Generating Robust Audio Adversarial Examples with Temporal Dependency

Hongting Zhang[1] , **Qiben Yan**[2] , **Pan Zhou**[1*] and **Xiao-Yang Liu**[3]
[1]Huazhong University of Science and Technology
[2]Michigan State University
[3]Columbia University
htzhang@hust.edu.cn, qyan@msu.edu, panzhou@hust.edu.cn, xl2427@columbia.edu

Abstract

Audio adversarial examples, imperceptible to humans, have been constructed to attack automatic speech recognition (ASR) systems. However, the adversarial examples generated by existing approaches usually incorporate noticeable noises, especially during the periods of silences and pauses. Moreover, the added noises often break temporal dependency property of the original audio, which can be easily detected by state-of-the-art defense mechanisms. In this paper, we propose a new Iterative Proportional Clipping (IPC) algorithm that preserves temporal dependency in audios for generating more robust adversarial examples. We are motivated by an observation that the temporal dependency in audios imposes a significant effect on human perception. Following our observation, we leverage a proportional clipping strategy to reduce noise during the low-intensity periods. Experimental results and user study both suggest that the generated adversarial examples can significantly reduce human-perceptible noises and resist the defenses based on the temporal structure.

1 Introduction

Due to the recent advancement in artificial intelligence (AI) and machine learning, automatic speech recognition (ASR) systems have been integrated into numerous commercial products. Recently, researchers [Vaidya *et al.*, 2015; Carlini *et al.*, 2016] demonstrated the possibility of creating adversarial examples to launch targeted attacks towards ASR systems. The goal of the attack is to force ASR systems to recognize the audio inputs as intelligible voice commands, while human perceives the audio inputs differently. Such attacks have proven to be effective towards ASR systems that use Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), while recently, ASR systems leveraging deep neural networks are also targeted by sophisticated adversarial examples generated by a slight perturbation of the original inputs [Cisse *et al.*, 2017; Kreuk *et al.*, 2018; Gong and Poellabauer, 2017; Yuan *et al.*, 2018].

Although remarkably effective, the generation of audio adversarial examples face two major challenges: 1) the incorporation of non-negligible perturbation during the periods of silences and pauses in adversarial audios; and 2) the lack of robustness in resisting data property based defenses. Silences or pauses are a part of human conversations. However, the adversarial examples generated by existing optimization-based approaches [Carlini and Wagner, 2018], which apply noises over the entire audio, incorporate a non-negligible perturbation in the period of silences and pauses and could alert the users on their existence. Moreover, existing approaches disregard the inherent voice data property when generating adversarial examples, which allows defense mechanisms [Yang *et al.*, 2018] that exploit voice data properties to spot such attacks. One crucial voice data property that has been used in defense mechanisms is *temporal dependency (TD).*

This research is motivated by an observation that TD in audios imposes a significant effect on human perception of audios. We perform a user study to investigate how TD affects our auditory perception of audio samples which even have an equivalent level of noise in decibels. We discover that the perturbed audio with a better preservation of TD presents a higher audio quality by human perception. And this observation motivates us to preserve temporal dependency while generating audio adversarial examples.

In this paper, we propose a new Iterative Proportional Clipping (IPC) algorithm that generates robust adversarial examples with TD. To the best of our knowledge, we are the first to consider TD in designing targeted attacks towards ASR systems. Specifically, we first extract Mel Frequency Cepstral Coefficient (MFCC) features of the input audio and propagate them through the ASR model to compute the objective loss. Then we perform backpropagation to retrieve the gradient, which will be used as the raw perturbation. We add the raw perturbation on the original input, and perform a data-driven proportional clipping operation on the updated signal based on the signal intensities of the original audio. After a series of iterations, the adversarial examples are generated. We demonstrate that our method is successful in attacking the state-of-the-art CNN based ASR system Wav2letter+. Notably, the proposed approach generates adversarial examples in the order of minutes instead of hours as seen in prior approaches [Carlini and Wagner, 2018]. Compared to one recent approach [Schönherr *et al.*, 2018] that leverages sound

[*] Corresponding Author

masking which requires the computation of the complex hearing threshold matrix, our approach is more generic and easier to implement in practice. The impact of the proposed attack lies in the generation of *human-imperceptible adversarial examples* to attack ASR systems without alerting the users.

Our contributions are summarized as follows:

- We propose a new Iterative Proportional Clipping (IPC) algorithm to generate robust audio adversarial examples with temporal dependency to attack ASR systems without alerting the users. We further propose two enhancements to the attacks by hiding noises in the high-intensity or high-frequency components to improve the imperceptiblity of adversarial examples.
- We implement a successful attack on the latest model of an end-to-end CNN based ASR system Wav2letter+ with a differentiable Mel Frequency Cepstral Coefficient (MFCC) features extraction.
- Experimental results show that the adversarial examples are effective even under temporal dependency based defense (TD defense). User study shows that our adversarial examples have the highest audio quality so far.

2 Related Work

2.1 Adversarial Examples

In their seminal work, [Biggio *et al.*, 2013] and [Goodfellow *et al.*, 2014] have shown that neural networks are vulnerable to adversarial examples. Compared with the image field, less efforts have been spent on studying the impact of audio adversarial examples towards neural networks based ASR systems. One type of attack approaches creates a waveform that ASR systems recognize as intelligible voice commands but humans perceive as noise. [Vaidya *et al.*, 2015] first explored adversarial examples against ASR systems, by integrating an audio command into an audio mangler while keeping most of MFCCs intact. This method leads to perceptible sound distortion due to the lossy inversion. [Carlini *et al.*, 2016] constructed white-box attacks via the hidden voice commands on CMU Sphinx speech recognition system, in which they demonstrated HMM-only ASR systems were subject to such targeted attacks. [Zhang *et al.*, 2017], [Yan *et al.*, 2020] proposed DolphinAttacks and SurfingAttack and showed the possibility to hide transcriptions by modulating the baseband audio signal with ultrasound higher than 20 kHz. However, all these methods above cannot generate the adversarial audio waveforms for an end-to-end ASR framework.

Another type of approaches is to deceive the neural networks by introducing minor perturbations on the input. [Carlini and Wagner, 2018] used CTC loss as an objective function and generated adversarial examples using a gradient-descent-based minimization scheme [Carlini and Wagner, 2017]. However, the generated adversarial examples tend to include widely distributed noises, which become noticeable by humans. One recent approach considers psychoacoustics to minimize human perception: [Schönherr *et al.*, 2018] proposed psychoacoustic hiding, which utilizes the hearing thresholds for designing proper perturbations of the input signals by curbing the signal variation below the threshold of human perception. [Abdullah *et al.*, 2019] utilized domain-specific knowledge of audio signal processing to achieve practical black-box attacks by leveraging the fact that humans interpret discontinuous signals as noisy and hardly discern differences in high-frequency signals. These techniques require domain-specific knowledge and complex signal processing, which are difficult to implement.

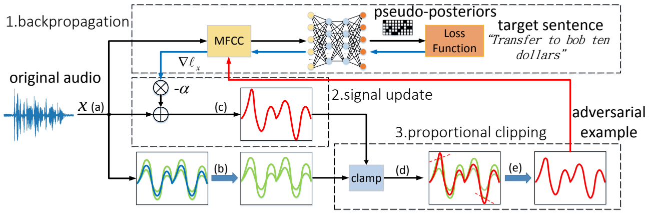

Figure 1: The creation of adversarial examples can be divided into three processes: (1) a backpropagation process to obtain a raw perturbation; (2) a signal update process based on the raw perturbation; (3) a proportional clipping process on the updated signal to constrain the signal within two boundary lines in (b). In the last component, we first calculate two boundary lines from the original waveform in (a). Then, we clip the disturbed waveform in (c) at those positions where the sound intensity goes beyond the area enclosed by the two boundary lines, as shown in (d). The modified waveform is displayed in (e).

2.2 Perceptual Assessment

Research efforts have been spent in developing computational methods for the perceptual assessment of transmission quality of lossy wide-band audio compression techniques as an alternative to costly listening tests [Huber and Kollmeier, 2006]. However, the mechanism of human auditory perception has not been fully explored and simulated, and there is no objective measurement that can completely replace subjective evaluation [Assembly, 1994; Recommendation, 2001]. Previous approaches for attacking ASRs all use noise measured in decibels for quantifying the human perceptibility. In this paper, we demonstrate that other factors could also affect human hearing of audio examples even with the same level of noises. Therefore, there is still room to substantially improve the auditory quality of the adversarial examples.

3 The Effect of Temporal Dependency

In this section, we aim to address the question: *"Does temporal dependency in audio affect human perception of the audio quality?"* Understanding the relationship between the two is pivotal for designing better adversarial examples.

3.1 Generating Testing Audio Samples

To compare audio samples with different degrees of temporal dependency, we construct two sets of audio samples by adding two types of noises on the original audio $\boldsymbol{x} \in \mathbb{R}^n$. The noises $\Delta\boldsymbol{x}$ and $\Delta\boldsymbol{x}'$ are generated as follows:

$$\Delta\boldsymbol{x} = \gamma \cdot \bar{x} \cdot \boldsymbol{r}, \ \Delta\boldsymbol{x}' = \gamma \cdot (\boldsymbol{x} \odot \boldsymbol{r}), \tag{1}$$

where $\boldsymbol{r} \in \mathbb{R}^n$ is a random vector from a uniform distribution in $[0, 1]$, $\bar{x}$ represents the average value of $\boldsymbol{x}$, $\odot$ represents element-wise product, and γ is a parameter to control

	a1	a2	a3	$\widehat{\text{a1}}$	$\widehat{\text{a2}}$	$\widehat{\text{a3}}$
$\boldsymbol{x}+\Delta\boldsymbol{x}$	0.9889	0.9872	0.9888	0.9249	0.9621	0.9622
$\boldsymbol{x}+\Delta\boldsymbol{x}'$	**0.9992**	**0.9980**	**0.9992**	**0.9917**	**0.9964**	**0.9965**

Table 1: Cross-correlation coefficient between the original audio (a1, a2, a3) and the perturbed audio samples.

$\gamma/\bar{x}$	Percentage of users choosing the second set
0.15	98.0%
0.2	100.0%
0.25	96.0%

Table 2: Results of the user study.

the degree of perturbation. For each set, we construct three sequences of audio samples with different degrees of perturbation, with each sequence containing 20 audio samples. In total, we have 60 pairs of audio samples for testing. In addition, we set the maximum intensity of noises in both sets as the same decibel value.

Since the added noise in the second set of samples is proportional to the original audio $\boldsymbol{x}$ as presented in Eq. (1), it resembles the original audio in terms of temporal dependency property. We use the cross-correlation coefficient to measure temporal dependency in these generated samples [Olden and Neff, 2001]. Table 1 presents the cross-correlation coefficient of three randomly chosen original audio samples, denoted as a1, a2 and a3 and their respective perturbed samples. The result shows that the perturbed samples in the second set has a higher cross-correlation coefficient with the original audio. Moreover, we extract the low-intensity components from the original audio, denoted as $\widehat{\text{a1}}$, $\widehat{\text{a2}}$ and $\widehat{\text{a3}}$, and the result shows the second set has a better preservation of temporal dependency in original audios, especially in their low-intensity components.

3.2 User Study

We generate 60 pairs of audio samples with samples from test-clean set of LibriSpeech dataset, which is introduced in Section 5, and perform a user study with 20 volunteers including seven postgraduates and thirteen undergraduates. We present each volunteer with the original audio and two perturbed samples from the first set and the second set. We then ask the volunteer to compare which of the two samples is closer to the original audio. For every volunteer, we execute the test with at least four different pairs of samples. The results in Table 2 show that most of the volunteers consider the audio samples in the second set are closer to the original audio with less perceptible noise, even with different degrees of perturbations. As a result, there does exist a positive correlation between the temporal dependency and human perception of the audio. The results imply that high temporal dependency in perturbed audio samples improve the audio quality.

4 Robust Audio Adversarial Example Generation

Motivated by the observation in Section 3, we propose an Iterative Proportional Clipping (IPC) algorithm to generate adversarial examples, and show how the audible noise can be reduced by preserving temporal dependency during the creation of adversarial examples. We further propose two enhancements similar to [Yakura and Sakuma, 2018] to further improve the imperceptibility of adversarial examples.

4.1 Iterative Proportional Clipping Algorithm

We formulate the problem of constructing an adversarial example as an optimization problem. Given a trained ASR model $f : \mathbb{R}^n \rightarrow \mathbb{R}^{u\times v}$ and a decoder $d : \mathbb{R}^{u\times v} \rightarrow \mathbb{R}^w$, we modify $\boldsymbol{x}$ with the minimal distortion so that the decoder recognizes the sample as a target sentence $\boldsymbol{t}$ rather than the original decoding result $\boldsymbol{y} \in \mathbb{R}^w$. The main challenge lies in the following dilemma: a larger distortion achieves better performance in altering the original phrase toward the target phrase, but it also leads to lower audio quality. To address this challenge, we propose an objective loss function ℓ for generating an audio adversarial example as follows:

$$\begin{aligned} &\text{minimize } \lambda_1 \cdot Loss(f(\phi(\boldsymbol{x}+\boldsymbol{\delta})), \boldsymbol{t}) + \lambda_2 \|\boldsymbol{\delta}\|_2^2, \\ &s.t. \ |\frac{\boldsymbol{\delta}_i}{\boldsymbol{x}_i}| < B, \ i \in [n], \end{aligned} \tag{2}$$

where $[n]$ denotes $\{1, 2, ..., n\}$, and the $\phi(\cdot)$ denotes a feature extractor which extracts MFCC features, log-mel filterbank energies features, or power spectrum features. The Connectionist Temporal Classification (CTC) loss is used in the first item in Eq. (2) to achieve adversarial attack and L2 distortion is used in the second item for reducing noise. The parameters λ_1, λ_2 are used to control the relative importance of being adversarial and remaining similar to the original audio. And $B \in [0, 1]$ is the intensity width of the perturbation. The goal of adversarial example generation is to minimize the loss function subject to the perturbation constraint.

The intensity width B in Eq. (2) is used to maintain the proportionality between the perturbation $\boldsymbol{\delta}$ and the original waveform $\boldsymbol{x}$. Because the intensity width makes the clipping thresholds of each position in the perturbed audio proportional to the original waveform, the clipped waveform has a higher similarity in terms of shape[1] with the original one, which better preserves the temporal dependency. Related experimental validation is presented in Section 5. Note that the proposed objective loss function does not need to be iterated many times as in previous approaches [Carlini and Wagner, 2018]. The novel constraint condition guarantees to reach a proper solution quickly.

As shown in Figure 1, our algorithm can be integrated into the CNN-based speech recognition process: during each iteration, we first apply backpropagation to get the raw perturbation to be added to the input, and then perform an proportional clipping operation on the modified input. We can see that the waveform in red in (e) after proportional clipping has a more similar trend with the original waveform. The pseudocode of the proposed algorithm is presented in Algorithm 1. In essence, IPC limits the perturbation within a certain range determined by the intensity width B in Eq. (2). The

[1] The shape of an audio waveform stands for the trend in audio time-series data, which is one of the three aspects of temporal dependency property: temporal closeness, period and trend.

Algorithm 1 IPC Attack

Input: Original sample $\boldsymbol{x} \in \mathbb{R}^n$, target sentence $\boldsymbol{t}$, trained ASR model f, decoder d, B

Output: Adversarial example $\boldsymbol{x}^{adv}$

```
Initialize $S \leftarrow False$, $\hat{\boldsymbol{x}} \leftarrow \boldsymbol{x}$, $\boldsymbol{\delta} \leftarrow \boldsymbol{0}$,
while $S = False$ do
    calculate the loss $\ell$,
    calculate the gradient $\nabla \ell_{\boldsymbol{x}} \in \mathbb{R}^n$ on the input $\boldsymbol{x}$,
    $\boldsymbol{\delta} \leftarrow \nabla \ell_{\boldsymbol{x}}$,
    for $i \in [n]$ do
        clip $\boldsymbol{\delta}_i$ within $[-\boldsymbol{x}_i \cdot B,\ \boldsymbol{x}_i \cdot B]$ as $\boldsymbol{\delta}'_i$,
    end for
    $\boldsymbol{x} \leftarrow \boldsymbol{\delta}' + \hat{\boldsymbol{x}}$,
    if $d(\boldsymbol{x}) = \boldsymbol{t}$ then
        $S \leftarrow True$,
        $\boldsymbol{x}^{adv} \leftarrow \boldsymbol{x}$,
    end if
end while
return $\boldsymbol{x}^{adv}$.
```

tolerable perturbation of the original audio will increase with a larger B value. When B is more than 1, IPC degenerates into Opt [Carlini and Wagner, 2018] .

4.2 Adversarial Examples with Noise Hidden in High-intensity Voice Components

The perturbation during silent and low-intensity periods such as the beginning and the end of a speech seriously degrades the quality of the speech. IPC decreases but does not completely remove such perturbation, which could alert the users once the perturbation exceeds a certain hearing threshold.

Therefore, we propose to hide commands into certain high-intensity voice segment based on IPC. This could avoid perceptual noise in the low-intensity periods. For a benign audio $\boldsymbol{x}$ that has high intensity in the first k timestamps and low intensity in the remaining timestamps. In essence, we only apply the perturbation on the high-intensity components (i.e., the first k timestamps) to minimize the objective loss function. Our attack can be formalized as follows:

$$\begin{aligned} \boldsymbol{\delta}^* &= \underset{\boldsymbol{\delta}}{argmin}\ \lambda_1 \cdot Loss(f(\phi(\boldsymbol{x}_{[k]} + \boldsymbol{\delta}), \boldsymbol{t}) + \lambda_2 \|\boldsymbol{\delta}\|_2^2, \\ s.t.\ &|\frac{\boldsymbol{\delta}_i}{\boldsymbol{x}_i}| < B,\ i \in [k], \end{aligned} \tag{3}$$

where $\boldsymbol{\delta}^* \in \mathbb{R}^k$ represents the perturbation of the audio segment $\boldsymbol{x}_{[k]}$.

$$\boldsymbol{x}^{adv} = [\boldsymbol{x}_{[k]} + \boldsymbol{\delta}^*\ \ \boldsymbol{x}_{[n]\setminus[k]}], \tag{4}$$

where $\boldsymbol{x}_{[n]\setminus[k]}$ denotes the components from $(k+1)$-th timestamp to n-th timestamp, [] represents the concatenation of the high-intensity and low-intensity segments. After concatenating the adversarial segment with the remaining ones, we derive the complete audio adversarial example $\boldsymbol{x}^{adv}$.

4.3 Adversarial Examples with Noise Hidden in High-frequency Voice Components

Humans have different perception capability to acoustic signals of different frequency bands [Fayek, 2016]. The noises within low frequency band is more likely to be better perceived by humans, allowing them to discover the existence of adversarial attacks. Moreover, the noises within low frequency band such as human voice frequency band from 300Hz to 3,400Hz also seriously influences the audio quality.

Therefore, we propose to perturb the audio input only in certain high frequency band. During each iteration, we only backpropagate the perturbation from certain frequency band to update the input and freeze the gradient from the other frequency bands. This is similar to the band-pass filter in [Yakura and Sakuma, 2018]. Based on empirical observations, we set the band to 3,500Hz to 8,000Hz, which results in less perceptual noise. And the adversarial example generation can be formalized as follows:

$$\begin{aligned} &\text{minimize}\ \ \lambda_1 \cdot Loss(f(\phi(\boldsymbol{x} + \boldsymbol{\delta})), \boldsymbol{t}) + \lambda_2 \|\boldsymbol{\delta}\|_2^2, \\ &s.t.\ |\frac{\boldsymbol{\delta}_i}{\boldsymbol{x}_i}| < B,\ i \in [n],\ where\ \boldsymbol{\delta} = \frac{\partial \ell}{\partial \boldsymbol{X}_{[q]\setminus[p]}} \frac{\partial \boldsymbol{X}_{[q]\setminus[p]}}{\partial \boldsymbol{x}}. \end{aligned} \tag{5}$$

Here, $\boldsymbol{X}$ denotes the output of the Fourier transform of the modified input in the previous iteration, and ℓ denotes the objective loss in Eq. (5). p and q represent the lower and upper frequency bounds of the considered frequency band, and the subscript range $[q] \setminus [p]$ along the frequency domain of $\boldsymbol{X}$, is used to explicitly limit the frequency range of the perturbed signals to be between p and q. By considering both the temporal dependency and frequency components, we generate high-quality adversarial examples.

5 Experiments

5.1 Dataset and Adversarial Model

Dataset. LibriSpeech [Panayotov *et al.*, 2015] is a corpus of approximately 1,000 hours of 16 KHz English speech derived from audiobooks from the LibriVox project. It comes with its own training, validation sets, test-clean and test-other sets. We use all available samples to train and validate our ASR system. We generate adversarial examples only using its test-clean set, which contains 2,620 waves with the average duration of 4.294s.

Adversarial model. Wav2letter [Collobert *et al.*, 2016] is an efficient end-to-end ASR system released by the Facebook AI research team. Based on the architectures in [Collobert *et al.*, 2016] and [Liptchinsky *et al.*, 2017], NVIDIA proposes Wav2letter+ which consists of 17 1D-Convolutional Layers and 2 Fully Connected Layers and uses CTC loss for training. During the training, it extracts log-mel filterbank energies as the input features to the model and outputs a pseudo-posteriors matrix, with each element representing the possibility of each alphabet label at each step. During the speech recognition, it decodes with beam search decoder and outputs a sequence of letters corresponding to the speech input.

5.2 Implementation Detail

In this research, we implement Wav2letter+ in Pytorch as our adversarial model. Different from the Wav2letter+ specification, we use a differentiate MFCC features extraction preceding the ASR model. We use "torch.rfft" to convert signal to the frequency domain. All experiments are carried out on

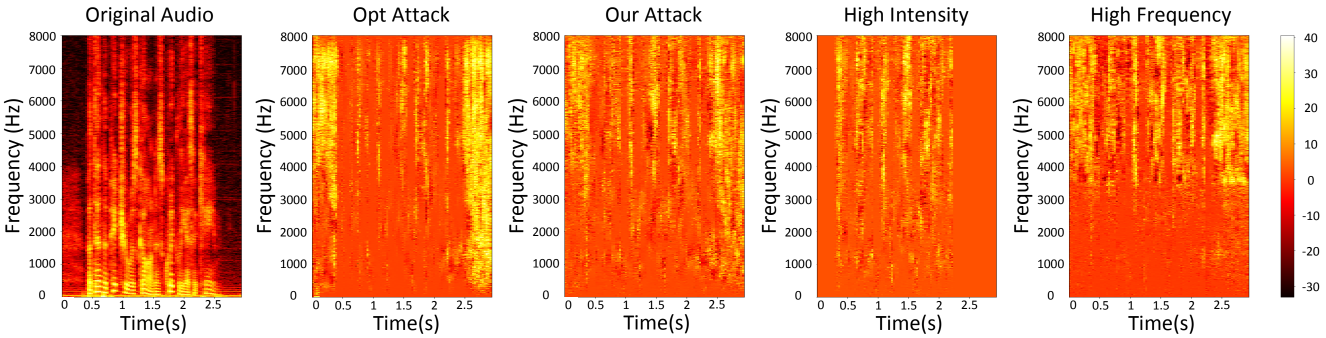

Figure 2: This figure displays the STFT of an audio sample, and Opt's perturbation, our perturbation, two types of perturbations on the audio sample when embedding noise in the high-intensity components and high-frequency components. The lighter colors indicates a higher intensity at a given frequency and time. Our adversarial perturbation has lower intensities in low-intensity segments than Opt.

Type	Transcribed results
Orig	but then the picture was gone as quickly as it came
half	but then the picture was
Adv	open alipay
half	open
Adv	transfer to bob ten dollars
half	transfer to bob
Adv	please delete the last transaction record
half	please delete the last

Table 3: Maliciously translated examples ('Adv') in different lengths from a general audio waveform('Orig'). 'half' denotes the result of TD defense by setting cut ratio as 0.5.

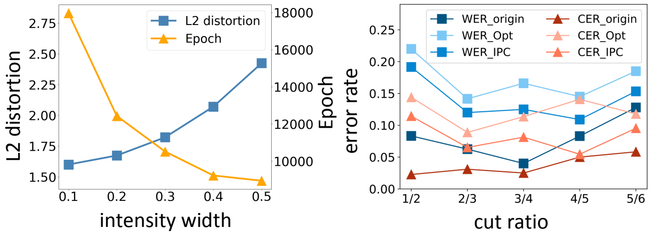

Figure 3: (a) L2 Distortion and epoch with different intensity widths. (b) Word Error Rate and Character Error Rate of audio examples.

an Ubuntu Server (16.04 LTS) with an Intel Core i5-6500@ 3.20GHz × 4, 16G Memory and GTX 1080 GPU.

For the input sample, we randomly select one audio sample from the test-clean set as the original audio. For the target sentence, we select the sentence corresponding to another audio sample or one random sentence as the target. We use a "Coarse-to-fine" strategy to reduce the time consumption of beam search decoder by splitting the generation process into two stages. In the first stage, we use a greedy decoder and generate a weak adversarial example under the train mode of the ASR system, which is an approximation of the true adversarial example; and then, in the second stage, we turn to beam search decoder and fine tune it to get the desired adversarial example under the eval mode of the ASR system. In our experiments, we set the learning rate as $1e^{-5}$ in the first stage and $5e^{-5}$ in the second stage.

5.3 Results

Table 3 shows the maliciously translated examples in different lengths from a general audio waveform. To investigate the impact of intensity width B in Eq. (2), we generate adversarial examples with different intensity widths and explore the trend of two key metrics in Figure 3(a): L2 distortion and Epoch. L2 distortion is chosen for quantifying the distortion introduced by the perturbation and the number of epochs implies the requisite time for generating an adversarial example. We can see that L2 distortion descends slowly while the number of requisite epochs grows as the B narrows down from 0.5 to 0.1. To strike a balance between epochs and distortion, we set the width B to 0.2 in all the following experiments to generate adversarial examples with a high quality while reducing runtime cost.

5.4 Comparison with Previous Work

First, our proposed algorithm makes adversarial audio samples resemble the original audio, which are difficult to discover by human hearing due to psychoacoustic principles. Figure 2 displays the Short-time Fourier Transform (STFT) of original audio sample, both Opt's and our perturbation on the original audio sample in the left three figures. We can see that the perturbation produced by Opt has greater intensities in the low-intensity and silent periods. In contrast, our perturbation has smaller intensities across the spectrogram especially in low-intensity periods, which makes our adversarial audio perceptually more appealing than Opt based adversarial audio.

We also show the STFT of the perturbation of our enhanced attacks in the right two plots in Figure 2. It can be seen that all perturbation exists in the high-intensity speech periods for the first enhanced attack, which greatly decreases widely distributed noises. For second enhanced attack, most perturbations are limited in the high frequency bands, while there is almost no perturbation during low frequency bands within 0Hz-3,500Hz, the most important frequency band for humans to interpret voice commands.

In addition, the proposed approach consumes less time costs to accomplish an attack. As the classical paper [Car-

		Hear target sentence	Hear anything abnormal	Hear nothing abnormal
Opt	b1	0.0%	40.0%	60.0%
	b2	0.0%	55.0%	45.0%
IPC	b1	0.0%	30.0%	70.0%
	b2	0.0%	35.0%	65.0%
Enh1	b1	0.0%	25.0%	75.0%
	b2	0.0%	10.0%	90.0%
Enh2	b1	0.0%	15.0%	85.0%
	b2	0.0%	25.0%	75.0%

Table 4: Results of the user study of adversarial examples generated from Opt, IPC and two enhanced attacks, denoted as Enh1 and Enh2 (b1 and b2 are two different original samples).

	c1	c2	c3
Opt	0.9160	0.9255	0.9170
IPC	0.9900	0.9970	0.9985
High-intensity	0.9941	0.9973	**0.9990**
High-frequency	**0.9987**	**0.9980**	0.9989

Table 5: Cross-correlation coefficient between original audio (c1, c2, c3) and adversarial examples based on different methods.

width (B)	c1	c2	c3
0.5	0.9873	0.9852	0.9889
0.4	0.9884	0.9857	0.9891
0.3	0.9886	0.9882	0.9895
0.2	0.9900	0.9970	0.9985
0.1	**0.9993**	**0.9989**	**0.9992**

Table 6: Cross-correlation coefficient between original audio (c1, c2, c3) and our adversarial examples with different intensity widths.

k	IPC (Opt)		
	WER	CER	LCP
$1/2$	0.524 (0.930)	0.507 (0.933)	0.609 (0.806)
$2/3$	0.770 (0.930)	0.700 (0.948)	0.885 (0.826)
$3/4$	0.573 (0.933)	0.510 (0.938)	0.835 (0.839)
$4/5$	0.575 (0.955)	0.553 (0.969)	0.772 (0.880)
$5/6$	0.755 (0.941)	0.680 (0.962)	0.766 (0.858)

Table 7: AUC scores of different k on adversarial examples based on IPC and Opt.

lini and Wagner, 2018] has demonstrated, generating a single adversarial example requires approximately one hour of computation time on commodity hardware (a single NVIDIA 1080Ti). [Schönherr *et al.*, 2018] requires less than two minutes to calculate the adversarial perturbations with 500 backpropagation steps but do not specify the exact time for a successful attack. In contrast, to accomplish a successful attack, our time consumption is in the minute-level, usually in 3-15 minutes on commodity hardware (a single NVIDIA 1080), without any complex computations.

5.5 User Study and Analysis

In order to confirm if humans could notice an attack, we conduct a user study. We let 20 volunteers to listen to six adversarial audio samples in three sets which correspond to IPC attack and two enhanced attacks respectively, with each set containing two adversarial examples. The results are shown in Table 4. Although a small number of volunteers felt the abnormality, most of them heard nothing abnormal, and all the participants could not hear the target sentences. We also show the result of Opt based adversarial examples for comparison.

To determine the amount of perceptible noise, we also calculate the cross-correlation coefficient between original audio samples and adversarial examples as our measure of the perceptibility of noise, which quantifies the temporal dependency property. The comparison of IPC with Opt and two enhanced attacks are shown in Table 5. The best solution tends to appear in two enhanced attacks. We also list the cross-correlation coefficient with different intensity widths on these three audio samples in Table 6 to investigate the effect of intensity width. It clearly demonstrates IPC results in better perception by leveraging a stricter intensity width.

5.6 Performance Against TD Defense

To validate the robustness of our adversarial examples, we evaluate their performance under TD defense [Yang *et al.*, 2018]. We follow the same experimental procedures as [Yang *et al.*, 2018], and adopt their evaluation metrics: the area under curve (AUC) of word error rate (WER), AUC of character error rate (CER), and AUC of longest common prefix (LCP). Table 3 lists some examples of translated results for benign and adversarial audios with the cut ratio $k = 0.5$. We can see that segments of our adversarial examples can be equally transcribed with the corresponding transcription of the whole adversarial examples. Moreover, Figure 3(b) shows that IPC has lower average WER and CER than Opt. We list the AUC scores of three basic metrics under different cut ratios on our adversarial examples in Table 7. The resulted AUC scores for our adversarial examples are mostly distributed between 0.5 and 0.7, which means that the classifier with TD defense has poor performance on our adversarial examples. It demonstrates the robustness of our adversarial examples against TD defense due to the preservation of temporal information.

6 Conclusion

In this paper, we propose a new Iterative Proportional Clipping algorithm to generate robust adversarial examples for ASR attacks. By iteratively performing proportional clipping on the perturbation which we compute from the backpropagated gradient through ASR model, we force the modified audio waveform to maintain the trend in the original audio and thus obtain adversarial examples with temporal dependency. The user study demonstrates that IPC significantly constraints the noise on the original audio. Our experiments show that IPC based adversarial examples can not only compromise the state-of-the-art Wav2letter+ model but also bypass the latest temporal information based defense mechanisms.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grant 61972448. The work of Yan is supported in part by National Science Foundation award CNS-1950171.

References

[Abdullah *et al.*, 2019] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. *arXiv preprint arXiv:1904.05734*, 2019.

[Assembly, 1994] ITU Radiocommunication Assembly. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1994.

[Biggio *et al.*, 2013] 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*, pages 387–402. Springer, 2013.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[Carlini and Wagner, 2018] 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.

[Carlini *et al.*, 2016] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 513–530, 2016.

[Cisse *et al.*, 2017] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

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

[Fayek, 2016] Haytham Fayek. Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between, 2016.

[Gong and Poellabauer, 2017] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280*, 2017.

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

[Huber and Kollmeier, 2006] Rainer Huber and Birger Kollmeier. Pemo-q—a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on audio, speech, and language processing*, 14(6):1902–1911, 2006.

[Kreuk *et al.*, 2018] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1962–1966. IEEE, 2018.

[Liptchinsky *et al.*, 2017] Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. based speech recognition with gated convnets. *arXiv preprint arXiv:1712.09444*, 2017.

[Olden and Neff, 2001] Julian D Olden and Bryan D Neff. Cross-correlation bias in lag analysis of aquatic time series. *Marine Biology*, 138(5):1063–1070, 2001.

[Panayotov *et al.*, 2015] 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.

[Recommendation, 2001] ITUR Recommendation. Methods for objective measurements of perceived audio quality. *ITU-R BS*, 13871, 2001.

[Schönherr *et al.*, 2018] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

[Vaidya *et al.*, 2015] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: exploiting the gap between human and machine speech recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, 2015.

[Yakura and Sakuma, 2018] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. *arXiv preprint arXiv:1810.11793*, 2018.

[Yan *et al.*, 2020] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided wave. In *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[Yang *et al.*, 2018] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing audio adversarial examples using temporal dependency. *arXiv preprint arXiv:1809.10875*, 2018.

[Yuan *et al.*, 2018] 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 *27th USENIX Security Symposium (USENIX Security 18)*, pages 49–64, 2018.

[Zhang *et al.*, 2017] 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, 2017.