Audio Adversarial Examples

Targeted Attacks on IBM Cloud Speech-to-Text API [1]

Xinyi Liu
Poornima Haridas
Priyank Pathak

Speech-To-Text Models

- Speech recognition is an interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers.

- Test Cloud APIs (Black-Box) - Google Cloud Speech-To-Text [9] & IBM Cloud Speech-To-Text [10]

Adversarial Attacks - Types

- White Box Attacks - The specifics of the Speech-To-Text model is known to the attacker and the attacker uses the same to malign the output.

- Black Box Attacks - The Speech-To-Text model is used as a black-box by the attacker. The only input available to the attacker is outputs from samples passed to the by the attacker to the model (active research topic)

- Targeted vs Untargeted Attacks

- Robust Attack - Attacks are transmitted across mediums (screenshot, microphone recordings etc.)

Audio Adversarial Attacks

- Relatively lesser explored field.
- Very difficult to generate convincing attacks.
- Best Case Scenario
  - Targeted Audio Attacks - Target transcript is generated by Speech-To-Text model instead of the original transcript.
  - Robust Attacks - The attack is robust across different medium.

Audio Adversarial Attacks - Evaluation

- ROUGE - 2(Recall-Oriented Understudy for Gisting Evaluation) Scores
  - Recall = $$\frac{number\_of\_overlapping\_2\_gram\_words}{total\_2\_grams\_words\_in\_reference\_summary}$$
  - Precision = $$\frac{number\_of\_overlapping\_2\_gram\_words}{total\_2\_grams\_words\_in\_system\_summary}$$
  - F1 score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
    - F1 = $$\frac{2*(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

Audio Adversarial Attacks - Evaluation

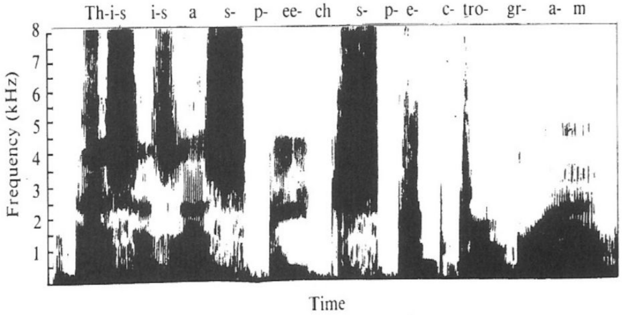
- The Bilingual Evaluation Understudy Score - or BLEU Scores

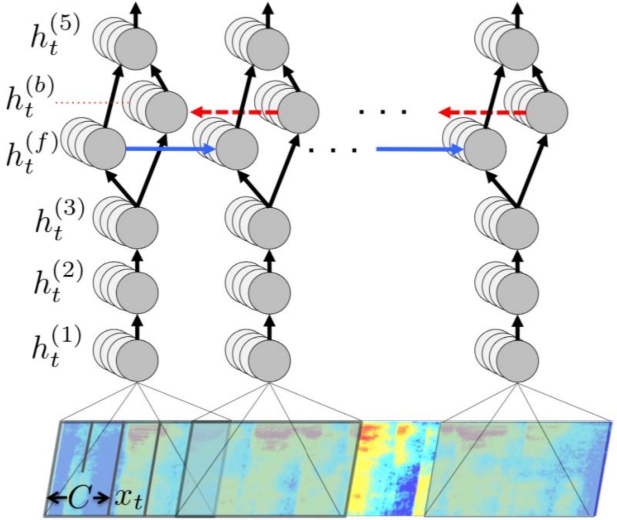
- Evaluation metric for given two sentences. BLEU's evaluation system requires two inputs: -
  - a numerical translation closeness metric, which is then assigned and measured against
  - a corpus of human reference translations

- We use the BLEU metric calculation provided by NLTK

Deep Speech Model - Scaling up end-to-end speech recognition [11]

- Recurrent neural network (RNN) trained to ingest speech spectrograms and generate English text transcriptions



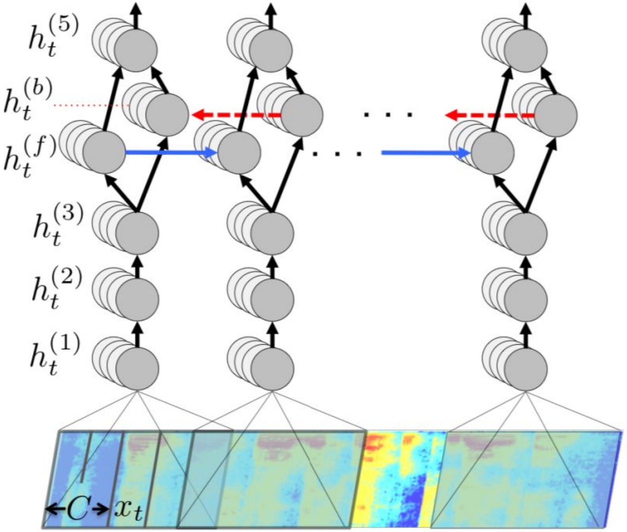


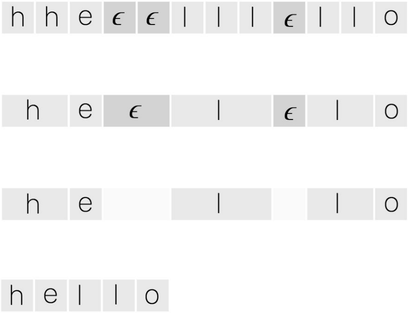
- IP - *x(i)* , a time-series of length *T(i)* where every time-slice is a vector of audio features, *x (i) t* , *t = 1, . . . , T(i)* + *C* , a set of context frames.

Image Source -[12], [11]

Deep Speech Model - Scaling up end-to-end speech recognition [11]

- OP of 1st 3 layers - $h(l)^t = g(W^{(l)} h_t^{(l-1)} + b^{(l)})$
- 4th Layer - Bidirectional RNN
- 5th - Takes $h_t^{(f)}$ and $h_t^{(b)}$ units as inputs and applies softmax




h h e ϵ ϵ l l l ϵ l l o

h e ϵ l ϵ l o

h e l l o

h e l l o

First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.

Audio Adversarial Attacks (Basic)

- Long audio
  - DEMO: 🔊
  - Original transcription: *“The accepted definition of creativity is production of something original and useful, and it is commonly thought that creativity occurs on the right side of the brain, and the arts play an important role in enhancing it. But according to a new research, creativity isn't about freedom from concrete facts.”*
  - IBM transcription: *“the accepted definition of creativity is production of something original and useful and it is commonly thought that creativity occurs on the right side of the brain and the arts play an important role in enhancing at but according to new research creativity isn't about freedom from concrete facts”*, **F1 = 0.928**
  - Google transcription: *“except the definition of creativity is production of something original and useful and it is commonly occurs on the play an important role in enhancing according to new research creativity isn't about Freedom From Concrete facts”*, **F1 = 0.690**

Audio Adversarial Attacks (Basic)

- Volume
  - DEMO: (original) 🔊 & (50 db lower) 🔊 & (50 db louder) 🔊
  - Original transcription: *"AND HENCE WE FIND THE SAME SORT OF CLUMSINESS IN THE TIMAEUS OF PLATO WHICH CHARACTERIZES THE PHILOSOPHICAL POEM OF LUCRETIUS"*
  - IBM transcription of 50 db lower: *"we find the same sort of clumsiness the to me is way too which characterizes the philosophical form of recreation"*, **F1 = 0.462** (Original: F1 = 0.878)
  - IBM transcription of 50 db louder: *"this this is a free service"*, **F1 = 0.0**
- Speed
  - DEMO: (0.5 times slower) 🔊 & (1.5 times faster) 🔊
  - IBM transcription of 0.5 times slower: *"and hens we find the same sort of clumsiness in the to me is the way to which characterizes the philosophical pool one of Lucretius"*, **F1 = 0.500**
  - IBM transcription of 1.5 times faster: *"and hence we find the same sort of clumsiness the today as of late %HESITATION which characterizes the philosophical poem of Lucretius"*, **F1 = 0.634**

Audio Adversarial Attacks (Basic)

- Pitch
  - DEMO: 🔊
  - IBM Transcription: *yes so it was yes right right those awful problem*, **F1 = 0.0** (Original: F1 = 0.878)

- Non-intelligent Noise
  - DEMO: (band limited noise) 🔊 & (music) 🔊
  - IBM transcription of band limited noise added: *“and hence we find the same sort of clumsiness in the two mayors of Plato which characterizes the philosophical poem of Lucretius”*, **F1 = 0.732**
  - IBM transcription of music added: *“and hence we find the same sort of clumsiness in the two main areas of Plato which characterizes the philosophical poem of the creation”*, **F1 = 0.698**

Audio Adversarial Attacks (Basic) - Evaluation

- Some basic attacks do succeed but are untargeted.
- Models are very likely to fail on words hard to pronounce when noise is added.
- Models may confuse when encountering words with similar pronunciation.

Attack Type	**F1 Score (IBM)**	**F1 Score (Google)**
Original	0.878	0.899
Volume - Lower	0.462	**NA**
Volume - Louder	**0**	**NA**
Speed - Slower	0.500	0.999
Speed - Faster	0.634	0.605
Pitch	**0**	**NA**
Noise - Band Limited	0.732	0.899
Noise - Music	0.698	0.899

Audio Adversarial Attacks (Original SOTA'18)

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text
(Nicholas Carlini & David Wagner, UCB**)**

- First of its kind, to work successfully on audio signals
- White box attack on *Deep Speech pretrained model*
- Fails on black box scenarios (*IBM API, Google Cloud API*)
    - Given original normalized signal (x)
    - Add "intelligent" noise to this signal x (**δ**)
    - Train this **δ** to force the model to predict the desired transcription
    - Only trainable parameter is the input noise (model weights freezed)
    - Training via back propagation
    - Limit the amount of noise to minimal via keeping a check on decibels of information added
    - Generalize the audio by introducing a random gaussian noise to the signal in addition to noise

Audio Adversarial Attacks (Original SOTA'18)

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text
(Nicholas Carlini & David Wagner, UCB**)**

- Limiting the perturbation : Decibel loss $dB_x(\delta) = dB(\delta) - dB(x)$

  where $dB(x) = \max_i \ 20 \cdot \log_{10}(x_i)$ i.e. minimising the relative decibel of noise with respect to original audio

- Desired Phrase prediction ensured via CTC Loss

Training is a 2 step process :

1) get the Desired decibel range $dB_x(\delta) \leq \tau$, keeping initial tau very large and subsequently decreasing
2) Minimize the L-2 norm (L infinity norm ideal case) and CTC loss

Audio Adversarial Attacks (Original SOTA'18)

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text
(Nicholas Carlini & David Wagner, UCB)

- DEMO: (w/o attack) 🔊 & (w/ attack) 🔊 & (microphone recording) 🔊
- Original transcription : *"AND HENCE WE FIND THE SAME SORT OF CLUMSINESS IN THE TIMAEUS OF PLATO WHICH CHARACTERIZES THE PHILOSOPHICAL POEM OF LUCRETIUS"*
- Target Phrase : *"this is cml class"*
  - IBM : (F1: **0.87804 vs 0.6829 vs 0.5116)**

    -> *"and hence we find the same sort of clumsiness in the to me as Plato which characterizes the philosophical form of recreation"*

    -> *"and hence we find the same sort of clumsiness the to me as a way to %HESITATION which characterizes the philosophical form of recreation"*

  - GOOGLE : (F1: **0.899 vs 0.4210 vs 0.799)**

    -> *"and how to recline the same sort of clumsiness in the team is Plato's which characterizes the clothes off"*

    -> *"and hence we find the same sort of clumsiness in the tomatoes of Plato's which characterizes the philosophical poem of lucretius"*

Audio Adversarial Attacks (Modified SOTA’18)

(Code publicly available on GitHub [5])

- Use L2 normalization in addition to whitening (0 mean and 1 standard deviation normalization)
- Use a scale factor of 5, to make the audio audible back again.
- 3 loss functions : L2 norm, CTC loss and DBx(**δ**) loss
- Adaptive Boosting for different loss functions

$$Loss = \begin{cases} 10 * CTC_{weight} * CTC_{loss} + 0.005 * (DBx(\delta) - \tau) + l2_{weight} * l2_{norm} & DBx(\delta) < \tau \\ CTC_{weight} * CTC_{loss} + 10 * 0.005 * (DBx(\delta) - \tau) + 10 * l2_{weight} * l2_{norm} & otherwise \end{cases}$$

It's easy to achieve desired phrase but difficult to quieten the noise

CTC-weight ↓ 0.1 if less than threshold (=2), l2-weight ↑ 0.2

Audio Adversarial Attacks (Modified SOTA’18)

- DEMO: 🔊 & (microphone recording) 🔊
- Target Phrase : *“this is cml class”*
  - IBM : **(**F1: **0.87804 vs 0.05263 vs 0.1250)**

    -> “Hey with fish finder things sort of close the two layers of clay to characterize it was awful for”

    -> “find things sort of close the clear way to characterize the philosophical issue”

  - GOOGLE : **(**F1: **0.899 vs vs 0.5555 vs 0.514285)**

    -> “required the same sort of clumsiness in the two layers of Plato which characterizes the clothes off.”

    -> “replying the same sort of clumsiness in the tomatoes with Plato which characterizes the clothes off.”

Audio Adversarial Attacks (Robust SOTA’19)

Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition [3] & [6]

(Yao Qin, Nicholas Carlini, **Ian Goodfellow**, Garrison Cottrell, Colin Raffel**)**

- First to introduce robustness for audio signals
- White box attack on 2019 SOTA Speech-Text model **lingvo (Google)**
- Continuation of 2018 paper (previous paper)
- Robustness:
  - Acoustic room simulator creating artificial utterances (speech with reverberations) that mimic playing the audio over-the-air
  - Room impulse response “**r**” based on the room configurations (the room dimension, source audio and target microphone location, and reverberation time)
  - Multiple room impulse responses “**r**”(s) to train the noise in robust room simulation

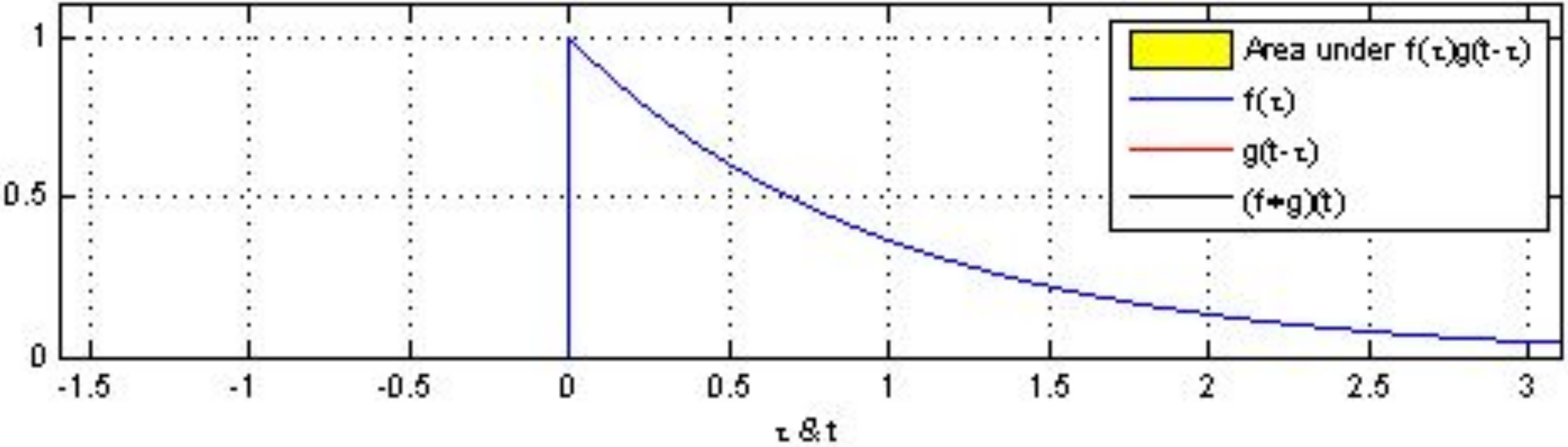
Audio Adversarial Attacks (Robust SOTA’19)

Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition [3] & [6]

(Yao Qin, Nicholas Carlini, **Ian Goodfellow**, Garrison Cottrell, Colin Raffel**)**

- Procedure :
    - Given original normalized speech signal (x)
    - Add “intelligent” noise to this signal (**δ**)
    - Perturb it in a simulated environment: r ✖ (x + **δ**)
    - Minimize the log-magnitude of power spectral density of **δ** (making it imperceptible, shadowing it with nearby original signal wavelets)
    - 2 step reduction :
        - Minimize infinity norm of noise in addition to CTC loss (lower audible noise to match target phrase)
        - Minimize normalized PSD loss and CTC loss (no bound on infinity norm)

Convolution



Audio Adversarial Attacks (Robust SOTA'19)

- DEMO: 🔊 & (microphone recording) 🔊
- Target Phrase : *"this is a cloud machine learning course"*
    - lingvo: 60% accuracy with simulated rooms

    - IBM : **(**F1: **0.87804 vs 0.0 vs 0.0)**

    "it's a crime scene the wait"

    "it's a class thing"

    - GOOGLE : **(**F1: **0.899 vs 0.0 vs None)**

    "Honda Kingsport"

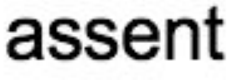
    N/A

Audio Adversarial Attacks - Discussion

- Tried to perform “blackbox” attacks on cloud APIs with limited queries.
- Map between patterns obtained from target set with DeepSpeech and transcription output by IBM.
    - A → till, with
    - B → **flames, female**, place, clean
    - C → (random words)
    - D → **fourth**, crew, please, later
    - E → premier, arrive, wait, **female**
    - F → four, **list, later**, please
- Robustness of attacks across different medium should be taken attention on.

Trials

- Targeted information may easily lose when transferring to new medium, same for image attacks like HopSkipJumpAttack [7] & [8]
    - Classifier: IBM Adversarial Robustness Toolbox, BlackBoxClassifier
    - Target: **assent**

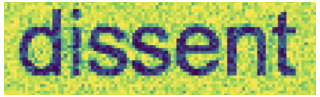
      

      assent

    - Input: **dissent**

      dissent

    - Output: **assent**, but **cissent** after screenshot

      dissent

    - Look into other robust image attacks

Audio Adversarial Attacks - Conclusion & Future Work

- Presented successful basic audio attacks on IBM and Google Cloud Speech-to-Text APIs
- Presented successful complex targeted black-box Audio Adversarial Attacks (our version of [1] & [3]) on IBM and Google Cloud Speech-to-Text APIs
- Developed and demonstrated robust version of [1]
- Started attacks on OCR and images

- Study the patterns observed in detail
- Study on how and why the model struggles with difficult/large words
- How to attack basic and most commonly occuring words (*a, the* etc)

References

1. Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.
2. GitHub - https://github.com/carlini/audio_adversarial_examples
3. Qin, Yao, et al. "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition." arXiv preprint arXiv:1903.10346 (2019).
4. Taori, Rohan, et al. "Targeted adversarial examples for black box audio systems." 2019 IEEE Security and Privacy Workshops (SPW). IEEE, 2019.
5. GitHub - https://github.com/ppriyank/Adveserial-Attacks/tree/master/paper2018
6. GitHub - https://github.com/tensorflow/cleverhans/tree/master/examples/adversarial_asr
7. Chen, Jordan, et al. "HopSkipJumpAttack: A Query-Efficient Decision-Based Attack" arXiv preprint arXiv:1904.02144 (2019).
8. OCR (GitHub) - https://github.com/IBM/adversarial-robustness-toolbox/blob/master/notebooks/classifier_blackbox_tesseract.ipynb
9. Google Cloud Speech-To-Text - https://cloud.google.com/speech-to-text/
10. IBM Cloud Speech-To-Text - https://www.ibm.com/cloud/watson-speech-to-text
11. Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567 (2014).
12. http://web.eecs.utk.edu/~leparker/Courses/CS594-fall13/Lectures/22-Slides-Speech-Recog-Dec-3.pdf
13. Guo, Chuan, et al. "Simple black-box adversarial attacks." arXiv preprint arXiv:1905.07121 (2019).

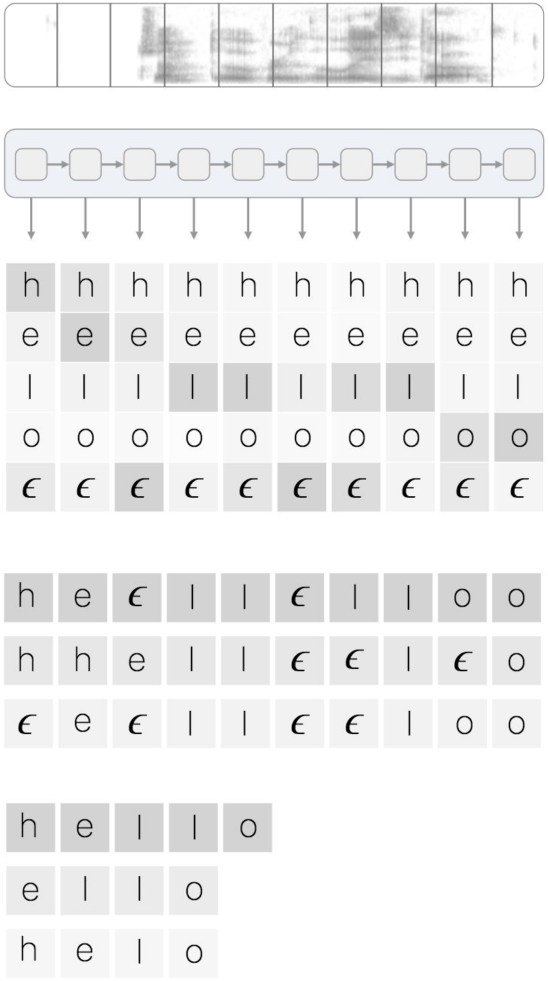
Audio Adversarial Attacks

Thank You!

Questions..

CTC Loss

1) Loss loss for correct alignment
2) Alignment probability is product of forward normalized probabilities and backward probabilities
3) Some more complications….



We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

The network gives $p_t(a \mid X)$, a distribution over the outputs {h, e, l, o, ϵ} for each input step.

With the per time-step output distribution, we compute the probability of different sequences

By marginalizing over alignments, we get a distribution over outputs.