Audio Adversarial Examples: Targeted Attacks on Speech-to-Text

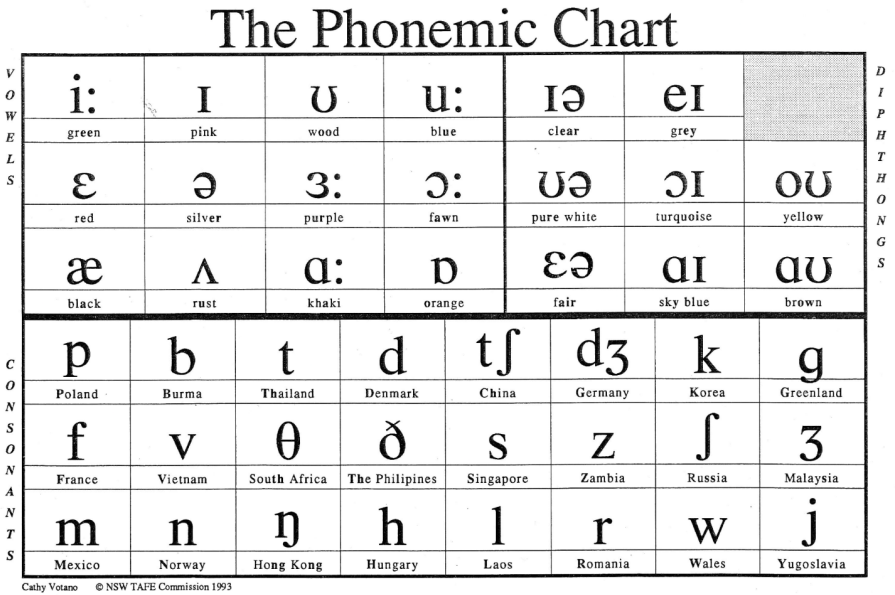
Nicholas Carlini, David Wagner University of California, Berkeley

Related work

- N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden voice commands. In 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, 2016.
- G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu. Dolphin attack: Inaudible voice commands. CCS, 2017.

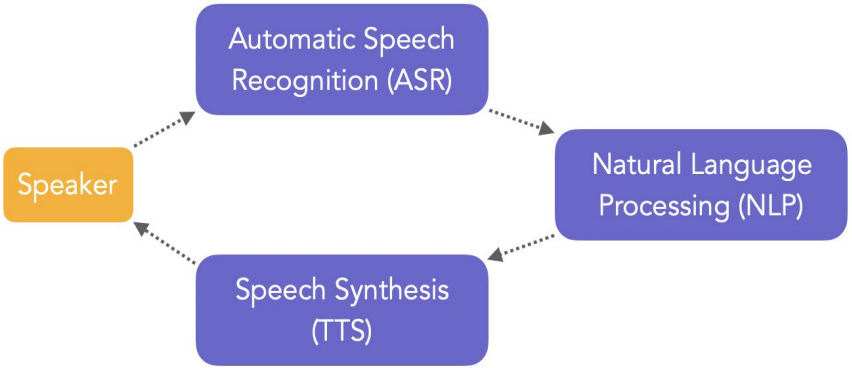
What are ASRs?

- Representation of words based on phonemes.
- Challenges of accents, room noise, reverberations etc.



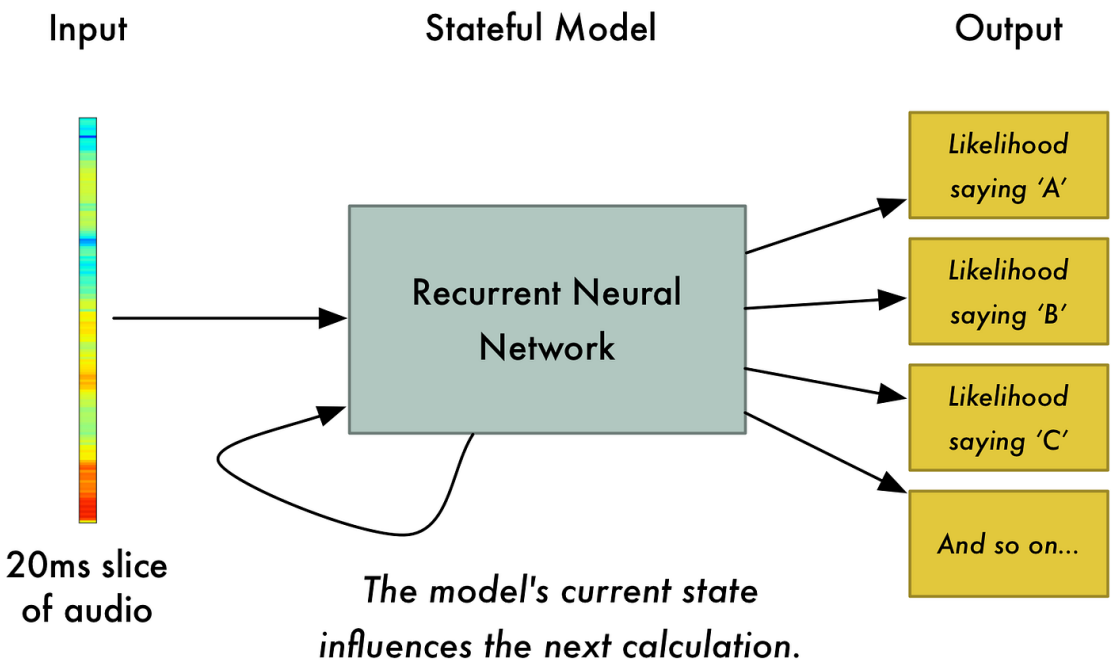
Multiple components

- Transformation from speech to phonemes.
- Probability distribution over the most probable words constructed from these phonemes.
- Probability distribution over most probable words to make a sentence.

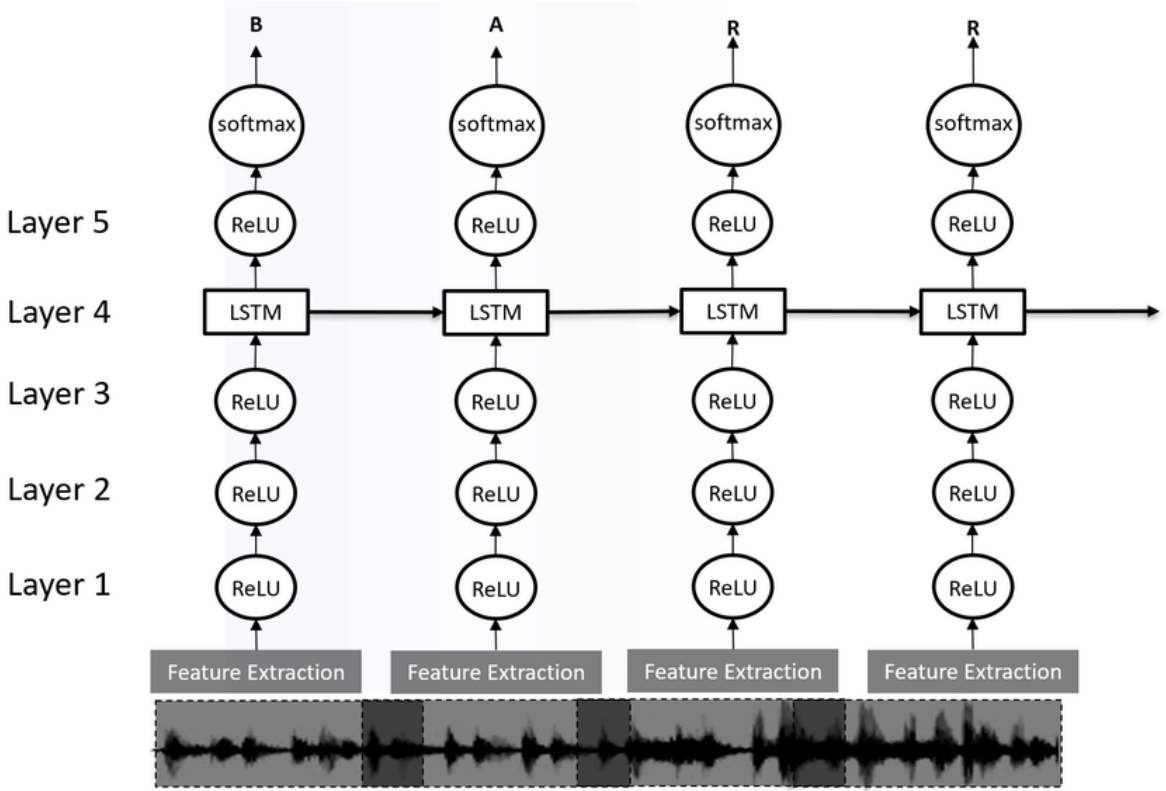


Speech-to-text

- Used in voice assistants, smart home devices…
- Architecture: CNN, RNN / LSTM, FC layers.



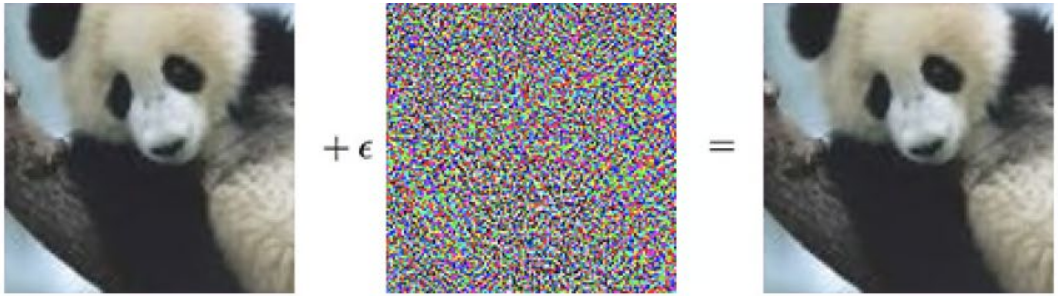
Mozilla DeepSpeech



Why should you care?

- Hiding information in audio files.
- Exploiting STT algorithms in home assistant devices.
- …

Adversarial examples for images

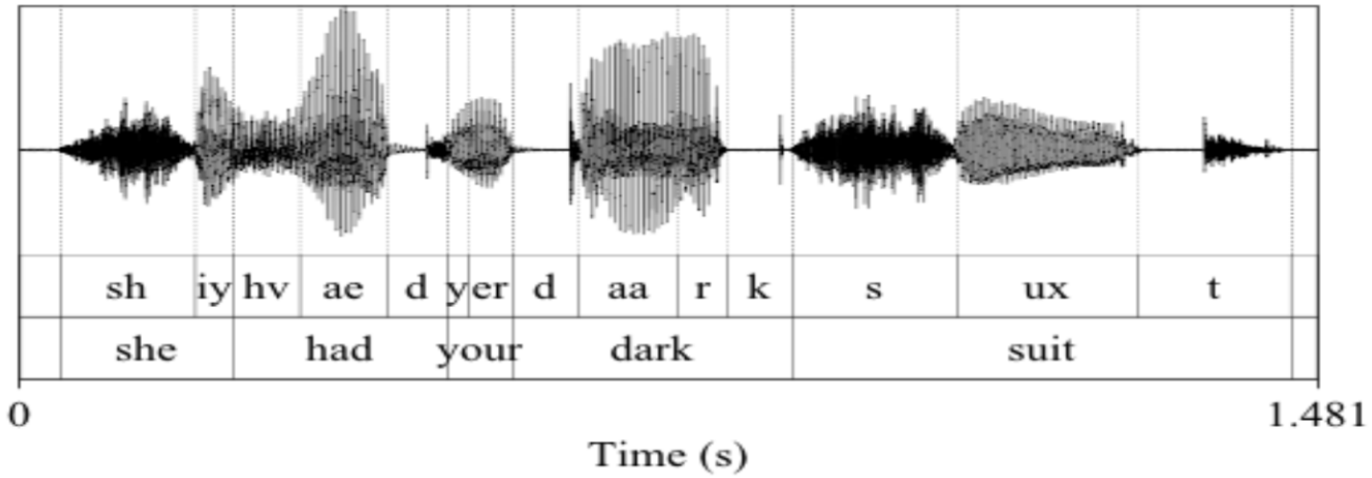


Source: https://openai.com/blog/adversarial-example-research/

Key findings: Everything that holds true in images about adversarial examples holds true for sound.

Challenges with audio training data

- Variable length (each sample is of different length).
- No alignment -> we don’t know which part of the audio corresponds to which phoneme.



Solution: Connectionist Temporal Classification

Connectionist Temporal Classification: how it works

We want to map X=[x1,x2,...,xT] to Y=[y1,y2,...,yT]

=> CTC gives us an output distribution for every possible Y given X.

=> **P(Y|X)**

Loss function

Function representing the difference between the output of the model and what the actual output should be.

Optimization is done through gradient descent (like always).

=> Let’s minimize it.

Constructing the adversarial example

- Definition of a distance metric d(x, x’) with x being the original input and x’ the adversarial example.
- Let f(x’) = T be the output of the model with T the output that the attacker has chosen.

The goal is to minimize the distance metric whilst achieving the output T.

CTC loss function

$$\text{CTC-Loss}(f(x), p) = -\log \Pr(p|f(x))$$ <- **Negative log likelihood**

Constructing the adversarial example

Distance metric used: Magnitude of perturbation in dB.

=> How much perturbation was introduced relative to the original signal.

=> Final formula that we want to minimize: **d(x, x') + g(x')** with g(x') being the CTC-Loss.

Examples

Original

That day the merchant gave the boy permission to build the display.

Now I would drift gently off to dreamland.

Adversarial example

Everyone seemed very excited.

My wife pointed out to me the brightness of the red green and yellow signal light.

Examples

Original

Adversarial example

Speech can be embedded in music.

Results

- 100% success rate with a mean perturbation of -31dB.
- Longer the target phrase is, the harder it is to generate a perturbation.
- Longer the input phrase is, the easier it is.
- Harder to generate new text than silence existing text.

Limitations

- Whitebox attack
- Over-the-air attacks do not work.

Future work

- Can these attacks be made playable over the air?
- Does the transferability property hold (like for images)?
- Do universal perturbations exist?
- Can any of the existing defenses be applied to audio?

Future work

- Can these attacks be made playable over the air? **Yes**
- Does the transferability property hold (like for images)?
- Do universal perturbations exist?
- Can any of the existing defenses be applied to audio?

Future work

- Can these attacks be made playable over the air? **Yes**
- Does the transferability property hold (like for images)? **Maybe**
- Do universal perturbations exist?
- Can any of the existing defenses be applied to audio?

Future work

- Can these attacks be made playable over the air? **Yes**
- Does the transferability property hold (like for images)? **Maybe**
- Do universal perturbations exist? **Yes**
- Can any of the existing defenses be applied to audio?

Future work

- Can these attacks be made playable over the air? **Yes**
- Does the transferability property hold (like for images)? **Maybe**
- Do universal perturbations exist? **Yes**
- Can any of the existing defenses be applied to audio? **Probably, further research needed**

Future work

- Can these attacks be made playable over the air? **Yes**
- Does the transferability property hold (like for images)? **Maybe**
- Do universal perturbations exist? **Yes**
- Can any of the existing defenses be applied to audio? **Yes, downsampling and upsampling**

Chen, Yuxuan, et al. "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices." (USENIX 2020)

Conclusion

- Audio adversarial examples are very similar to image adversarial examples.
- This type of attack is very easy to conduct assuming a whitebox environment.
- A lot of work remains for audio adversarial examples to be as effective as their image counterparts.

Great questions

- In the paper, the proposed attack was successfully carried out in a white-box setting, and the authors mentioned that it could also be possible in a black-box environment. If conducted in a black-box setting, what would be the most challenging aspect of implementing this attack?
- Similar to the MSF-ADV paper, we are again starting with a white-box model. Can this attack be carried out on a black-box model? Would it be effective enough? Would it be worthwhile, especially considering the fact that generating adversarial examples took quite a lot of time?
- This paper must have been done on English speech-to-text system, but would the results be different on other languages, depending on pronunciation for example?

Great questions

- Recently, crimes involving AI-generated voices, such as voice phishing or spreading false information, are on the rise. Given the findings in this paper—that certain words can be made indiscernible to humans but detectable by AI—could this approach be used to track such crimes or determine whether a voice was generated by AI (such as watermark)?
- Given that the attack relies on a white-box setting with full access to the model, how feasible would it be to extend this attack to a black-box scenario where only query access is available?

Best questions

- **Taeha Kim:** Compared to image-based examples, audio adversarial examples exhibit different properties, such as a lack of linearity in the audio domain. How might these differences affect the effectiveness of defense mechanisms originally developed for images?
- **YoungHyo Kang:** How could attacks that cause speech recognition systems to return incorrect results be used in real-world attack scenarios? Even if it's not effective over-the-air, could this be an attractive attack surface for attackers' perspective?
- **Sihun Yang:** Could the proposed attacks be generalized to other types of audio recognition systems beyond speech-to-text, such as speaker identification or keyword spotting systems?