Our Goals

- **Targeted**
  Given an input audio $x$, a targeted transcription $y$, an automatic speech recognition system $f(\cdot)$, our target is to find a perturbation $\delta$, that $f(x + \delta) = y$ and $f(x) \neq y$.

- **Imperceptible**
  Humans cannot differentiate $x$ and $x + \delta$ when listening to these examples.

- **Robust**
  Played by a speaker and recorded by a microphone (over-the-air).
  (We don’t achieve this goal completely, but succeed at simulated rooms.)
Our Settings

- Threat Model
  White-box Attack

- ASR Model
  Lingvo ASR system (state-of-the-art) [1]

Imperceptibility

- Frequency Masking

A louder signal (the “masker”) can make other signals at nearby frequencies (the “maskees”) imperceptible.
Imperceptibility

- **Loss function** $\ell(x, \delta, y) = \ell_{net}(f(x + \delta), y) + \alpha \cdot \ell_\theta(x, \delta)$
  - $\ell_{net}(f(x + \delta), y)$ is the cross-entropy loss function;
  - $\ell_\theta(x, \delta) = \max\{\bar{p}_\delta(k) - \theta_x(k), 0\}$ is the imperceptibility loss

Where $\delta$ is the perturbation, $\bar{p}_\delta(k)$ is the psd of $\delta$ and $\theta_x(k)$ is the masking threshold.
Robustness

- Room Simulator
  - Simulate room impulse $r$ based on room configurations
  - Convolve speech with reverberation $t(x) = x * r, \ t \sim T$

- Robustness Loss Function
  - Minimize $\ell(x, \delta, y) = E_{t \sim T} [\ell_{net}(f(t(x + \delta)), y)]$ such that $|\delta| < \epsilon$
Imperceptible and Robust Attacks

- **Combination Loss Function (imperceptibility & robustness)**
  
  Minimize $\ell(x, \delta, y) = E_{t \sim T} [\ell_{\text{net}}(f(t(x + \delta)), y)] + \alpha \cdot \ell_{\theta}(x, \delta)$

  - Robustness loss
  - Imperceptibility loss
Conclusions

- Construct *effectively imperceptible* adversarial examples using frequency masking.
- Develop robust adversarial examples that remain effective after playing over-the-air in the simulated rooms.
- Generate adversarial examples for non-$\ell_p$-based metrics.
Thanks!
Come to our poster #65!

Project Webpage:
http://cseweb.ucsd.edu/~yaq007/imperceptible-robust-adv.html

Code:
https://github.com/tensorflow/cleverhans/tree/master/examples/adversarial_asr