Supplementary Material for Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition

Yao Qin [1] **Nicholas Carlini** [2] **Ian Goodfellow** [2] **Garrison Cottrell** [1] **Colin Raffel** [2]
[1] **University of California, San Diego** [2] **Google Brain**

1. Frequency Masking Threshold

In this section, we detail how we compute the frequency masking threshold for constructing imperceptible adversarial examples. This procedure is based on psychoacoustic principles which were refined over many years of human studies. For further background on psychoacoustic models, we refer the interested reader to (Lin & Abdulla, 2015; Mitchell, 2004).

Step 1: Identifications of Maskers

In order to compute the frequency masking threshold of an input signal $x(n)$, where $0 \leq n \leq N$, we need to first identify the maskers. There are two different classes of maskers: tonal and nontonal maskers, where nontonal maskers have stronger masking effects compared to tonal maskers. Here we simply treat all the maskers as tonal ones to make sure the threshold that we compute can always mask out the noise. The normalized PSD estimate of the tonal maskers $\bar{p}_x^m(k)$ must meet three criteria. First, they must be local maxima in the spectrum, satisfying:

$$\bar{p}_x(k-1) \leq \bar{p}_x^m(k) \quad \text{and} \quad \bar{p}_x^m(k) \geq \bar{p}_x(k+1), \tag{1}$$

where $0 \leq k < \frac{N}{2}$.

Second, the normalized PSD estimate of any masker must be higher than the threshold in quiet ATH(k), which is:

$$\bar{p}_x^m(k) \geq \text{ATH}(k), \tag{2}$$

where ATH(k) is approximated by the following frequency-dependency function:

$$\begin{aligned}\text{ATH}(f) = 3.64(\frac{f}{1000})^{-0.8} - 6.5\exp\{-0.6(\frac{f}{1000} - 3.3)^2\} \\ + 10^{-3}(\frac{f}{1000})^4.\end{aligned} \tag{3}$$

The quiet threshold only applies to the human hearing range of 20Hz $\leq f \leq$ 20kHz. When we perform short time Fourier transform (STFT) to a signal, the relation between the frequency f and the index of sampling points k is

$$f = \frac{k}{N} \cdot f_s, \quad 0 \leq f < \frac{f_s}{2} \tag{4}$$

where f_s is the sampling frequency and N is the window size.

Last, the maskers must have the highest PSD within the range of 0.5 Bark around the masker's frequency, where bark is a psychoacoustically-motivated frequency scale. Human's main hearing range between 20Hz and 16kHz is divided into 24 non-overlapping critical bands, whose unit is Bark, varying as a function of frequency f as follows:

$$b(f) = 13\arctan(\frac{0.76f}{1000}) + 3.5\arctan(\frac{f}{7500})^2. \tag{5}$$

As the effect of masking is additive in the logarithmic domain, the PSD estimate of the the masker is further smoothed with its neighbors by:

$$\bar{p}_x^m(\bar{k}) = 10\log_{10}[10^{\frac{\bar{p}_x(k-1)}{10}} + 10^{\frac{\bar{p}_x^m(k)}{10}} + 10^{\frac{\bar{p}_x(k+1)}{10}}] \tag{6}$$

Step 2: Individual masking thresholds

An individual masking threshold is better computed with frequency denoted at the Bark scale because the spreading functions of the masker would be similar at different Barks. We use $b(i)$ to represent the bark scale of the frequency index i. There are a number of spreading functions introduced to imitate the characteristics of maskers and here we choose the simple two-slope spread function:

$$\text{SF}[b(i), b(j)] = \begin{cases} 27\Delta b_{ij}, & \text{if } \Delta b_{ij} \leq 0. \\ G(b(i)) \cdot \Delta b_{ij}, & \text{otherwise} \end{cases} \tag{7}$$

where

$$\Delta b_{ij} = b(j) - b(i), \tag{8}$$

$$G(b(i)) = [-27 + 0.37\max\{\bar{p}_x^m(b(i)) - 40, 0\}] \tag{9}$$

where $b(i)$ and $b(j)$ are the bark scale of the masker at the frequency index i and the maskee at frequency index j respectively. Then, $T[b(i), b(j)]$ refers to the masker at Bark index $b(i)$ contributing to the masking effect on the maskee at bark index $b(j)$. Empirically, the threshold $T[b(i), b(j)]$ is calculated by:

$$T[b(i), b(j)] = \bar{p}_x^m(b(i)) + \Delta_m[b(i)] + \text{SF}[b(i), b(j)], \tag{10}$$

where $\Delta_m[b(i)] = -6.025 - 0.275b(i)$ and SF$[b(i), b(j)]$ is the spreading function.

Step 3: Global masking threshold

The global masking threshold is a combination of individual masking thresholds as well as the threshold in quiet via addition. The global masking threshold at frequency index i measured with Decibels (dB) is calculated according to:

$$\theta_x(i) = 10\log_{10}[10^{\frac{ATH(i)}{10}} + \sum_{j=1}^{N_m} 10^{\frac{T[b(j),b[i]]}{10}}], \quad (11)$$

where N_m is the number of all the selected maskers. The computed θ_x is used as the frequency masking threshold for the input audio x to construct imperceptible adversarial examples.

2. Stability in Optimization

In case of the instability problem during back-propagation due to the existence of the log function in the threshold $\theta_x(k)$ and the normalized PSD estimate of the perturbation $\bar{p}_\delta(k)$, we remove the term $10\log_{10}$ in the PSD estimate of $p_\delta(k)$ and $p_x(k)$ and then they become:

$$p_\delta(k) = \left|\frac{1}{N}s_\delta(k)\right|^2, \quad p_x(k) = \left|\frac{1}{N}s_x(k)\right|^2 \quad (12)$$

and the normalized PSD of the perturbation turns into

$$\bar{p}_\delta(k) = \frac{10^{9.6}p_\delta(k)}{\max_k\{p_x(k)\}}. \quad (13)$$

Correspondingly, the threshold $\theta_x(k)$ becomes:

$$\theta_x(k) = 10^{\frac{\theta_x}{10}} \quad (14)$$

3. Notations and Definitions

The notations and definitions used in our proposed algorithms are listed in Table 1.

4. Implementation Details

The adversarial examples generated in our paper are all optimized via Adam optimizer (Kingma & Ba, 2014). The hyperparameters used in each section are displayed below.

4.1. Imperceptible Adversarial Examples

In order to construct imperceptible adversarial examples, we divide the optimization into two stages. In the first stage, the learning rate lr_1 is set to be 100 and the number of iterations T_1 is 1000 as (Carlini & Wagner, 2018). The max-norm bound ϵ starts from 2000 and will be gradually reduced during optimization. In the second stage, the number of iterations T_2 is 4000. The learning rate lr_2 starts from 1 and will be reduced to be 0.1 after 3000 iterations. The adaptive parameter α which balances the importance between ℓ_{net} and ℓ_θ begins with 0.05 and gradually updated based on the performance of adversarial examples. Algorithm 1 shows the details of the two-stage optimization.

Algorithm 1 Optimization with Masking Threshold

Input: audio waveform x, target phrase y, ASR system $f(\cdot)$, perturbation δ, loss function $\ell(x, \delta, y)$, hyperparameters ϵ and α, learning rate in the first stage lr_1 and second stage lr_2, number of iterations in the first stage T_1 and second stage T_2.

```
# Stage 1: minimize ||δ||
Initialize δ = 0, ε = 2000 and α = 0.
for i = 0 to T_1 − 1 do
    δ ← δ − lr_1 · sign(∇_δ ℓ(x, δ, y))
    Clip ||δ|| ≤ ε
    if i % 10 = 0 and f(x + δ) = y then
        if ε > max(||δ||) then
            ε ← max(||δ||)
        end if
        ε ← 0.8 · ε
    end if
end for

# Stage 2: minimize the perceptibility
Reassign α = 0.05
for i = 0 to T_2 − 1 do
    δ ← δ − lr_2 · ∇_δ ℓ(x, δ, y)
    if i % 20 = 0 and f(x + δ) = y then
        α ← 1.2 · α
    end if
    if i % 50 = 0 and f(x + δ) ≠ y then
        α ← 0.8 · α
    end if
end for
```

Output: adversarial example $x' = x + \delta$

4.2. Robust Adversarial Examples

To develop the robust adversarial examples that could work after played over-the-air, we also optimize the adversarial perturbation in two stages. The first stage intends to find a relative small perturbation while the second stage focuses on making the constructed adversarial example more robust to random room configurations. The learning rate lr_1 in the first stage is 50 and δ will be updated for 2000 iterations. The max-norm bound ϵ for the adversarial perturbation δ starts from 2000 as well and will be gradually reduced. In the second stage, the number of iterations is set to be 4000

x	The clean audio input
δ	The adversarial perturbation added to clean audio
x'	The constructed adversarial example
y	The targeted transcription
$f(\cdot)$	The attacked neural network (ASR)
$\mathscr{F}(\cdot)$	Fourier transform
k	The index of the spectrum
N	The window size in short term Fourier transform
$s_x(k)$	The k-th bin of the spectrum for audio x
$s_\delta(k)$	The k-th bin of the spectrum for perturbation δ
$p_x(k)$	The log-magnitude power spectral density (PSD) for audio x at index k
$\bar{p}_x(k)$	The normalized PSD estimated for audio x at index k
$p_\delta(k)$	The log-magnitude power spectral density (PSD) for audio δ at index k
$\bar{p}_\delta(k)$	The normalized PSD estimated for audio δ at index k
$\theta_x(k)$	The frequency masking threshold for audio x at index k
$\ell(x, \delta, y)$	Loss function to optimize to construct adversarial examples
$\ell_{net}(\cdot, y)$	Loss function to fool the neural network with the input $(\cdot)$ and output y
$\ell_\theta(x, \delta)$	Imperceptibility loss function
α	A hyperparameter to balance the importance of ℓ_{net} and ℓ_θ
$\|\cdot\|$	Max-norm
ϵ	Max-norm bound of perturbation δ
$\nabla_\delta(\cdot)$	The gradient of $(\cdot)$ with regard to δ
lr_1, lr_2, lr_3	The learning rate in gradient descent
r	Room reverberation
$t(\cdot)$	The room transformation related to the room configuration
$\mathcal{T}$	The distribtion from which the transformation $t(\cdot)$ is sampled from
δ^*_{im}	The optimized δ in the first stage in constructing imperceptible adversarial examples
ϵ^*_r	The optimized ϵ in the first stage in constructing robust adversarial examples
δ^*_r	The optimized δ in the first stage in constructing robust adversarial examples
ϵ^{**}_r	The max-norm bound for δ used in the second stage in constructing robust adversarial examples
δ^{**}_r	The optimized δ in the second stage in constructing robust adversarial examples
Δ	The difference between $\epsilon^{**}_r - \epsilon^*_r$
Ω	A set of transformations sampled from distribution $\mathcal{T}$
M	The size of the transformation set Ω

Table 1. Notations and Definitions used in our algorithms.

Original phrase 1	the more she is engaged in her proper duties the less leisure will she have for it even as an accomplishment and a recreation
Targeted phrase 1	old will is a fine fellow but poor and helpless since missus rogers had her accident
Original phrase 2	a little cracked that in the popular phrase was my impression of the stranger who now made his appearance in the supper room
Targeted phrase 2	her regard shifted to the green stalks and leaves again and she started to move away

Table 2. Examples of the original and targeted phrases on the LibriSpeech dataset.

and the learning rate lr_2 is 5. In this stage, ϵ is fixed and equals the optimized ϵ_r^* in the first stage plus Δ. The size of transformation set Ω is set to be $M = 10$.

4.3. Imperceptible and Robust Attacks

To construct imperceptible as well as robust adversarial examples, we begin with the robust adversarial examples generated in Section. 4.2. In the first stage, we focus on reducing the imperceptibility by setting the initial α to be 0.01 and the learning rate is set to be 1. We update the adversarial perturbation δ for 4000 iterations. If the adversarial example successfully attacks the ASR system in 4 out of 10 randomly chosen rooms, then α with be increased by 2. Otherwise, for every 50 iterations, α will be decreased by 0.5.

In the second stage, we focus on improving the less perceptible adversarial examples to be more robust. The learning rate is 1.5 and α starts from a very small value 0.00005. The perturbation will be further updated for 6000 iterations. If the adversarial example successfully attacks the ASR system in 8 out of 10 randomly chosen rooms, then α will be increased by 1.2.

5. Transcription Examples

Some examples of the original phrases and targeted transcriptions from the LibriSpeech dataset (Panayotov et al., 2015) are shown in Table 2.

References

Carlini, N. and Wagner, D. A. Audio adversarial examples: Targeted attacks on speech-to-text. *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Mitchell, J. L. Introduction to digital audio coding and standards. *Journal of Electronic Imaging*, 13(2):399, 2004.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210. IEEE, 2015.