PHASEFOOL: PHASE-ORIENTED AUDIO ADVERSARIAL EXAMPLES VIA ENERGY DISSIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio adversarial attacks design perturbations onto inputs that lead an automatic speech recognition (ASR) model to predict incorrect outputs. Current audio adversarial attack algorithms optimize perturbations with different constraints (e.g. l_p-norm for waveform or auditory masking for magnitude spectrogram) to achieve their imperceptibility. Since phase is not relevant for speech recognition, the existing audio adversarial attacks neglect the influence of phase spectrogram. In this work, we propose a novel phase-oriented algorithm named PhaseFool to efficiently construct imperceptible audio adversarial examples. Specifically, we leverage the phase spectrogram and the spectrogram consistency of short-time Fourier transform (STFT) to adversarially dissipate the energy that is crucial for ASR systems. Since the magnitude spectrogram plays a dominant role in human perception, the phase-oriented perturbations cause tiny auditory influences. Experimental results demonstrate that PhaseFool can inherently generate full-sentence imperceptible audio adversarial examples with 100% targeted success rate within 500 steps on average (2.5x speed-up over current state-of-the-art imperceptible counterparts), which is verified through a human study. Most importantly, our attack is the first to exploit the phase-oriented energy dissipation rather than add perturbations on the audio waveform like most previous works.[1]

1 INTRODUCTION

Adversarial examples are specifically designed instances that cause a machine-learning algorithm to produce a misclassification (Szegedy et al., 2013; Biggio et al., 2013). Since researchers have successfully created adversarial examples to make automatic speech recognition (ASR) system transcribe the audio as any different sentence (Carlini & Wagner, 2018; Vaidya et al., 2015; Carlini et al., 2016), audio adversarial attacks have become more and more popular. Those attack algorithms usually optimize perturbations added on the benign audio waveform with different constraints based on l_p-norm (Carlini & Wagner, 2018; Zhang et al., 2020). Although the perturbations introduced is small in magnitude, noise in the audio adversarial example is usually perceptible. Later, algorithms based on psychoacoustic principle (Qin et al., 2019; Schönherr et al., 2019) successfully generate effectively imperceptible adversarial examples by hiding the adversarial perturbations to regions of magnitude spectrogram where it will not be heard by a human.

In current imperceptible targeted audio adversarial attacks, only waveform and magnitude spectrogram are taken into account leaving phase and its nature unexplored. Since the structure of natural speech is highly complicated, these two types of algorithms face several challenges:

- Waveform-oriented attacks are usually not imperceptible enough. Although constraining perturbations with l_2-norm or l_∞-norm (Carlini & Wagner, 2018; Liu et al., 2020), these attacks introduce audible noises to the adversarial examples. Later works focused on the temporal dependency (Zhang et al., 2020) successfully reduce human-perceptible noises but their perturbations are still perceptible, due to the uncorrelated relationship between psychoacoustic perception and l_p-norm.
- The optimization of attacks based on magnitude spectrogram is highly time-consuming and hard to converge. Due to the determining influence of magnitude spectrogram on human

[1] https://xxx.github.io/xxxexample/

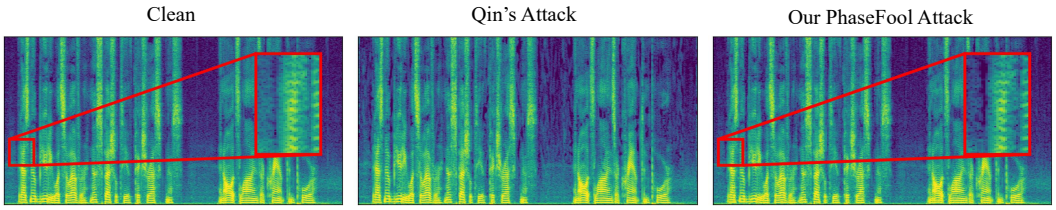


Figure 1: Visualizations of the ground-truth mel-spectrograms and the mel-spectrograms generated by Qin's attack (Qin et al., 2019) and our PhaseFool attack. Qin's attack leverage the phychoacoustic principle of auditory masking and hide noises in the inaudible regions while PhaseFool attack dissipates the energies that is important for classification while less important for human perception.

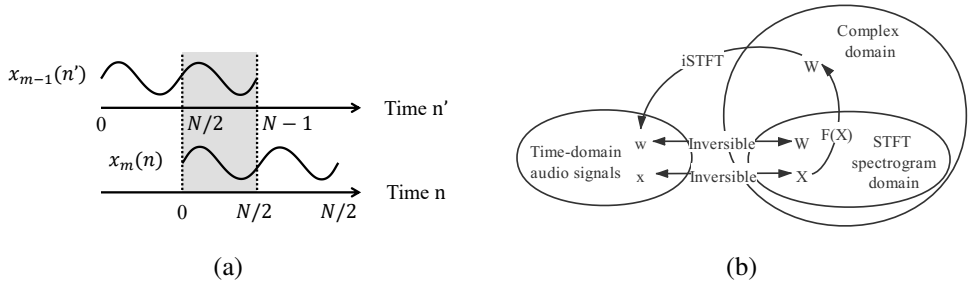
perception, the attack that leverages the psychoacoustic principle of auditory masking (Qin et al., 2019) successfully generates imperceptible audio adversarial examples. However, it needs two-stage optimization where the first stage of optimization focuses on finding a relatively small perturbation to fool the network (as was done in prior work (Carlini & Wagner, 2018)) and the second stage makes the adversarial examples imperceptible. The search process of the optimal balance between hiding the adversarial perturbations in the inaudible regions and fooling the ASR systems is empirically difficult. Generating a single full-sentence adversarial example requires approximately 5000 iterations until it finally converges to an imperceptible solution.

This research is motivated by a psychoacoustic principle that phase manipulation in one-channel audios imposes tiny effects on human perception (Vary & Eurasip, 1985). Besides, although phase is not relevant for speech recognition, the inherent spectrogram consistency of short-time Fourier transform (STFT) reveals that the adversarial perturbations in a frame of phase spectrogram will transfer to adjacent frames of magnitude spectrogram. And these observation motivates us to exploit phase spectrogram for generating audio adversarial examples.

Considering the relationship between phase spectrogram and the spectrogram consistency of STFT, in this work, we propose a novel phase-oriented algorithm called PhaseFool that efficiently generates imperceptible audio adversarial examples via energy dissipation. We prove that the phase-oriented perturbations always reduce the magnitude of the spectrogram in the corresponding areas via spectrogram consistency, which is called energy dissipation. Leveraging such a phenomenon, we bring new directions for audio adversarial attacks and are able to perform a bi-directional attack that dissipates the crucial energy for ASR models and meanwhile, construct the perturbations for the adversarial attack. To the best of our knowledge, we are the first to consider phase spectrogram in the audio adversarial attacks. Our contributions can be summarized as follows:

- We leverage the spectrogram consistency of STFT to construct phase-oriented audio adversarial examples, which make further steps towards imperceptibility and efficiency.
- We further investigate the relationship between phase perturbations and magnitude spectrogram, which is called energy dissipation. Instead of the common l_p distance metrics or magnitude-oriented metrics based on the psychoacoustic principle, we propose a phase-oriented metric based on the phenomenon of energy dissipation.
- Further investigating properties of PhaseFool, we observe additional insights about the vulnerability of ASR systems. It brings a better interpretation of audio adversarial attacks, i.e. the perturbation positions reveal that which part of one magnitude spectrogram is important but also fragile for the corresponding ASR system.

We conduct experiments on the LibriSpeech dataset (Panayotov et al., 2015) and implement a successful attack against a Transformer-based ASR system (Winata et al., 2020b). The results show that in terms of imperceptibility, PhaseFool surpasses current state-of-the-art algorithms (verified by a human study). Furthermore, it can generate full-sentence audio adversarial examples with the 100% targeted success rate within 2 minutes and achieves 2.5x speedup over the current state-of-the-art imperceptible counterpart. We attach some audio adversarial examples generated by our method in the supplementary materials.



(a) (b)

Figure 2: The Illustration of spectrogram consistency.

2 BACKGROUNDS

In this section, we briefly overview the background of this work, including audio adversarial examples and phase-aware audio signal processing.

Audio Adversarial Examples. After the early works (Gong & Poellabauer, 2017; Cissé et al., 2017) towards neural networks based ASR systems successfully generate untargeted adversarial examples that produce incorrect but arbitrary transcriptions. Recently, more and more efforts have been made in the audio domain (Alzantot et al., 2018; Carlini & Wagner, 2018). Among them, one of the most outstanding studies is the iterative optimization-based attack (Carlini & Wagner, 2018), which introduces a two-stage attack and directly generates adversarial examples using a gradient-descent minimization method. Nevertheless, the generated adversarial examples tend to include noticeable noise. Thus, two recent approaches (Qin et al., 2019; Schönherr et al., 2019) introduce hearing thresholds for designing inaudible perturbations by curbing the signal variation below the threshold of human perception. However, most audio adversarial attacks generate perturbations into audio waveform which suffers from a slow convergence rate.

Phase-aware Audio Signal Processing. The interest of utilizing phase information in the field of audio signal processing is constantly increasing (Mowlaee et al., 2016). The STFT phase has been widely used in signal enhancement (Miyahara & Sugiyama, 2014; Choi et al., 2019; Hu et al., 2020) and signal reconstruction (Mowlaee & Kulmer, 2015). The phase in harmonic domain is used in the speech synthesis (Degottex & Erro, 2014; Ai & Ling, 2020) and the phase in perceptual bands is used in speech watermarking (Wang et al., 2021). The important information in the phase brings positive contributions to the above fields. Nevertheless, the existing audio adversarial attacks directly neglect the phase spectrogram, due to the fact that phase is not relevant for speech recognition. In this work, instead of constructing adversarial examples based on l_p-norm, we propose to inherently constrain the perturbations with the phase spectrogram and spectrogram consistency.

3 PHASE-ORIENTED AUDIO ADVERSARIAL EXAMPLES VIA ENERGY DISSIPATION

In this section, we first define the preliminaries of phase-oriented audio adversarial attack. Then we investigate the relation between phase spectrogram and magnitude spectrogram under spectrogram consistency and prove that phase perturbations always reduce the magnitude of spectrogram in the corresponding areas under spectrogram consistency, which is called energy dissipation (illustrated in Figure 1). Finally, we design a novel phase-oriented audio adversarial attack based on energy dissipation.

3.1 PRELIMINARIES OF PHASE-ORIENTED AUDIO ADVERSARIAL ATTACK

Denote the ASR classification model as $f : \{X\} \to Y$, with $X \in [0,1]^{t\times c}$ being the audio space. $Y \in \{0,1,...,k\}^{n\times 1}$ is the sequence of the k-class output space corresponding to the transcription. Specifically, denote $x \in X$ as the input audio, $y \in Y$ as the ground-truth label, and $y_t \in Y$ as

the target transcription, and $\mathcal{L}_{net}(f(x), y_t)$ be a loss function for the threat model that is minimized when $f(x) = y_t$. Then the general formulation of adversarial attack for ASR system is:

$$\underset{\delta}{\operatorname{argmin}} \quad \mathcal{L}_{net}(f(x+\delta), y_t) + \lambda \|\delta\|_p^p \ , \ \text{s.t.} \|\delta\|_\infty < \epsilon \ , \tag{1}$$

where $\mathcal{L}_{net}$ is the loss function used in the ASR model, $\lambda > 0$ is the trade-off parameter between the imperceptibility and adversarial loss (where in some formulations λ = 0), ϵ control the maximum perturbation value, and p could be specified as different values, according to the attacker's requirement (e.g. $p = 2$ in C&W's attack Carlini & Wagner (2018)).

Denote ϕ as the STFT and ϕ^{-1} as the inverse STFT and $X = \phi(x) = |X| \cdot e^{i\theta_x}$, where $|X|$ and $\theta_x) = \angle X$ representing the spectral magnitude and the instantaneous phase spectrogram, respectively. We assume that window function w, window length N and window shift R is fixed and a signal can be perfectly reconstructed from its STFT spectrogram through the inverse STFT in this setting. We propose to generate perturbations on the instantaneous phase spectrogram and reconstruct the adversarial signals with the inverse STFT as the following formulation:

$$\underset{\delta}{\operatorname{argmin}} \quad \mathcal{L}_{net}(f(\phi(\phi^{-1}(|X| \cdot e^{i\theta_x} \cdot e^{j\delta}))), y_t), \ \text{s.t.} -\pi \leq \delta \leq \pi \ . \tag{2}$$

As shown in the Figure 2, the perturbed spectrogram $X' = |X| \cdot e^{i\theta_x} \cdot e^{j\delta}$ is usually located outside the STFT spectrogram domain. Considering the phenomenon of spectrogram consistency (Le Roux et al., 2010), it is easy to prove the following two conclusions: 1) $X' \neq \phi(\phi^{-1}(X'))$ mathematically; 2) $\phi^{-1}(X')$ and $\phi^{-1}(X)$ represent the same time-domain audio signals. Since the STFT and iSTFT functions are differentiable, these conclusions enable us to adversarially manipulate the reconstructed magnitude spectrogram by introducing perturbations to the original phase spectrogram, which we called phase-oriented audio adversarial attack. Then we detailed how the perturbations on a frame of phase spectrogram influence the adjacent frames of corresponding magnitude spectrogram and reveal the energy dissipation nature of phase-oriented audio adversarial attack.

3.2 Relation Analysis between phase and magnitude

This subsection mainly discusses the essence of the relationship between phase perturbations and magnitude spectrogram frames. Reconstructing the time-domain samples after introducing perturbations to the phase spectrogram includes an overlap-add process. For simplicity, we assume that the window shift length R is equal to half of the window length N, which means the adjacent frames have 50% overlapped signals. As shown in Figure 2, the signals $x_{m-1}(n')$ and $x_m(n)$ in adjacent frames $m-1$ and m overlap, where n' and n represent the time indices and $n' = n + \frac{N}{2}$. $x_{m-1}(n')$ and $x_m(n)$ are obtained by an inverse Fourier transform as the following:

$$x_{m-1}(n') = \phi^{-1}(X_{m-1}) = \frac{1}{N} \sum_{k=0}^{N-1} X_{m-1}[k] e^{j\frac{2\pi k n'}{N}} \ , \tag{3}$$

$$x_m(n) = \phi^{-1}(X) = \frac{1}{N} \sum_{k=0}^{N-1} X_m[k] e^{j\frac{2\pi k n}{N}} \ , \tag{4}$$

where j is the imaginary unit, $X_{m-1}[k]$ and $X_m[k]$ are spectrogram samples in frames $m-1$ and m at k-th frequency bin. Then we introduce perturbations δ_{m-1}, δ_m to the frame $m-1$, m and the corresponding signals becomes:

$$\hat{x}_{m-1}(n') = \frac{1}{N} \sum_{k=0}^{N-1} X_{m-1}[k] e^{j\frac{2\pi k n'}{N}} e^{j\delta_{m-1}[k]} \ , \tag{5}$$

$$\hat{x}_m(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_m[k] e^{j\frac{2\pi k n}{N}} e^{j\delta_m[k]} \ . \tag{6}$$

Since the human's auditory system is relatively less sensitive to the tiny time delay resulted by phase rotation, the phase perturbations are inherently more imperceptible than the perturbations on

the audio waveform. Then we further investigate how phase perturbations affect the magnitude spectrogram. The reconstructed signal $\tilde{x}_m(n)$ in the frame m is obtained as:

$$\begin{aligned}\tilde{x}_m(n) &= \hat{x}_{m-1}(n') + \hat{x}_m(n) \\ &= \frac{1}{N}\sum_{k=0}^{N-1} X_{m-1}[k] e^{j\frac{2\pi k n'}{N}} e^{j\delta_{m-1}[k]} + \frac{1}{N}\sum_{k=0}^{N-1} X_m[k] e^{j\frac{2\pi k n}{N}} e^{j\delta_m[k]} \ ,\end{aligned} \tag{7}$$

We substitute $X_{m-1}[k]$ and $X_m[k]$ with the Fourier transform of $x_{m-1}(n')$ and $x_m(n)$, then the equation 7 will become:

$$\begin{aligned}\tilde{x}_m(n) &= \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{n'=0}^{N-1} x_{m-1}(n') e^{-j2\pi\frac{n'}{N}k}\right) e^{j2\pi\frac{n'}{N}k} e^{j\delta_{m-1}[k]} \\ &+ \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{n=0}^{N-1} x_m(n) e^{-j2\pi\frac{n}{N}k}\right) e^{j2\pi\frac{n}{N}k} e^{j\delta_m[k]} \\ &= \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{n'=0}^{N-1} x_{m-1}(n') e^{j\delta_{m-1}[k]} + \sum_{n=0}^{N-1} x_m(n) e^{j\phi_m[k]}\right) \\ &= \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{n'=N/2}^{N-1} x_{m-1}(n') e^{j\delta_{m-1}[k]} + \sum_{n=0}^{N/2-1} x_m(n) e^{j\phi_m[k]}\right) \ .\end{aligned} \tag{8}$$

The overlapping area of the $x_{m-1}(n')$ and $x_m(n)$ in Figure 2 satisfy the following relationship based on spectrogram consistency:

$$x_{m-1}(t + \frac{N}{2}) = x_m(t), \ s.t. \ 0 \le t \le \frac{N}{2} \ . \tag{9}$$

With equation 9, we can deduce the final formulation for the reconstructed signal $\tilde{x}_m(n)$ as following:

$$\begin{aligned}\tilde{x}_m(n) &= \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{n=0}^{N/2-1} x_m(n) e^{j\delta_{m-1}[k]} + \sum_{n=0}^{N/2-1} x_m(n) e^{j\delta_m[k]}\right) \\ &= \frac{1}{N}\sum_{n=0}^{N/2-1} x_m(n) \sum_{k=0}^{N-1}\left(e^{j\delta_{m-1}[k]} + e^{j\delta_m[k]}\right) \ ,\end{aligned} \tag{10}$$

where the clean audio can be represented by setting $\delta_{m-1}[k] = 0$ and $\delta_m[k] = 0$. Therefore, the ratio $\gamma_m[k]$ of the reconstructed signal's spectrogram energy to the clean audio's satisfies:

$$\gamma_m[k] = \frac{\left|e^{j\delta_{m-1}[k]} + e^{j\delta_m[k]}\right|}{\left|e^{j0} + e^{j0}\right|} = \frac{\sqrt{2 + 2\cos(\delta_{m-1}[k] - \delta_m[k])}}{2} \ . \tag{11}$$

Since $\gamma_m[k]$ is ranged within $[0, 1]$, it means that the phase perturbations $\delta_{m-1}[k], \delta_m[k]$ always cause loss of the reconstructed magnitude spectrogram, which is called 'energy dissipation'. Till now, we reveal the bi-directional essence of phase-oriented audio adversarial attack: it dissipates the critical energy for the classification model and constructs the adversarial perturbations in the meantime.

3.3 Optimization with Energy Dissipation

Simply minimizing the l_p distortion between $1 - \gamma_m[k]$ and its upper bound 1 is a possible solution. However, the l_p based metrics yield more audible perturbations in the practical attacks due to the differences between l_p distance and human perceptibility. In this paper, we use the psychoacoustic model of non-simultaneous masking to constrain the perturbations in an imperceptible range. As is described in Lin & Abdulla (2015), the non-simultaneous masking refers to the phenomenon that louder audible sound may mask another preceding or following audible sound, which includes pre-masking and post-masking. Since the duration of pre-masking is quite short (about 20 ms), we only utilize the post-masking phenomenon.

According to the characteristics of human auditory systems, we divide the spectrogram into four parts: 1) the inaudible signals; 2) the signals masked by post-masking (i.e. they are inaudible either); 3) the signals that did not masked by post-masking; 4) the maskers of post-masking, which are the local maxima in the spectrum. For 1) and 2), we do not constrain the corresponding γ since the energy dissipation in these areas can not be perceived. And for 3) and 4), we assign trade-off parameters β_1 and β_2 in the loss function, respectively. The details of dividing criteria can be referred at Lin & Abdulla (2015).

Loss function We formulate the construction of imperceptible phase-oriented adversarial examples as minimizing the following loss function:

$$\mathcal{L} = \mathcal{L}_{net}(f(\phi(\phi^{-1}(|X| \cdot e^{i\theta_x} \cdot e^{j\delta}))), y_t) + \alpha \mathcal{L}_{imp}(\gamma) \ , \tag{12}$$

$$\mathcal{L}_{imp}(\gamma) = \sum_{m=0}^{N-1} \sum_{k=0}^{N-1} \mathcal{I}_3(\beta_1 \cdot max\{1 - \gamma_m[k], 0\}) + \sum_{m=0}^{N-1} \sum_{k=0}^{N-1} \mathcal{I}_4(\beta_2 \cdot max\{1 - \gamma_m[k], 0\}) \ , \tag{13}$$

4 Experimental Setup

4.1 Datasets

We conduct experiments on the LibriSpeech dataset (Panayotov et al., 2015), which contains 1000 hours of 16KHz English speech derived from audiobooks. The train-clean-100 and train-clean-360 datasets are used to train the threat ASR model. Then we randomly select 100 audio examples from the test-clean dataset as source examples, and 1000 separate transcriptions from the same dataset to be the targeted transcriptions. We ensure that each target transcription is around the same length as the original transcription because it is unrealistic and overly challenging to perturb a short audio clip (e.g., 10 words) to have a much longer transcrip- tion (e.g., 20 words). Examples of the original and targeted transcriptions are available in the supplementary material.

4.2 Settings and Metrics

We implement an end-to-end speech recognition model based on the transformer architecture following Winata et al. (2020a); Zhou et al. (2018); Winata et al. (2019). The transformer model consists of four encoder layers and four decoder layers with a dim_{inner} of 2048, dim_{model} of 512, and dim_{emb} of 512. The number of attention heads is set to 8. The encoder module uses a VGG net [17] with a 6-layer CNN architecture. The VGG consists of convolutional layers that are added to learn a universal audio representation and generate input embedding. The input of the unit is a spectrogram. Since the feature extraction is differentiable, our phase attack can be easily extended to the ASR models that take other acoustic features as input.

As is done in most previous works, we assume a white-box threat model, i.e. the adversary is allowed to compute gradients through the model in order to generate adversarial examples. Considering that there are many ways to convert the black-box model to a white-box model Chen et al. (2020) and to improve the robustness against distortions Yakura & Sakuma (2019); Liu et al. (2020) and defencesZhang et al. (2020), our work focus on the white-box model, as is done in most previous works.

Following Qin et al. (2019), we calculate the Word error rate (WER) metric which is defined as $WER = \frac{S+D+I}{N} \times 100\%$, where S, D, and I are the numbers of substitutions, deletions, and insertions, respectively, and N is the total number of words. Particularly, we choose the target transcription as the label for calculating WER in the adversarial attack.

5 Results

In this section, we evaluate the performance of PhaseFool in terms of imperceptibility, generation speedup and interpretation of attacks.

Method	Q1	Q2	Q3	Mel.Err
C&W's Attack				
IPC				
Imperceptible	43.75% / 41.75%	39.75%	75.25%	
PhaseFool	38.25% / 42.25%	4.25%	74.00%	

Table 1: Results of the user study for imperceptibility. Q1, Q2, Q3 is the three of adversarial examples generatedfrom Opt, IPC and two enhanced attacks, denoted as Enh1 and Enh2(b1 and b2 are two different original samples)

5.1 IMPERCEPTIBILITY ANALYSIS

We construct 100 full-sentence targeted adversarial examples sampled from the LibriSpeech test-clean dataset for our PhaseFool attack. For comparison, we then generate adversarial examples using the prior state-of-the-art works as baselines, including: 1) *C&W's attack (Carlini & Wagner, 2018)*, a two-stage adversarial attack for audio based on CTC loss and l_2-norm; 2) *IPC attack Zhang et al. (2020)*, a new Iterative Proportional Clipping (IPC) algorithm to generate robust audio adversarial examples with temporal dependency; 3) *Qin's attack (Qin et al., 2019)*, the imperceptible attack that leverages the psychoacoustic principle of auditory masking. We ensure that all attacks has a 100% successful rate and the target transcription is consistent among the experiments to exclude other interference factors.

The user study is conducted to verify the imperceptibility of these attacks. We recruit 24 users online and give them one of the three types of questions, including: 1) clean or noisy, we give users a set of audio samples and ask them to tell us if they have any background noise (e.g., static, echoing, people talking in the background). The users are asked to listen to each sample with headphones on, and answer questions about a set of audio samples (one user can only answer one type of the questions). In all experiments, users have the ability to listen to audio samples multiple times when they are unsure of the answer, making it as difficult as possible for our adversarial examples to pass as clean data.

We perform three experiments to validatethat our adversarial examples are imperceptible, especiallycompared to prior work

5.2 GENERATION SPEEDUP

We evaluate the generation speech of PhaseFool compared with several state-of-the-art audio adversarial attacks in the previous experiment. We generate 100 adversarial examples for each algorithm with 100% successful rate and ensure that the original audio and the target transcription is consistent among the experiments to exclude other interference factors. As shown in Table 2, it can be seen that PhaseFool can generate a successful full-sentence audio adversarial example in around 30 seconds on average, which is a 9.24x speedup compared with the imperceptible counterpart Qin's attack (Qin et al., 2019). PhaseFool is a one-stage attack without complex computations and only requires 500 steps to obtain a successful and imperceptible adversarial example, which demonstrates the effectiveness of phase-oriented audio adversarial attack.

Method	Steps	Efficiency(s)
C&W's Attack	4000	211.87 ± 12.31
IPC	3000	158.75 ± 10.99
Imperceptible	5000	305.91 ± 8.02
PhaseFool	500	33.10 ± 3.27
PhaseFool w/o constrain	212	17.80 ± 19.56

Table 2: Comparison of the generation speed with 95% confidence intervals. Efficiency(s) means that the algorithm can generate a successful full-sentence audio adversarial examples within the listed seconds on average.

Table 3: Hearing test

Method	**Q1**	**Q2**	**Q3**	**Mel.Err**
PhaseFool	38.25% / 42.25%	4.25%	74.00%	
PhaseFool w/o PWM	38.25% / 42.25%	4.25%	74.00%	
PhaseFool w/o constrain	38.25% / 42.25%	4.25%	74.00%	

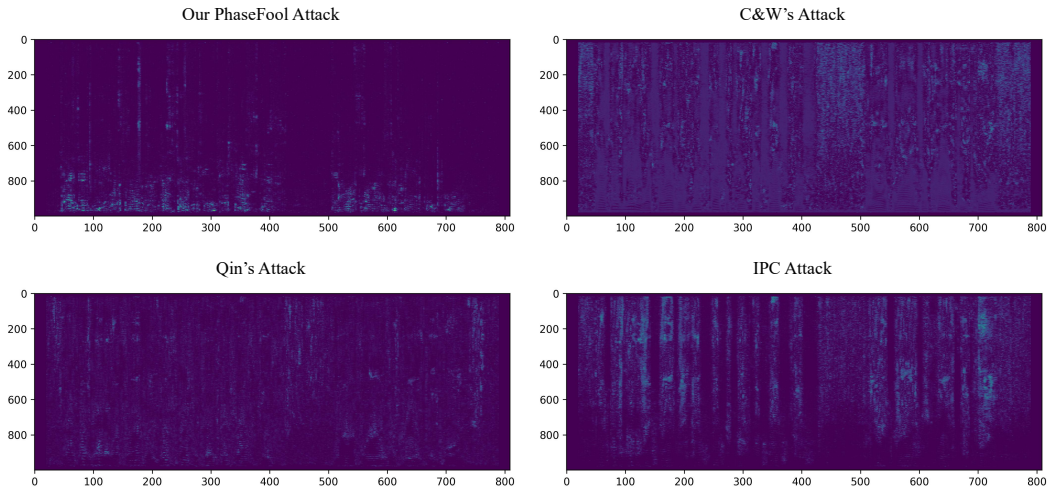


Figure 3: Pixel-wise difference in the mel-spectrogram domain between generated waveforms and a mel-spectrogram from Tacotron2. Before fine-tuning, HiFi-GAN generates waveforms corresponding to input conditions accurately. After fine-tuning, the error of the mel-spectrogram level increased, but the perceptual quality increased.

5.3 Interpretation of Attacks

For further investigating the interpretation of different attacks, we present the pixel-wise difference of an adversarial example in the mel-spectrogram domain to visualize the learned adversarial perturbations in Figure 3.

5.4 Ablation Study

We also conduct ablation studies to explore the influences of the weighted loss function proposed in subsection 3.3.

6 Conclusions

In this work, we proposed PhaseFool, a phase-oriented audio adversarial examples, to address the issues in FastSpeech and ease the one-to-many mapping problem: 1) we directly train the model with ground-truth mel-spectrograms to simplify the training pipeline and also avoid information loss compared with FastSpeech; and 2) we improve the duration accuracy and introduce more variance information including pitch and energy to ease the one-to-many mapping problem, and improve pitch prediction by introducing continuous wavelet transform. Moreover, based on FastSpeech 2, we further developed FastSpeech 2s, a non-autoregressive text-to-waveform generation model, which enjoys the benefit of fully end-to-end inference and achieves faster inference speed. Our experimental results show that FastSpeech 2 and 2s outperform FastSpeech, and FastSpeech 2 can even surpass autoregressive models in terms of voice quality, with much simpler training pipeline while inheriting the advantages of fast, robust and controllable speech synthesis of FastSpeech.

High quality, fast and fully end-to-end training without any external libraries is definitely the ultimate goal of neural TTS and also a very challenging problem. To ensure high quality of FastSpeech

2, we use an external high-performance alignment tool and pitch extraction tools, which may seem a little complicated, but are very helpful for high-quality and fast speech synthesis. We believe there will be more simpler solutions to achieve this goal in the future and we will certainly work on fully end-to-end TTS without external alignment models and tools. We will also consider more variance information (Zhang et al., 2021) to further improve the voice quality and speed up the inference with more light-weight model (Luo et al., 2021).

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Yang Ai and Zhen-Hua Ling. A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:839–851, 2020. doi: 10.1109/TASLP.2020.2970241. URL https://doi.org/10.1109/TASLP.2020.2970241.

Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

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*, pp. 387–402. Springer, 2013.

Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pp. 1–7. IEEE Computer Society, 2018. doi: 10.1109/SPW.2018.00009. URL https://doi.org/10.1109/SPW.2018.00009.

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)*, pp. 513–530, 2016.

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 Srdjan Capkun and Franziska Roesner (eds.), *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pp. 2667–2684. USENIX Association, 2020. URL https://www.usenix.org/conference/usenixsecurity20/presentation/chen-yuxuan.

Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex u-net. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkeRTsAcYm.

Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *CoRR*, abs/1707.05373, 2017. URL http://arxiv.org/abs/1707.05373.

Gilles Degottex and Daniel Erro. A measure of phase randomness for the harmonic model in speech synthesis. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *CoRR*, abs/1711.03280, 2017. URL http://arxiv.org/abs/1711.03280.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 2472–2476. ISCA, 2020. doi: 10.21437/Interspeech.2020-2537. URL https://doi.org/10.21437/Interspeech.2020-2537.

Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama. Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency. In *Proc. DAFx*, volume 10, pp. 397–403, 2010.

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

Xiaolei Liu, Kun Wan, Yufei Ding, Xiaosong Zhang, and Qingxin Zhu. Weighted-sampling audio adversarial example attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4908–4915, 2020.

Ryoji Miyahara and Akihiko Sugiyama. An auto-focusing noise suppressor for cellphone movies based on phase randomization and power compensation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2199–2203. IEEE, 2014.

Pejman Mowlaee and Josef Kulmer. Phase estimation in single-channel speech enhancement: Limits-potential. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8): 1283–1294, 2015.

Pejman Mowlaee, Rahim Saeidi, and Yannis Stylianou. Advances in phase-aware signal processing in speech communication. *Speech communication*, 81:1–29, 2016.

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)*, pp. 5206–5210. IEEE, 2015.

Yao Qin, Nicholas Carlini, Garrison W. Cottrell, Ian J. Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5231–5240. PMLR, 2019. URL http://proceedings.mlr.press/v97/qin19a.html.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. URL https://www.ndss-symposium.org/ndss-paper/adversarial-attacks-against-automatic-speech-recognition-systems-via-psychoacoustic-hiding/.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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.

Peter Vary and Member Eurasip. Noise suppression by spectral magnitude estimation—mechanism and theoretical limits—. *Signal processing*, 8(4):387–400, 1985.

Shengbei Wang, Weitao Yuan, Zhen Zhang, Jianming Wang, and Masashi Unoki. Synchronous multi-bit audio watermarking based on phase shifting. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pp. 2700–2704. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9414307. URL https://doi.org/10.1109/ICASSP39728.2021.9414307.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. Code-switched language models using neural based synthetic data from parallel sentences. In Mohit Bansal and Aline Villavicencio (eds.), *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pp. 271–280. Association for Computational Linguistics, 2019. doi: 10.18653/v1/K19-1026. URL https://doi.org/10.18653/v1/K19-1026.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 6144–6148. IEEE, 2020a. doi: 10.1109/ICASSP40776.2020.9053878. URL https://doi.org/10.1109/ICASSP40776.2020.9053878.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6144–6148. IEEE, 2020b.

Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. In Sarit Kraus (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 5334–5341. ijcai.org, 2019. doi: 10.24963/ijcai.2019/741. URL https://doi.org/10.24963/ijcai.2019/741.

Hongting Zhang, P Zhou, Q Yan, and Xiao-Yang Liu. Generating robust audio adversarial examples with temporal dependency. In *Proc. Int. Joint Conf. Artif. Intell.*, pp. 1–5, 2020.

Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. In B. Yegnanarayana (ed.), *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pp. 791–795. ISCA, 2018. doi: 10.21437/Interspeech.2018-1107. URL https://doi.org/10.21437/Interspeech.2018-1107.

A Appendix

You may include other additional sections here.