MP3 Compression as a Means to Improve Robustness against Adversarial Noise Targeting Attention-based End-to-End Speech Recognition

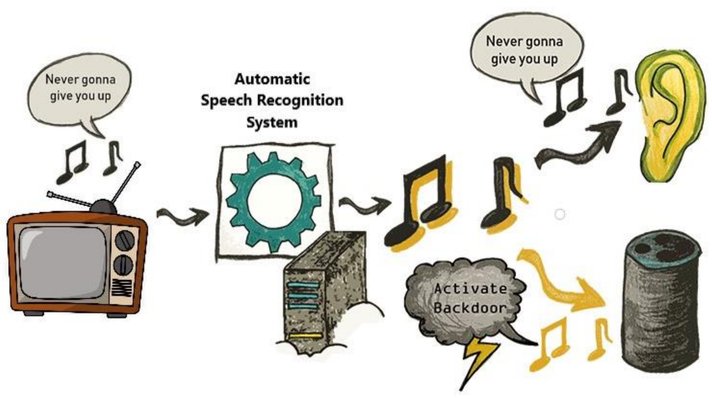
Iustina Andronic[1], Ludwig Kürzinger[2], Bernhard U. Seeber[3]

[1]Elite Master Program in Neuroengineering, Department of Electrical and Computer Engineering, Technical University of Munich
[2]Chair of Human-Machine Communication, Department of Electrical and Computer Engineering, Technical University of Munich
[3]Professorship of Audio Information Processing, Department of Electrical and Computer Engineering, Technical University of Munich

Abstract

Adversarial Examples represent an imminent security threat to any Machine Learning system. This thesis addresses this issue by proposing MP3 compression as a potential measure to defend Automatic Speech Recognition (ASR) systems against being mislead by Audio Adversarial Examples (AAEs). We generated untargeted AAEs in the form of adversarial noise added to original speech samples and used a feature inversion procedure to convert the adversarial examples from the feature into the audio domain. Different from prior work, we targeted an end-to-end, fully neural ASR system, namely ESPnet. We found that MP3 compression of AAEs indeed reduces the recognition errors when compared to raw, uncompressed adversarial inputs (result was validated by experiments with four ASR models trained on four types of audio data). Finally, a statistical test performed on the estimated Signal-to-Noise Ratio (SNR) of adversarial inputs confirmed that MP3-compressed adversarial samples had higher SNRs (hence less adversarial noise) than uncompressed adversarial inputs. This last finding consolidates the previous one in showing that MP3 compression is effective in diminishing only the adversarial noise.

1. Introduction

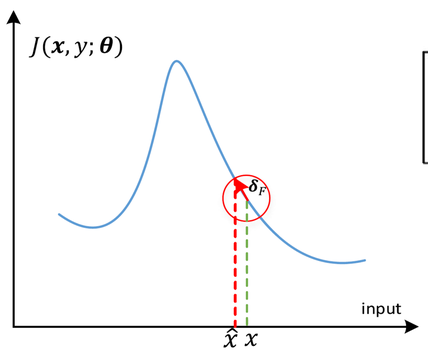


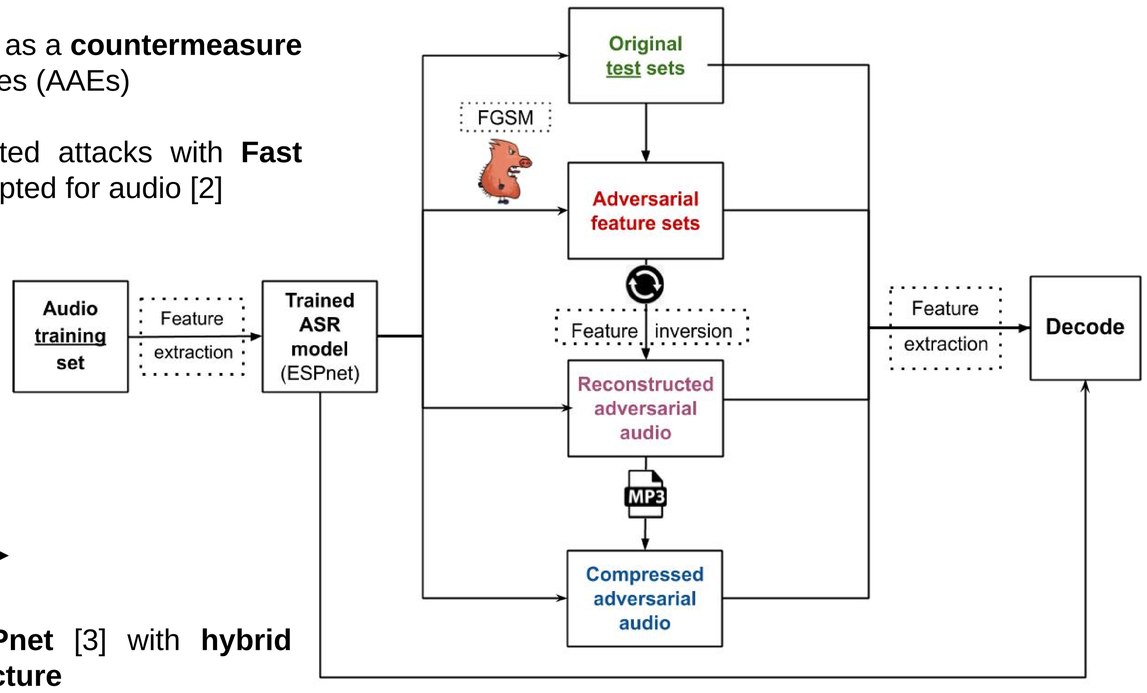
Adapted from https://adversarial-attacks.net/

- Machine Learning systems have **blind spots** - they can be purposefully mislead by malicious agents via **adversarial examples**
- **Adversarial examples** = special inputs created by adding customized noise to genuine inputs in order to induce misclassification by recognition systems
- The adversarial noise is usually imperceptible to humans [1]
- **Goal:** improve robustness of end-to-end automatic speech recognition (ASR) systems against adversarial noise

2. Methods and Experimental Setup

- Explore **MP3 compression** as a **countermeasure** to audio adversarial examples (AAEs)
- Build white-box & untargeted attacks with **Fast Gradient Sign Method** adapted for audio [2]





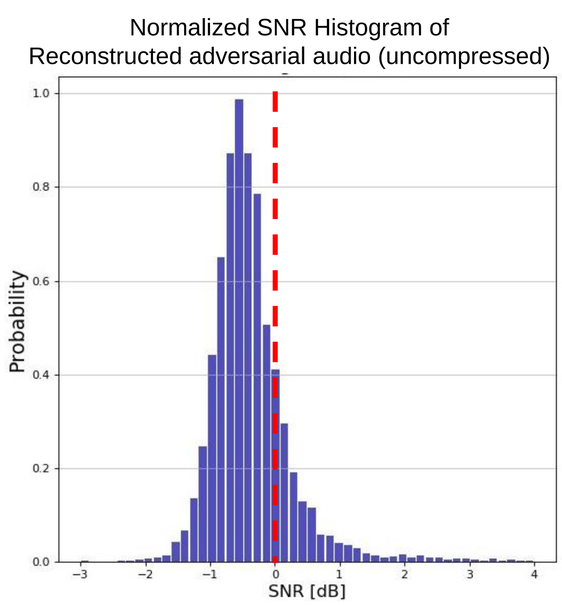
- Target ASR system: **ESPnet** [3] with **hybrid encoder-decoder architecture**
- Experiments performed on the **VoxForge speech corpus** [4]

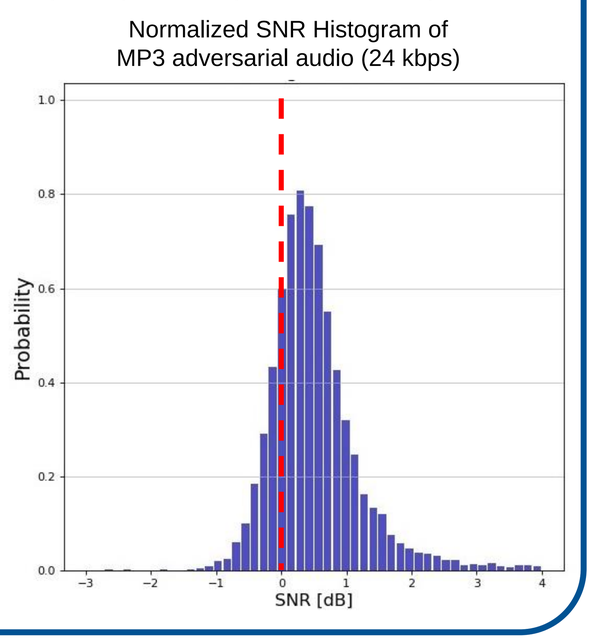
3. Results

Speech recognition results to *original* vs. *adversarial inputs*

Character Error Rates (CER) [%]		Test data			
		Original audio	Adv. features	Reconstructed Adv. audio	Compressed Adv. Audio (24 kbps)
ESPnet model (source format of input data)	#1 (uncompressed)	17.8	70.5	62.2	**57.4**
	#2 (128 kbps-MP3)	18.8	72.3	64	**58.4**
	#3 (64 kbps-MP3)	18.6	71.8	63.1	**56.5**
	#4 (24 kbps-MP3)	20.2	69	60.5	**55.3**

MP3-compressed adversarial samples have **higher SNRs** (hence less adversarial noise) than uncompressed adversarial inputs (KS test p-value: 0.019)





4. Conclusions & Outlook

- **MP3 compression:**
  - partially reduces the error rates to adv. samples (original transcription still not fully recovered)
  - has reverse effects for inputs augmented with non-adversarial noise (higher error rates)
  - has a significant effect in increasing the SNR of adversarial inputs, implying that MP3 reduces the adversarial noise
- **Future directions:** explore MP3 compression in **more complex attack** paradigms**:** targeted, black-box, over-the-air setting

5. References

[1] Schönherr, Lea, et al. (2018). *Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding,* arXiv preprint: 1808.05665

[2] Goodfellow, I. J., Jonathon S., and C. Szegedy (2014). *Explaining and harnessing adversarial examples* arXiv preprint: 1412.6572

[3] Watanabe, S., et al. (2018). *Espnet: End-to-end speech processing toolkit*, arXiv preprint: 1804.00015

[4] VoxForge Speech Corpus, available online *http://www.voxforge.org/home/downloads*