{x}\}$. The STFT output is a $m \times n$ matrix of complex values and m and n are the time and frequency indices respectively.

For many applications, it is convenient to modify the STFT representation of the signal directly (as opposed to its time-domain representation). These include noise cancellation Boll (1979), source separation Virtanen (2007), time-scale and pitch-scale modifications Laroche and Dolson (1999). Typically, only the squared magnitude of an STFT is modified, and the phase is either left unaltered or discarded Perraudin et al. (2013). It is important to note that, due to the overlapping frames of the STFT, modifications of the magnitude prevent a perfect STFT inversion (i.e., STFT is non-invertible).

The most fundamental and widely used Shen et al. (2018); Wang et al. (2017) technique for spectrogram inversion is the Griffin-Lim algorithm Griffin and Lim (1984). This is an iterative algorithm that *approximates* the time-domain representation of the spectrogram. This technique is widely regarded as the most effective method of reconstructing an audio signal from its modified spectrogram. We denote this operation as: $\boldsymbol{x}' = \text{Griffin-Lim}(S, k)$. Here, $\boldsymbol{x}'$ is the approximated time-domain signal for spectrogram $S(\boldsymbol{x})$, and k is a parameter of the algorithm that controls the quality of the reconstructed signal.

3.2. Feature Extraction in ASRs

Several techniques have been used in the ASR pipeline. Here, we distinguish between *traditional* approaches that use signal processing algorithms for features extraction, and *fully end-to-end* approaches that directly learn features from data.

Traditional approaches use signal processing algorithms (e.g., STFT and Mel-frequency Cepstral Coefficients (MFCC) Mannell (1994)) to extract features from raw audio samples (Figure 1 (a)). The extracted features are then passed to a model for inference, which can either be a neural network Hannun et al. (2014); Amodei et al. (2016) or a Hidden Markov Model-Gaussian Mixture Model Povey et al. (2011); Lamere et al. (2003). However,

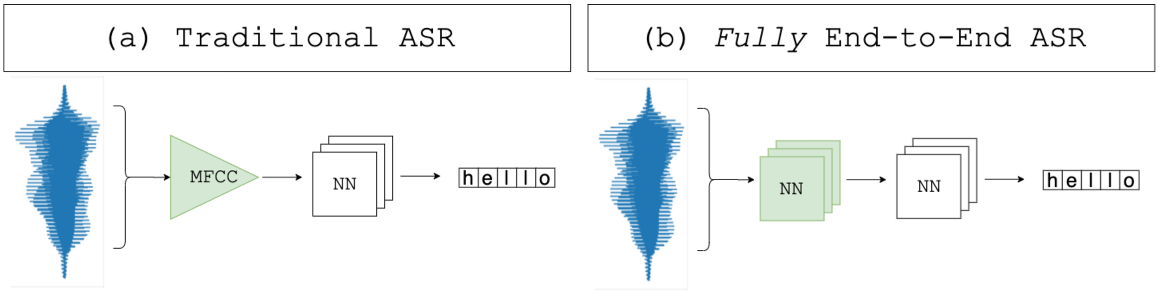

Figure 1: The architectures of (a) traditional and (b) fully end-to-end ASRs. The primary difference between the two is that fully end-to-end ASRs do not include any signal processing-based feature extraction algorithms (e.g., MFCC). Instead, they use NNs that learn to extract the appropriate features during model training.

hand-crafting features may not maximize model accuracy Sainath et al. (2015), as important features may be considered unimportant and discarded Ilyas et al. (2019).

In contrast, more modern approaches do not directly use traditional signal processing for feature extraction (Figure 1 (b)). Instead, the ASR pipeline incorporates additional (trainable) neural network layers in the model which *learn* to extract features to maximize accuracy Sainath et al. (2015); Jaitly and Hinton (2011); Palaz et al. (2013); Tüske et al. (2014); Hoshen et al. (2015); Zeghidour et al. (2018a,b); Fu et al. (2018). Thus, a *fully* end-to-end ASR model consists of a single element (i.e., a neural network) that operates directly on the raw audio waveform. Since the currently published psychoacoustic attacks depend on the existence of a signal processing-based feature extraction layer, none of them work against the newer fully end-to-end models. In this work, we propose a new psychoacoustic attack that overcomes this limitation and produces high-quality audio against fully end-to-end models.

3.3. Psychoacoustic model

Audio Perception Psychoacoustic models have been developed to capture the functioning of the human auditory system. In particular, psychoacoustic models tell us that there are two phenomena that affect the perception of changes made to an audio sample.

1. **Frequency Masking:** Consider a signal that is composed of two tones. Frequency masking is when the softer tone (maskee) is inaudible in the presence of the louder tone (masker). However, increasing the tone of the maskee beyond a certain threshold can make it audible, despite the presence of a masker. As the intensity of the maskee is below the threshold, it will remain largely imperceptible.

2. **Variable Perception:** The human ear perceives audio in a variable manner. Studies in the field of human audio perception have resulted in widely familiar models of frequency perception; the Mel Scale Stevens et al. (1937) and the BARK Scale Zwicker (1961) [1] ,

[1] A visual plot of the BARK is provided in our supplementary materials.

the later of which we use in our masking frequency calculations described in Algorithm 1. Additionally, studies in this field have also produced models of perceived loudness, with the most popular being the ISO 226 Equal-Loudness Contours International Organization for Standardization (ISO) (2003).These auditory models demonstrate that lower frequencies (e.g., 100Hz to 200Hz) are perceived as less intense than higher ones (e.g., 10,000Hz to 20,000Hz) at the same actual intensity. In addition to this, it is also more difficult for humans to perceive small differences in frequency in high frequency tones as opposed to low frequency tones. As a result, lower frequencies of an audio sample can be perturbed to a relatively larger degree without resulting in perceptible distortion. In this case, distortion is an alteration that degrades the quality of the original signal.

Algorithm 1 Psychoacoustic PGD Attack Steps

```
$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x}$ ;
$S_{mag} = |\text{STFT}(\boldsymbol{x})|$ ;
$M = \text{generate\_thresholds}(S_{mag})$ ;
while $f(\boldsymbol{x}_{\text{adv}}) = t$ do
    $\boldsymbol{\delta} = \text{sign}(\nabla l(\theta, \boldsymbol{x}_{\text{adv}}, t))$
    $\Delta = \text{STFT}(\boldsymbol{\delta})$ ;
    for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$: do
        $\Delta'_{i,j} = |M_{i,j}| \frac{\Delta_{i,j}}{|\Delta_{i,j}|}$
    end
    $\boldsymbol{\delta}' = \text{Griffin-Lim}(\Delta', k)$ ;
    $\boldsymbol{x}_{\text{adv}} = \boldsymbol{x}_{\text{adv}} + \boldsymbol{\delta}'$
end
```

Masking Thresholds Both phenomena can be accounted for through the use of *masking thresholds*. These define the maximum degree by which each frequency can be perturbed without any perceivable audible difference. For example, if a frequency intensity is increased beyond what is defined in the threshold, it will become audible to the human ear. Therefore, one can use this to ensure that the frequencies of the perturbations remain below the masking thresholds, thereby maintaining audio quality.

Masking thresholds are determined using the frequencies present in an audio signal, and thus they vary throughout the audio [2]. We can calculate these thresholds using a sliding window, which is moved by fixed-sized increments. This is done by taking the magnitude of each complex coefficient of the STFT representation of a signal, which we refer to as: $[S_{\text{mag}}]_{i,j} = |[S]_{i,j}| \in \mathbb{R}^{m \times n}$

This is then followed by the calculation of the masking thresholds:

$$M = \text{generate\_thresholds}(S_{\text{mag}})$$

[2] In most cases audio, specially recorded speech, the signal will consist of multiple frequencies. However, we note that it is technically possible for a recorded speech signal to consist of a singular frequency, but is highly uncommon

Here $\boldsymbol{x}$ is an audio sample, S_{mag} is its magnitude spectrum and M is the matrix of the corresponding masking thresholds. For more details, interested readers can refer to Lin and Abdulla (2015).

Psychoacoustic Attacks Attacks that leverage psychoacoustic information can generate higher quality attack audio (i.e., ones with lesser audible distortion) Schönherr et al. (2019); Qin et al. (2019). By ensuring that the perturbation remains below the masking thresholds, the attacks produce audio whose distortions are largely imperceptible.

Psychoacoustic Attacks vs Fully End-to-End ASRs However, a major limitation of current psychoacoustic attacks Schönherr et al. (2019); Qin et al. (2019) is that they require the presence of the traditional signal-processing-based feature extraction layer in the target ASR[3]. This layer could be based on either STFT or the MFCC algorithms. The attack calculates and psychoacoustically scales the gradients, which are in the frequency domain at this layer. Since psychoacoustic scaling operates in the frequency domain, the ASR pipeline must contain a traditional feature extraction layer. However, there is no such layer in an end-to-end ASR, rendering existing psychoacoustic attacks obsolete.

4. L_p Clipping and Psychoacoustics

One popular method to control adversarial perturbation is the L_p clipping Madry et al. (2017); Kurakin et al. (2016); Abdullah et al. (2020). It uses the `clip` to "cut" the amplitude of a signal beyond a maximum threshold (Figure 2(b)). L_p clipping to control the magnitude of attack perturbation Abdullah et al. (2020), which It has been widely successful in the space of adversarial images Goodfellow et al. (2016); Madry et al. (2017). However, we argue that this method should not be applied to the audio domain by evaluating it with regards to the principles of psychoacoustics. This will motivate an alternate method to L_p clipping.

4.1. Hard Clipping & Psychoacoustics

The popular L_p clipping method can be categorized as "hard clipping" technique. Though such techniques have seen popular use in the image space Madry et al. (2017); Kurakin et al. (2016), they should not be used in the adversarial audio space as they do not create imperceptible perturbations. This is due to two reasons. First, hard clipping is considered to be extreme since it cuts the signal at a maximum threshold (Figure 2(b)). This process modifies the waveform of a signal, which distorts an audio signal and may introduce undesirable artifacts such as harmonics. The artifacts that are introduced by hard clipping are typically unpleasant or jarring to human listeners.

Second, hard clipping does not account for human audio perception, thus failing to produce adversarial audio that capitalizes on the human perception of audio. As discussed in the previous section, human audio perception is a function of psychoacoustic phenomena such as frequency masking and variable perception of loudness. Therefore, adversarial audio samples generated via hard clipping will always have greater perceivable distortion than attacks that account for psychoacoustics.

[3] We contacted the authors of current psychoacoustic attacks Schönherr et al. (2019) who confirmed our hypothesis.

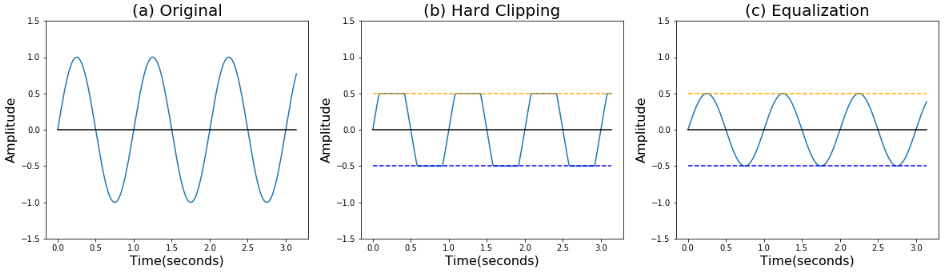

Figure 2: The figure above demonstrates the effects on an of (b) Hard Clipping and (c) Equalization on an (a) Original signal. We can see that (b) Hard Clipping results in a distortion of the audio signal, making the audio noisier. In contrast, (c) Equalization reduces the intensity of a single frequency in the signal, avoiding any audible distortion. It is important to note that equalization does not just reduce the intensity of the whole signal, but only a particular frequency of the signal.

4.2. Equalization Approach

To overcome these limitations, we propose an equalization-based method. This method will help control the adversarial audio quality, but will not introduce auditory artifacts that can degrade quality. Instead of cutting audio sample at a fixed amplitude value (Figure 2(b)), equalization can change the intensity of individual frequencies in a signal (Figure 2(c)). This helps avoid the undesirable audio effects introduced by the L_p clipping method. If used in accordance with the psychoacoustic model, equalization can help control the adversarial perturbation so that it conforms to the human model of hearing. As a result, the perturbation can remain largely imperceptible, resulting in better quality adversarial audio.

4.3. Attack Details:

Provided is a brief overview of how we compute the frequency masking threshold based on psychoacoustics for imperceptible adversarial examples. We begin the process by first identifying all of the maskers as tonal in order to ensure that the threshold that we compute can always mask out the noise. We then calculate the individual masking thresholds, which is done using frequencies with respect to the BARK scale, as the spreading functions of the masker would be similar at different "Barks". Finally, we compute the global masking threshold which is a combination of the individual masking thresholds and the threshold of moments of silence and pauses in speech.

5. Adversarial Audio Attack

We now describe an equalization-based psychoacoustic attack against ASRs. It overcomes the limitations of the existing psychoacoustic attacks and successfully exploits a fully end-to-end model (by ignoring the feature extraction layer). Additionally, it capitalizes on the psychoacoustics of the human ear, thereby producing audio that is of better quality than L_p-clipping algorithms.

5.1. Problem Statement and Threat Model

Consider a time domain audio sample $\boldsymbol{x}$, a target transcription y and an ASR model $f(\cdot)$. The attacker's goal is to produce an adversarial audio sample $\boldsymbol{x}_{\text{adv}}$ that sounds like $\boldsymbol{x}$ to the human ear, but is transcribed as y by f. Specifically, the attacker is successful if the two following goals are achieved.

1. **Targeted Transcription:** the ASR transcribes $\boldsymbol{x}_{\text{adv}}$ as y (i.e., $f(\boldsymbol{x}) \neq f(\boldsymbol{x}_{\text{adv}}) = y$).
2. **Imperceptibility:** the human ear is unable to semantically distinguish between $\boldsymbol{x}_{\text{adv}}$ and $\boldsymbol{x}$.

Following prior work Carlini et al. (2016); Carlini and Wagner (2018); Qin et al. (2019); Schönherr et al. (2019), we consider a white-box threat model. However, in contrast to existing psychoacoustic attacks, we do *not* need the target ASR to contain a hand-crafted feature extraction layer (i.e., fully end-to-end model).

5.2. Attack Overview

We craft the adversarial audio using an iterative process. Upon each iteration, we generate a perturbation $\boldsymbol{\delta}$ to a benign audio sample $\boldsymbol{x}$, resulting in $\boldsymbol{x}_{\text{adv}}$: $\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \epsilon\boldsymbol{\delta}$ (2). Here, ϵ is a scalar between 0 and 1 that controls the magnitude of the perturbation. The perturbation $\boldsymbol{\delta}$ shares the dimensions of the audio $\boldsymbol{x}$ and is generated using the following two steps:

1. **Generating a candidate perturbation:** Using the gradient of the model loss (on the target transcription y) with respect to the input audio, we produce a *candidate* perturbation $\boldsymbol{\delta}$. Adding this perturbation to the current version of the audio results in an output closer to the target transcription in the decision space.
2. **Constraining the perturbation:** We constrain the candidate perturbation $\boldsymbol{\delta}$ according to the masking thresholds to ensure that the perturbation is largely imperceptible.

5.3. Attack Formulation

For the first step, we generate a candidate perturbation by looking at the sign of the gradient of the loss with respect to the input audio. For the second step, we propose a novel set of equalization-based techniques to constrain the perturbation according to psychoacoustics.

Generating Perturbation To produce a candidate perturbation $\boldsymbol{\delta}$, we use the Fast Gradient Sign Method (FGSM) Goodfellow et al. (2014): $\boldsymbol{\delta} = \text{sign}(\nabla l(\theta, \boldsymbol{x}, y))$. Here, θ denotes the model's parameters, $l(\cdot)$ is the loss function, and ∇ denotes the gradient so that $\nabla l(\theta, \boldsymbol{x}, y)$ is the gradient of the loss with respect to the input.

Constraining Perturbation As done in prior work Abdullah et al. (2020), a candidate perturbation $\boldsymbol{\delta}$ could simply be added to the audio sample, scaled by some parameter ϵ (Equation 2). This scalar multiplication results in equal scaling of all frequencies in the perturbation signal by ϵ. However, this would ignore psychoacoustics as the human ear perceives frequencies variably, resulting in a noisy perturbation. To better reflect psychoacoustics, lower frequencies should be scaled with a higher constant since lower frequencies are perceived as softer than higher ones. This can be achieved by first constraining the frequencies of the candidate perturbation $\boldsymbol{\delta}$ according to the masking thresholds of the psychoacoustic model. This is followed by the application of Equation 2. This will help achieve the desired quality of audio:

1. **Frequency Representation:** We constrain the perturbation $\boldsymbol{\delta}$ using the masking thresholds M, which we precompute according to the psychoacoustic model. The thresholds will define the maximum allowable intensity by which we can perturb each frequency in the audio without having an impact on the audio quality. However, we cannot directly use the masking threshold for two reasons. The masking thresholds are in the (magnitude-only) time-frequency domain (due to the application of S_{mag} shown in Equation 1), whereas the perturbation $\boldsymbol{\delta}$ is in the time domain (Equation 2).

   We apply the STFT to transform $\boldsymbol{\delta}$ to the time-frequency domain. The STFT maintains the phase information, which is an integral part of accurately reconstructing an audio sample from its frequency representation. Reconstructing a time-domain audio sample using only the magnitude information can result in distortion. To produce adversarial audio that best adheres to the principals of psychoacoustics, the phase information must be maintained while making modifications to the spectrogram representation of an audio sample. Therefore, we use the S instead of S_{mag}: $\Delta = S(\boldsymbol{\delta}) \in \mathbb{C}^{m \times n}$. Here, Δ is the spectrogram for perturbation $\boldsymbol{\delta}$.

2. **Equalization:** Now, we use equalization (4.2) to constrain the perturbation according to the masking thresholds M in the frequency domain. We equalize the spectrogram of the perturbations (Δ) for the indices i, j that violate the masking thresholds. We scale the real-imaginary pair of the complex number, thereby maintaining the phase:

$$\Delta'_{i,j} = |M_{i,j}| \frac{\Delta_{i,j}}{|\Delta_{i,j}|} \in \mathbb{C}^{m \times n} \tag{5}$$

   where $\Delta'_{i,j}$ is the scaled value, i the frequency bin, and j is the time index. Since Equation 5 is equalizing using the masking thresholds M, this method will produce higher quality audio, while avoiding the distortion that is a result of L_p based clipping methods. Additionally, this method will work in the absence of a traditional feature extraction layer, thereby overcoming the limitation of existing psychoacoustic attacks.

3. **Time Domain Reconstruction:** Since Δ' is a *modified* time-frequency representation of Δ, it does not have a perfect time-domain representation (Section 3.1). We obtain an approximate reconstruction via the Griffin-Lim algorithm (Section 3.1). The reconstructed time-domain perturbation $\boldsymbol{\delta}'$ is then added to the original audio sample to produce an adversarial sample.

	Which one is noisier?			*How does the noisy audio sample (CW) differ from the other one (us)?*			
Audio Samples	*Noisy CW (%)*	*Noisy Us (%)*	*Both sound same (%)*	*Perceptible but not annoying (%)*	*Slightly annoying (%)*	*Annoying (%)*	*Very annoying (%)*
That is comparatively nothing	83	10	7	30.12	48.19	20.48	1.20
Talking later is beneath us	89	6	5	5.62	29.21	44.94	20.22
But there seemed no	97	3	0	10.31	36.08	41.24	12.37
Been looking up tombs county	83	3	14	32.53	50.60	16.87	0
Foul mouth fellow at the top	93	6	1	2.15	27.96	37.63	32.26
Tied to a woman	85	6	9	27.06	49.41	17.65	5.88

Table 1: Breakdown of participants' selection of noisier audio across the audio samples and subjective assessment of noise difference in a pair of audio samples. More than 80% of participants selected each of the CW's audio samples to be noisier than ours (left). Participants' subjective audio perception rating between CW and our audio samples are recorded(right). About 70% - 98% (aggregating ratings for slightly annoying, annoying, and very annoying) of the participants found CW's audio samples to be annoying.

6. Experimental Setup

6.1. ASR Model

To demonstrate that our attack exhibits cross-architecture generalization, we run it against both a traditional model and a fully end-to-end one. For the traditional ASR, we use DeepSpeech[4] and for the fully end-to-end one, we employ Wav2Letter Collobert et al. (2016).

6.1.1. DeepSpeech

The model pipeline consists of an STFT stage followed by convolutions and bi-directional RNNs. The model was trained on the LibriSpeech data set to achieve the state-of-the-art word error rate of 8% moz (2017).

DeepSpeech is not a fully end-to-end model because it includes a signal processing-based feature extraction layer (i.e., STFT). However, we can treat it as such by simply ignoring the STFT stage (as if it does not exist) and applying the attack directly on the raw audio. This highlights that our attack works against both types of models (i.e., those that are fully end-to-end and those that are not). In other words, our attack *does not require the signal processing based feature extraction layer.*

6.1.2. Wav2Letter

This is a fully end-to-end model consisting of stacked layers of convolutions (without any recurrent layer)[5]. We trained this model on the LibriSpeech data set to achieve the state-of-the-art word error rate of 7%. Since there is no signal processing-based feature extraction layer, this model is not vulnerable to existing psychoacoustic attacks Qin et al. (2019); Schönherr et al. (2019).

[4] Code available at: `https://github.com/mozilla/DeepSpeech`

[5] Despite the lack of a recurrent layer, which is commonly found in most ASR architectures, we specifically chose Wav2Letter due to an explicit request by a previous reviewer. This is important to note, since the lack of the RNN might reduce attack success for optimization attacks (Section 7.1)

6.2. Dataset and Attack Parameters

Following the experimental setup outlined in Qin et al Qin et al. (2019), we randomly sample 100 audio samples from the LibriSpeech Panayotov et al. (2015) test set, each of which is perturbed through our attack using 1000 iterations. Each audio sample is perturbed to produce 50 random target phrases sampled from the test set. The attack is successful if the ASR transcription *exactly* matches the target transcription.

We also repeat this experiment for different values of k, which is the number of iterations of the Griffin-Lim algorithm[6]. This parameter influences the quality of audio reconstruction and thus has an impact on perceptibility and attack success. Specifically, we repeat the above-described attack experiment for $k = 1, 2, 4, 8$. This results in a total of 20,000 attack audio samples (100 audio samples × 50 target transcriptions × 4 iteration values) for the entire experimental setup. Such a large number of attack audio files were generated to thoroughly test the viability of our attack.

6.3. User Study

We ran a user study to determine whether our attack perturbation are less noisy than other attacks by comparing it to that of state-of-the-art CW Carlini and Wagner (2018). We chose this attack since it is the only other one that does work against the fully end-to-end model. However, unlike our attack, CW's attack ignores psychoacoustics when perturbing adversarial audio samples.

We hypothesized that participants who listened to both audio samples would find our audio to be less noisy. To test this, 100 participants were recruited using Amazon's Mechanical Turk (MTurk) and compensated two USD upon completion of a Qualtrics survey. They had to read and digitally sign a consent form before they could continue. Then, participants were asked to disclose any hearing issues; however, this did not disqualify them from our study.

The survey contained ten questions about audio samples from our and CW's attack Carlini and Wagner (2018). Participants were presented with two audio samples (Audio Sample A and Audio Sample B) and asked which of the two samples was noisier. Participants were given the option to choose (A) Audio Sample A; (B) Audio Sample B; or (C) The two pieces of audio are the same. If users selected A or B, they were asked a follow-up question: "How does the noisy audio differ from the other one?" Options were (A) Perceptible but not annoying; (B) Slightly Annoying; (C) Annoying: (D) Very Annoying.

In four of the ten questions, Audio Sample A and Audio Sample B were identical, where two were our audio, and two were Carlini's. This was done to test the attentiveness of participants while they were answering the questions. The remaining six questions compared CW's audio to our audio. In these cases the transcription of the audio was the same; only the technique to produce the audio was different. Finally, we asked participants to disclose the device (e.g., laptop, desktop, or other) and headphones types.

7. Experimental Results

To demonstrate the viability of our attack against ASRs, we needed to demonstrate that the two previously outlined attacker goals are met (Section 5.1). Specifically, that the rate

[6] Code available at: `https://librosa.github.io/ librosa/generated/librosa.core.istft.html`.

of successful target transcription and the imperceptibility of the perturbations used in our attack are either equal to or better than other attacks in current literature.

7.1. Attack Success Rate

For all the attack audio files generated with $k = 1$ and 2, the ASR transcribed them as malicious text. When attacking DeepSpeech, our attack was able to achieve a success rate of 100%, which is on par with other attacks in this space Carlini et al. (2016); Carlini and Wagner (2018); Schönherr et al. (2019). However, the success rate with $k = 1$ against Wav2Letter model fell to 76%. Upon further investigation, we observed that the success rate for CW also failed to reach 100%. We believe that this was due to the lack of a recurrent layer or the MFCC in the Wav2Letter architecture, which lead to reduced attack success. Additionally, we discuss why smaller values of k had a higher success rate than larger ones in the supplementary material.

The number of iterations of the Griffin-Lim algorithm influences both the rate of attack success and the time it takes to generate the audio file. We discuss in the supplementary materials how we converge to the optimum value for this variable via empirical tests and manual listening experiments. We found this value to be $k = 1$, and will be using it for the remainder of the experiments.

These results demonstrate the steps of our psychoacoustic attack still converge to the desired solution, with the same accuracy as that of other attacks. This is true even though our attack requires switching between domains (time-frequency) and involves approximating a modified STFT. The time-frequency conversion of the STFT is lossy and equalization is happening in the frequency domain. As a result, if the attack steps are structured the way they are, they would have introduced additional loss and have had prevented convergence.

7.2. User Study

In our MTurk study, 100 participants gave consent and completed the survey entirely. Among them, 61 were male, and 39 were female. The average age of the participants was 35.5. All participants reported being native English speakers. Each of the participants used either a laptop or a desktop for the study and wore a headset for the audio experiment. The median completion time of the study was 3.62 minutes. During the study, each participant was presented with ten pairs of audio samples in random order. If not stated otherwise, the rest of this section presents results made from observations from comparisons of the six pairs of audio samples, one from CW's attack and one from ours. (For details, please see Section 6.3.)

Table 1 shows the results of our study. More than 80% of participants considered all of CW's attack audio samples as noisy. Note that across all samples, the number of participants to consider the CW audio noisier was at least 83%. Table 1 also shows the breakdown of audio perception difference for participants who selected CW to be noisier among the two audio samples. We can see that about 70% - 98% (aggregating ratings for slightly annoying, annoying and very annoying) of the participants found CW's audio samples to be annoying. Meanwhile, only a small proportion (2% - 32%) perceived the audio sample generated by CW's technique to be distinguishable from ours but they did not find it annoying.

To measure the statistical significance of our observations, we ran a χ^2 with the null hypothesis: there is no audio quality difference between CW's and our attack audio samples.

With a *p-value* of less than 0.01, the null hypothesis is rejected and thus shows that there exists a significant noise quality difference between CW and our audio samples.

8. Conclusion

ASRs are vulnerable to adversarial examples. The more potent attacks in this space are the targeted ones based on psychoacoustics. These can produce clean audio samples (i.e., largely imperceptible perturbation) that are transcribed to a targeted transcription. Though these attacks are effective against traditional ASRs, they are obsolete against newer, fully end-to-end ones. This is due to the absence of the traditional, signal-processing-based feature extraction layer. In this paper, we propose an equalization-based attack that leverages signal processing and psychoacoustics to produce clean adversarial audio. As a result, our attack produces less noisy audio than the current state-of-the-art attacks. In the process of developing our attack, we discovered that the L_p clipping method, which has been used in the past Abdullah et al. (2020), is a poor technique for controlling the imperceptibility of a perturbation. Our work has a 100% success rate, produces high-quality audio, and is applicable to both traditional and fully end-to-end ASRs.

9. Acknowledgments

This work is partly supported by the National Science Foundation under CNS-1933208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Deep Speech 0.4.1, 2017. Available at https://github.com/mozilla/DeepSpeech/releases/tag/v0.4.1.

Sajjad Abdoli, Luiz G Hafemann, Jerome Rony, Ismail Ben Ayed, Patrick Cardinal, and Alessandro L Koerich. Universal Adversarial Audio Perturbations. *arXiv preprint arXiv:1908.03173*, 2019.

Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin Butler, and Joseph Wilson. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS)*, 2019a.

Hadi Abdullah, Muhammad Sajidur Rahman, Washington Garcia, Logan Blue, Kevin Warren, Anurag Swarnim Yadav, Tom Shrimpton, and Patrick Traynor. Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. *arXiv preprint arXiv:1910.05262*, 2019b.

Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. *In Submission*, 2020.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

Mary K Bispham, Ioannis Agrafiotis, and Michael Goldsmith. A Taxonomy of Attacks via the Speech Interface. 2018.

Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden Voice Commands. In *USENIX Security Symposium*, pages 513–530, 2016.

Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6980–6990, 2017.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems. *arXiv preprint arXiv:1901.07846*, 2019.

Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1570–1584, 2018.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

Awni Y. 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. *CoRR*, abs/1412.5567, 2014. URL http://arxiv.org/abs/1412.5567.

Yedid Hoshen, Ron J Weiss, and Kevin W Wilson. Speech acoustic modeling from raw multichannel waveforms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4624–4628. IEEE, 2015.

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*, pages 125–136, 2019.

International Organization for Standardization (ISO). Acoustics — Normal equal-loudness-level contours. Technical Report ISO 226:2003, International Organization for Standardization (ISO), Aug 2003.

Navdeep Jaitly and Geoffrey Hinton. Learning a better representation of speech soundwaves using restricted boltzmann machines. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5887. IEEE, 2011.

Corey Kereliuk, Bob L Sturm, and Jan Larsen. Deep Learning and Music Adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071, 2015.

Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling End-to-end Speaker Verification by Adversarial Examples. *arXiv preprint arXiv:1801.03339*, 2018.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Paul Lamere, Philip Kwok, William Walker, Evandro Gouvea, Rita Singh, Bhiksha Raj, and Peter Wolf. Design of the CMU Sphinx-4 decoder. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Jean Laroche and Mark Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio processing*, 7(3):323–332, 1999.

Yiqing Lin and Waleed H Abdulla. Principles of Psychoacoustics. In *Audio Watermark*, pages 15–49. Springer, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083*, 2017.

R.H. Mannell. The perceptual and auditory implications of parametric scaling in synthetic speech. *Macquarie University*, page (Chapter 2), 1994. URL "http://clas.mq.edu.au/speech/acoustics/auditory_representatiopitchdiscrim.html".

Dimitri Palaz, Ronan Collobert, and Mathew Magimai Doss. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv preprint arXiv:1304.1018*, 2013.

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

Nathana el Perraudin, Peter Balazs, and Peter L Søndergaard. A fast griffin-lim algorithm. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

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 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011. IEEE Catalog No.: CFP11SRW-USB.

Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. *arXiv preprint arXiv:1903.10346*, 2019.

Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. The Internet Society, 2019. URL https://www.ndss-symposium.org/ndss2019/.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney. Acoustic modeling with deep neural networks using raw time signal for lvcsr. In *Fifteenth annual conference of the international speech communication association*, 2014.

Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. *WOOT*, 15:10–11, 2015.

Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

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 *Proceedings of the USENIX Security Symposium*, 2018.

Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux. End-to-end speech recognition from the raw waveform. *arXiv preprint arXiv:1806.07098*, 2018a.

Neil Zeghidour, Qiantong Xu, Vitaliy Liptchinsky, Nicolas Usunier, Gabriel Synnaeve, and Ronan Collobert. Fully convolutional speech recognition. *arXiv preprint arXiv:1812.06864*, 2018b.

Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. DolphinAttack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117. ACM, 2017.

Eberhard Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961.