# Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition

## Paper Review

## By Team Tri-Vachan

Summary:

The paper improves upon the current state-of-the-art in generating imperceptible audio adversarial examples by using the principle of frequency masking instead of $\ell_p$ distance measures to minimize the total energy which have been used in the past and only adding the adversarial perturbation to regions of the audio where it will not be heard by a human ear. The paper constructs both imperceptible and robust adversarial examples (verified by a human study) with the help of a cross entropy and a hinge loss function, while retaining 100% targeted success rate on arbitrary full-sentence targets. This paper also details initial steps towards developing audio which can be played over-the-air and still retain the adversarial property after being processed by random room environment simulators.

---

Reasons for score:

Overall, we vote for acceptance. The authors look at adversarial attacks to ASR systems from the new lens of frequency masking and make novel contributions towards generating imperceptible and robust adversarial examples. Furthermore, they take the first step towards developing adversarial audio examples that can be played in an over-the-air setting where they are played by a speaker and recorded by a microphone. We do have some concerns regarding clarity on certain areas which we have mentioned in limitations and comments sections.

---

Pros:

- The paper is easy to read and has been structured well.
- The paper makes meticulous attempt to justify each choice like giving theoretical reasons and convenience benefits to explain selection preferences, such as using random subsets at each gradient steps during lr2 phase and choosing $l_\theta(x, \delta)$ over $l_\theta(t(x), t(\delta))$ while framing loss function for imperceptible and robust attack training.
- Their experimental analysis is complete and rigorous.
- We especially found the human verification commendable. In human study, users have the ability to listen to the audio file multiple times and also have the added benefit of hearing 20 examples for “training” them.

---

Cons:

- It would have been appreciated if the authors could mention the average time taken to generate one adversarial sample and compared it with Wagner's method. Currently, there is no analysis on the time taken by the proposed method.
- During analysis, an actual experiment for robustness in a noisy environment could be reported without using a simulator. This would have been more practical and gave a sense of better measure of robustness. Since, the problem formulation and adversarial audio generation method allows the authors to test their robust environment hypothesis, this concern can be addressed in the rebuttal period.
- More details regarding the acoustic room simulator need to be provided. The distribution model for the acoustic room simulator and the method of sampling from it is unclear from the paper.
- In Experiment 2, the imperceptible adversarial examples are not preferred 50% of the time, which would appear in case of random sampling. There is still room for improvement.
- In Experiment 3, there is still some difference between the adversarial examples and the clean audio, which can be picked up by “trained” humans.
- In Table 2, the WER’s for clean and adversarial audio side by side may create confusion due to similar values, even though the ground truth considered is different in both cases.
- $\Delta$ = 300 and $\Delta$ = 400 have been used in the paper. Adding more datapoints would help generalize the readings better, rather than relying on just 2 datapoints.
- The results have been shown against the Lingvo system. However, there are other benchmarks that the adversarial system could have been compared against, for example - Kaldi system, etc.

---

Comments:

- In robustness analysis, the line “the WER is smaller than that of the clean audio” suggests a comparison between WERs but it does not make sense because target transcriptions are different in both cases, former being adversarial targeted transcription and latter the original ones. This could have been explained better in the paper. Overall, there are indeed few parts which create a bit of confusion which can be improved in the rebuttal period.
- Understandability could have been improved with a block diagram or a pictorial presentation of the idea.
- Attacks on different ASR systems can be tried in future, eg. where $l$ is CTC loss.
- Instead of crowdsourcing with the amazon turk, experiments with more controlled groups (to include diversity) and calibration of results could be done.

---

Concluding remarks:

The results are better than the ones used as a baseline and it looks like a great approach that can be expanded and improved in the future as the authors say. There are some minor possible improvements that have been mentioned above, but they don’t detract from the key message of the paper, and would only improve the quality of the paper.