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

~ Master‘s Thesis presentation ~

Author: Iustina Andronic

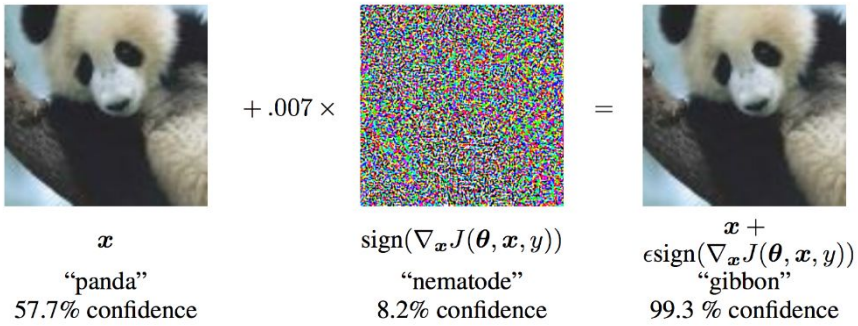
Supervisor: Prof. Dr. Ing. Bernhard Seeber
Scientific Advisor: Dipl.-Ing. (Univ.) Ludwig Kürzinger

March 23rd, 2020

Motivation and goals

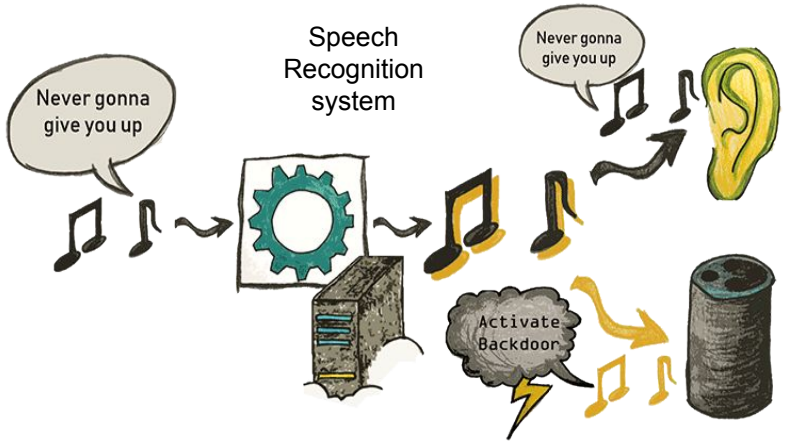
- Recognition systems have **blind spots** - they can be purposefully mislead

Image Domain



Goodfellow et al. 2015

Audio Domain



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

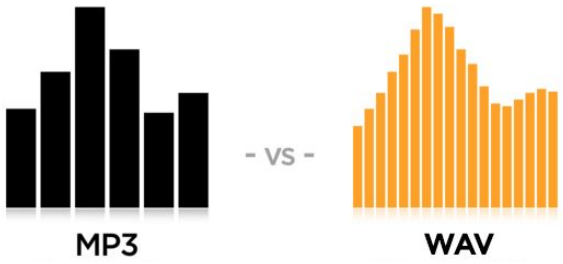
- The **adversarial noise**/perturbation is usually **imperceptible** to humans

Goal:

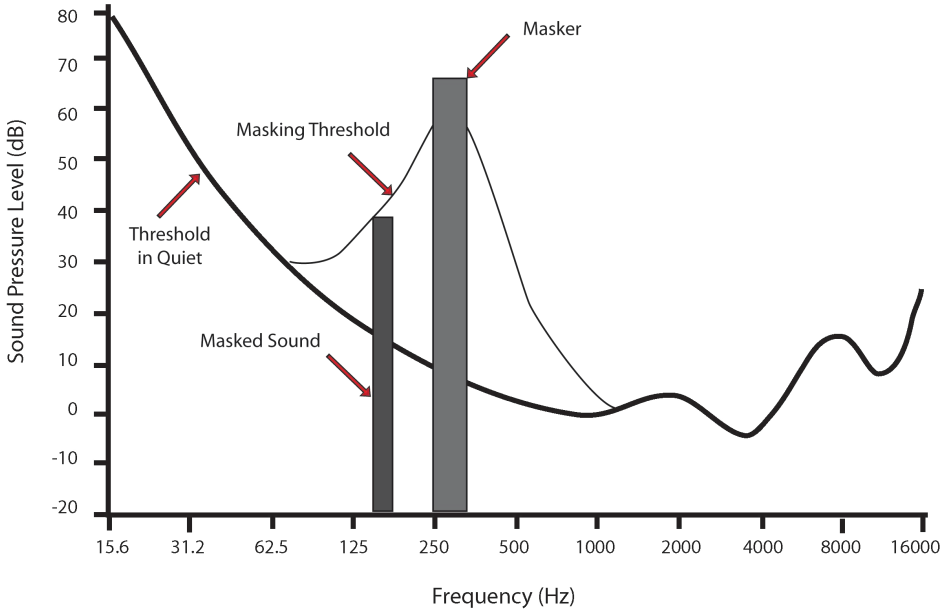
- **improve robustness** of end-to-end automatic speech recognition (ASR) systems **against adversarial noise**

Methods

- Explore **MP3-compression** as a countermeasure to adversarial examples (AdvEx)
- leverage the **mp3 perceptual audio encoder** that **discards inaudible frequency content**
- Compare ASR systems trained on .wav and .mp3 audio and assess their susceptibility/robustness to compressed adversarial noise

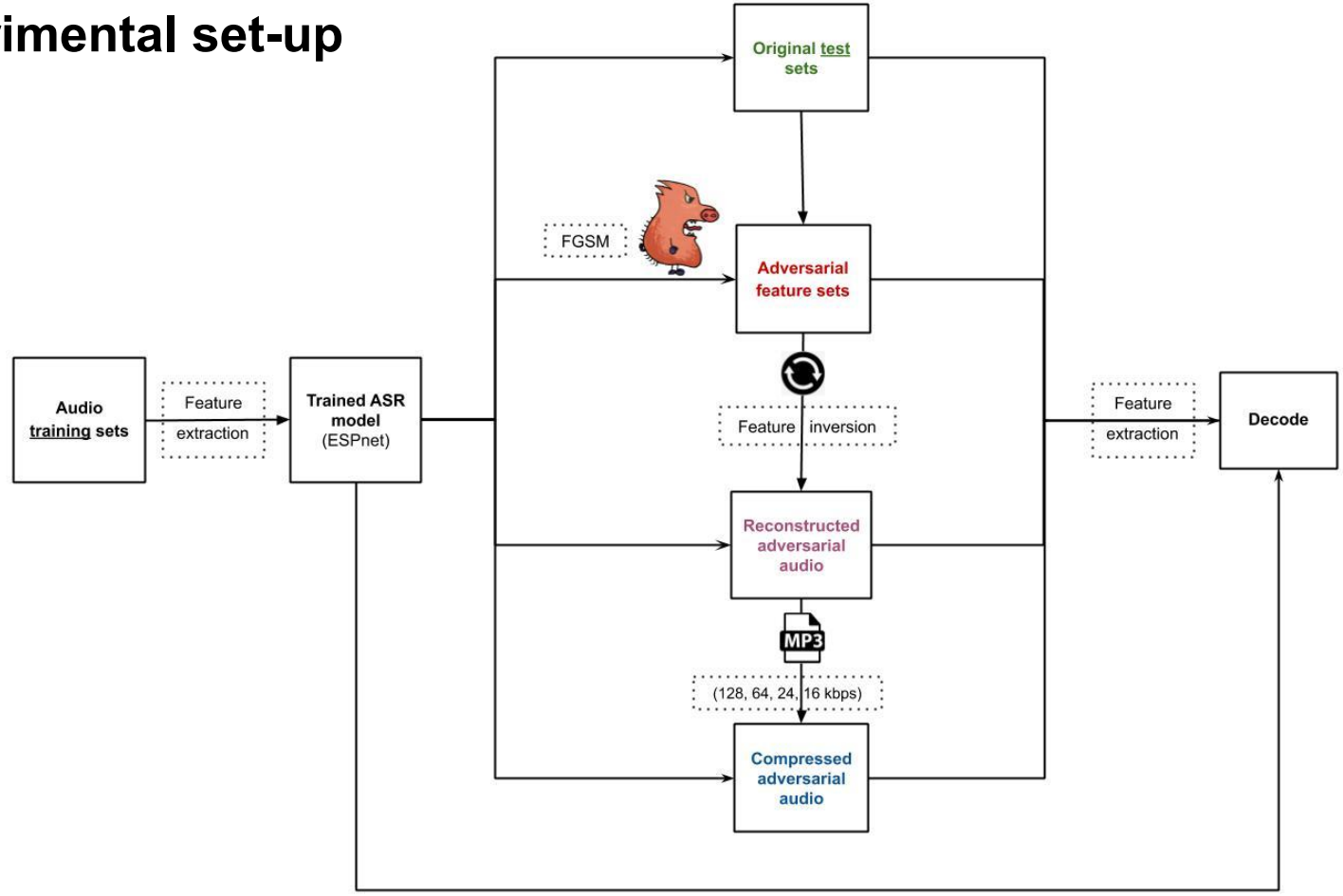


https://vox.rocks/resources/wav-vs-mp3



https://dsp.stackexchange.com/questions/37442/mp3-filterbank-mdct-why

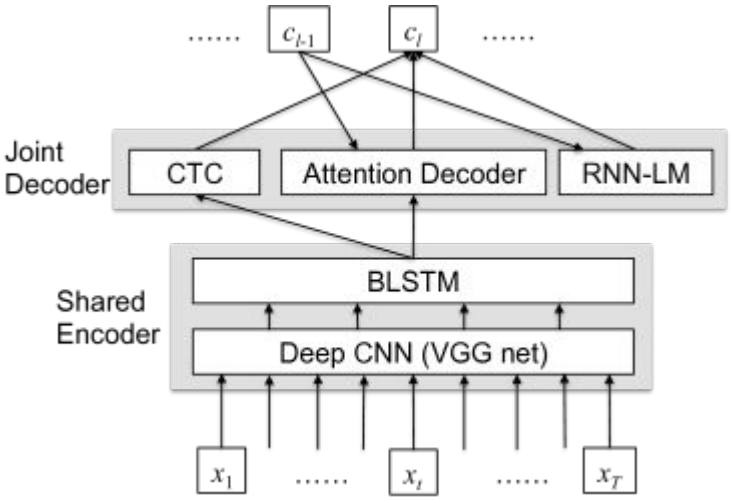
Experimental set-up



- **White-box** attack: the adversary has full access to the model and its parameters
- **Untargeted** attack: AdvEx are wrongly transcribed by the ASR system to any random phrase

ASR System

- **ESPnet** – End-to-End Speech Processing Toolkit - neural network directly mapping a sequence of input acoustic features into a sequence of graphemes/words
- hybrid **Connectionist Temporal Classification (CTC) / attention-based encoder-decoder** architecture



Cho et al. 2018

1. **Encoder** (analogous to Acoustic Model)
2. **Attention + CTC** (alignment model)
3. **Decoder** (analogous to Pronunciation and Language Models)

- training based on **multi-task learning:**

$$L_{hybrid} = \lambda L_{CTC} + (1 - \lambda) L_{Attention}$$

$\lambda = 0.3$

Speech corpus

- **VoxForge** - **small** open-source **uncompressed** speech database
- 131 hours of read English sentences (~5s / utterrance), limited vocabulary
- recorded by ~1200 speakers
- Audio specs: mono, 16-bit signed integer PCM, 16 kHz, 256 kbps

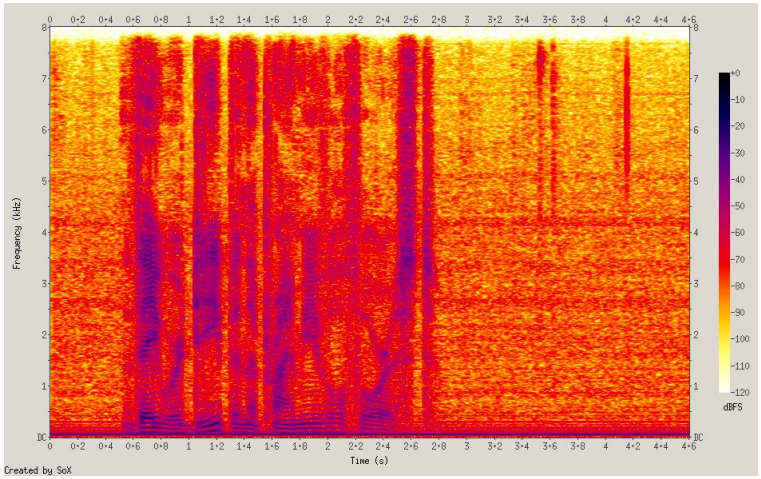
{ Train: 80%
Validation: 10%
Test: 10%

- Trained 4 hybrid ESPnet models:

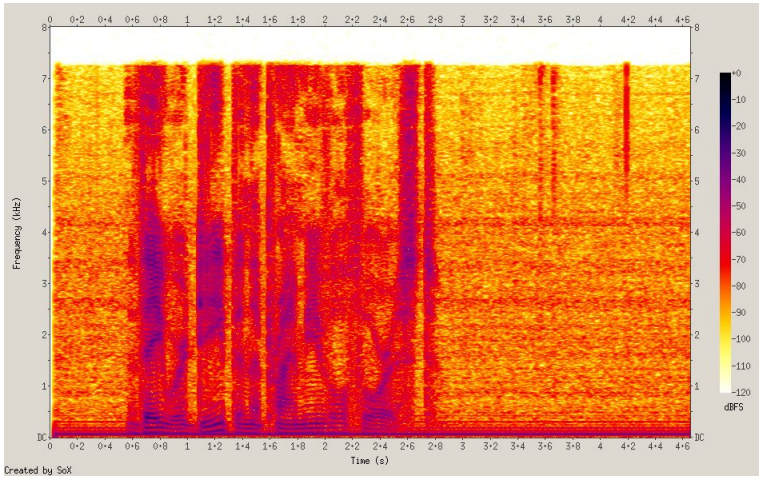
http://www.voxforge.org/

ESPnet model	Training data
#1	uncompressed (original)
#2	128 kbps - mp3
#3	64 kbps - mp3
#4	24 kbps - mp3

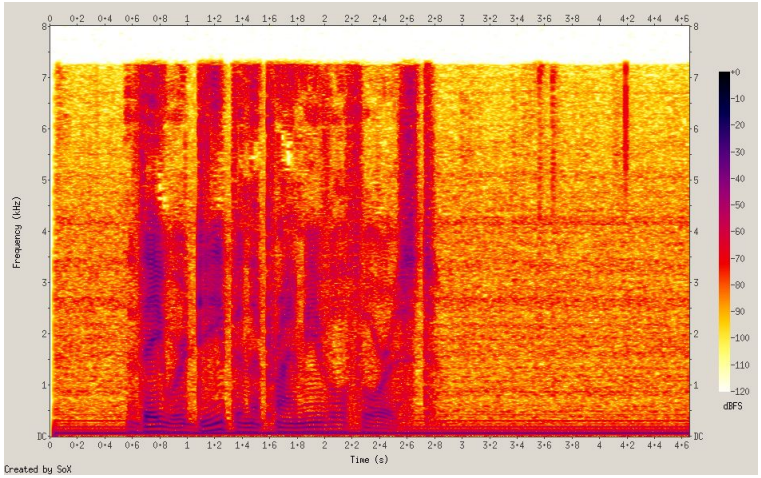
MP3-compression artefacts



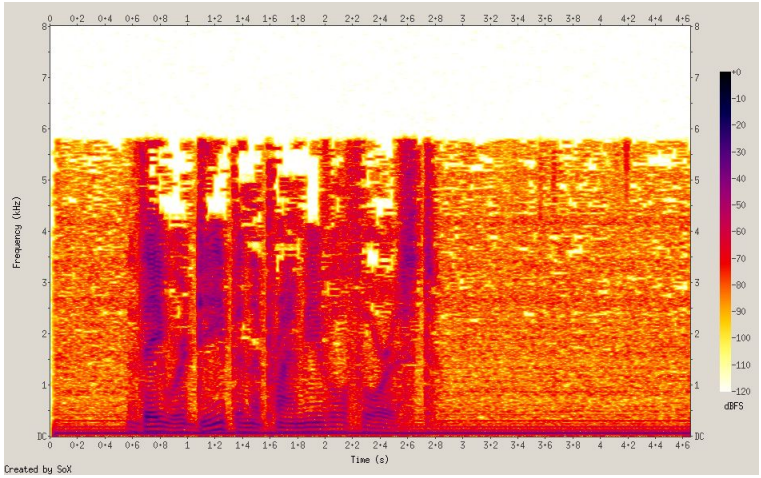
Original **.wav** utterance (*They were deep in the primeval forest*)



128 kbps .mp3



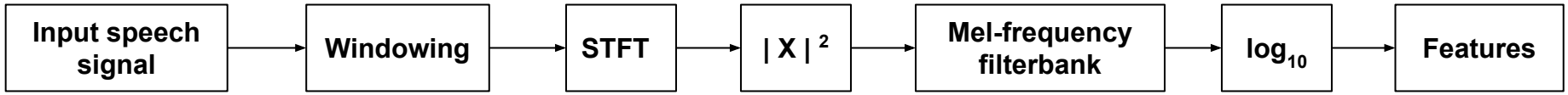
64 kbps .mp3



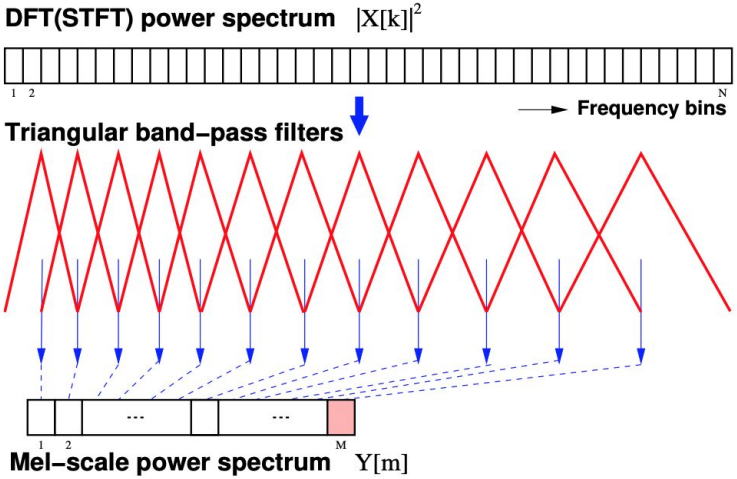
24 kbps .mp3

Feature extraction

Goal: simple, representative features that can easily be inverted



$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$



https://labrosa.ee.columbia.edu/doc/HTKBook21/node54.html

	Previously	Now
# of features	83	80
Type of features	mel fbanks + pitch	mel fbanks
CMVN	✔	X
Toolbox	Kaldi	Librosa

ASR baseline results

- Character/Word Error Rate (CER / WER)
- Results for original (clean, **non-adversarial**) test sets

Old features (Kaldi)

ESPnet model	Test data	CER [%]	WER [%]
#1	uncompressed (raw)	**17.3**	**40.7**
#2	128 kbps mp3	18.2	42.6
#3	64 kbps mp3	18	42
#4	24 kbps mp3	19	43.8

New features (Librosa)

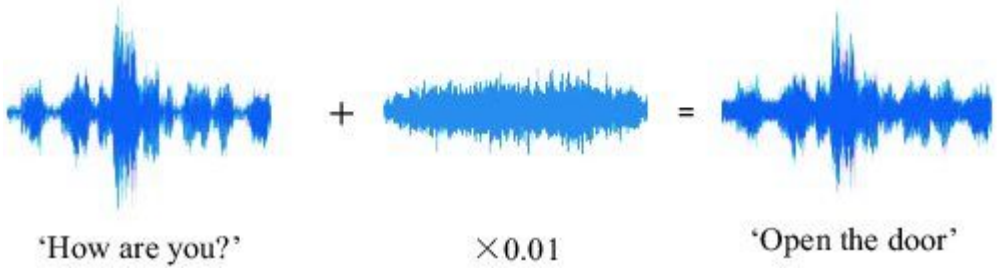
ESPnet model	Test data	CER [%]	WER [%]
#1	uncompressed (raw)	**17.8**	**41.4**
#2	128 kbps mp3	18.8	43.5
#3	64 kbps mp3	18.6	43
#4	24 kbps mp3	20.2	45.5

Takeaways:

- new features have *comparable performance* with previous Kaldi features (but are *easier* to invert!)
- Same trend: errors for compressed data are *slightly* higher

Generation of Adversarial Examples

- Start from the **speech** signal and compute the adversarial noise to change the transcription



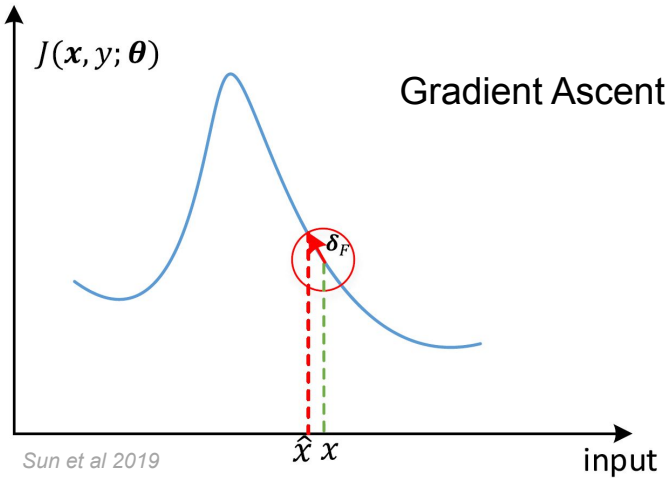
Gong, Poellabauer 2019

- Untargeted attack: **Fast Gradient Sign Method** (FGSM) - finds the worst possible noise

$$\hat{x} = x + \delta$$

$$\delta_F = \epsilon \, \text{sign}(\nabla_x J(x, y; \boldsymbol{\theta}))$$

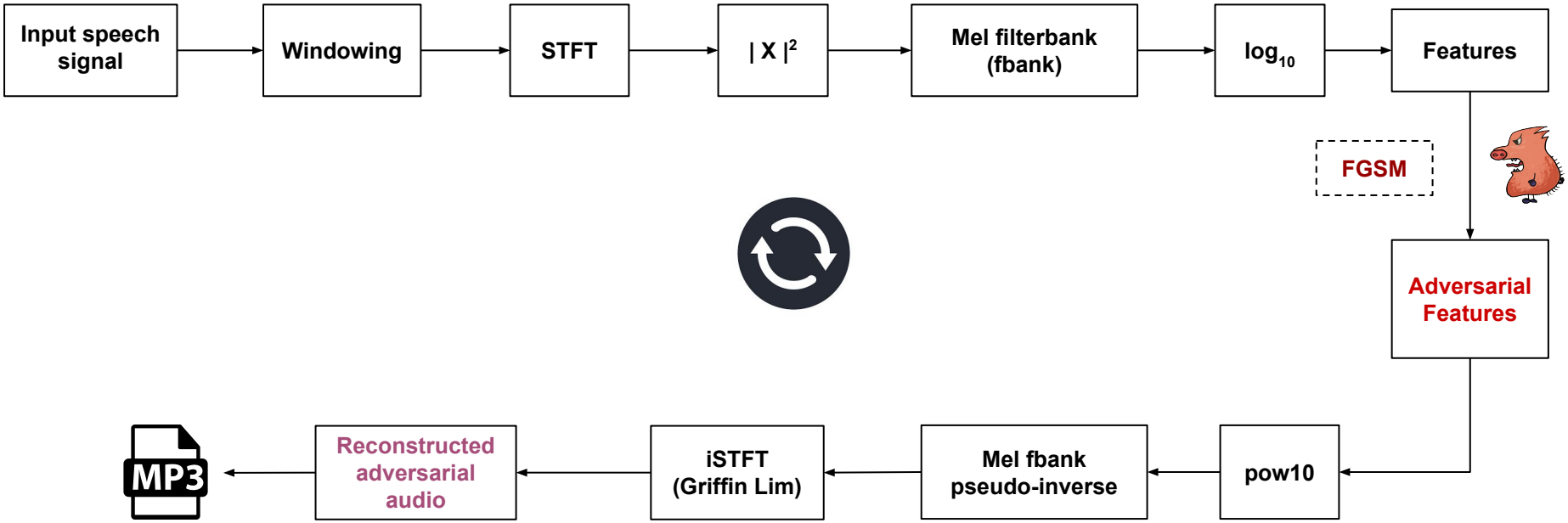
$$y \neq f(\hat{x}; \boldsymbol{\theta})$$



Sun et al 2019

Adversarial audio reconstruction

- FGSM creates **adversarial features**, but we need **adversarial audio**
- Feature inversion (*Griffin Lim* algorithm - random phase estimation for iSTFT)



- **Audio samples**

Original audio

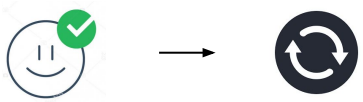
Reconstructed audio from **original** features

Reconstructed audio from **adversarial** features

ASR results for reconstructed features

- Validation of the reconstruction process on the original (clean) test sets

CER [%]	Test data	
	Original audio	Reconstructed audio
ESPnet model #1 (uncompressed data)	**17.8**	19.1



Takeaway

- comparable ASR performace when using reconstructed audio from feature inversion

ASR results for Adversarial Input

- Reconstruct audio from the adversarial features

CER [%]		Test data		
		Original audio	Adv. features	Reconstructed adv. audio
ESPnet model (test data type)	#1 (uncompressed)	**17.8**	70.5	62.2
	#2 (128 kbps-mp3)	18.8	72.3	64
	#3 (64 kbps-mp3)	18.6	71.8	63.1
	#4 (24 kbps-mp3)	20.2	**69**	**60.5**

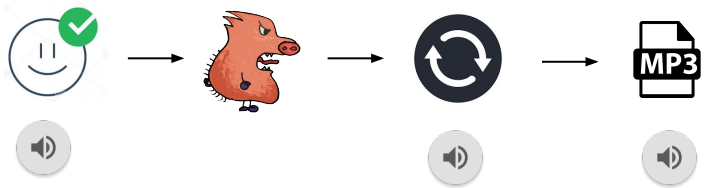
- **Takeaway:**
  - adversarial reconstruction already has a beneficial effect in reducing CER

- **Next:** re-encode the reconstructed adversarial audio with mp3

ASR results for compressed Adversarial Input

- Apply mp3-compression (24kbps) to the adversarial reconstructed audio

CER [%]		Test data				Relative CER difference [%]
		Original audio	Adv. features	Reconstructed Adv. audio	Compressed Adv. Audio (24 kbps)	
ESPnet model (format of train data and adversarial origin)	#1 (uncompressed)	17.8	70.5	62.2	**57.4**	-18.58
	#2 (128 kbps-mp3)	18.8	72.3	64	**58.4**	-19.23
	#3 (64 kbps-mp3)	18.6	71.8	63.1	**56.5**	**-21.31**
	#4 (24 kbps-mp3)	20.2	69	60.5	**55.3**	-19.86



- **Takeaway:**
  - mp3-compression further reduces error rates to AdvEx, regardless of the model and adversarial origin data type

ASR results for compressed Adversarial Input originating from compressed data

- Results only for **ESPnet #1** - model trained on **uncompressed data**

CER [%]		Adversarial test sets	
		Reconstructed adv. audio	24 kbps-mp3
Source format of test data	uncompressed	62.2	57.4
	128 kbps-mp3	51.7	47.5
	64 kbps-mp3	50.7	**46.5**
	24 kbps-mp3	53.8	49.3

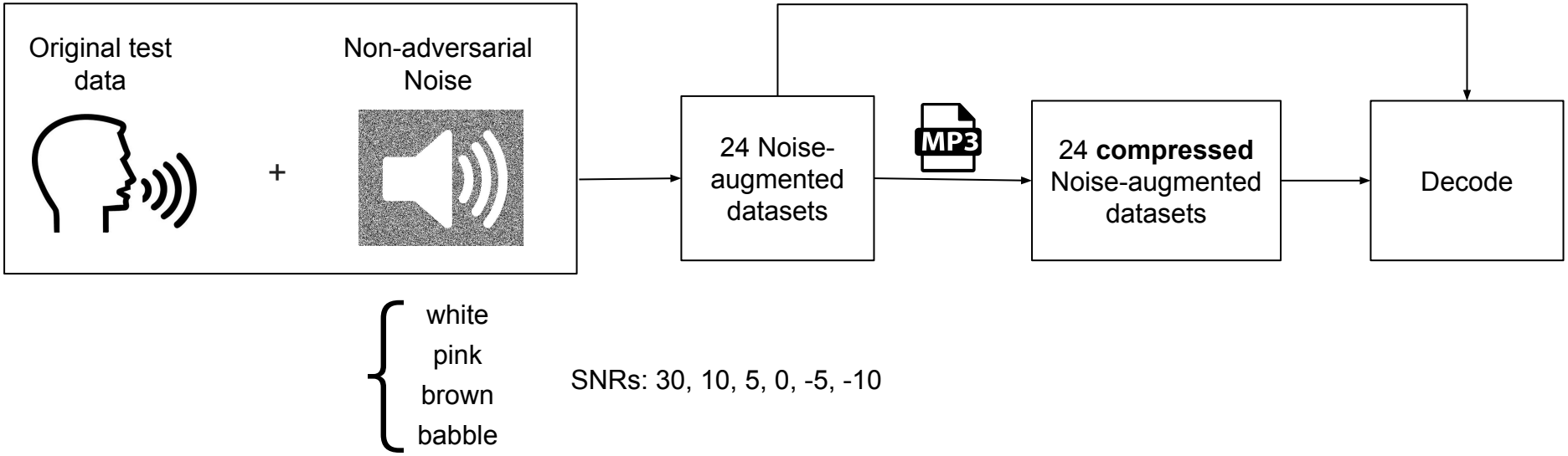
-25% relative CER reduction

Takeaway:

- compressed AdvEx originating from **already compressed data** further *decrease the error rate* of a model trained on uncompressed data

Noise Augmentation experiments

- What happens when we compress samples augmented with non-adversarial noise?
- non-adversarial noise as comparison baseline



Noise Augmentation results

- Results for **ESPnet model #1** (trained on uncompressed data)

Uncompressed data augmented with noise

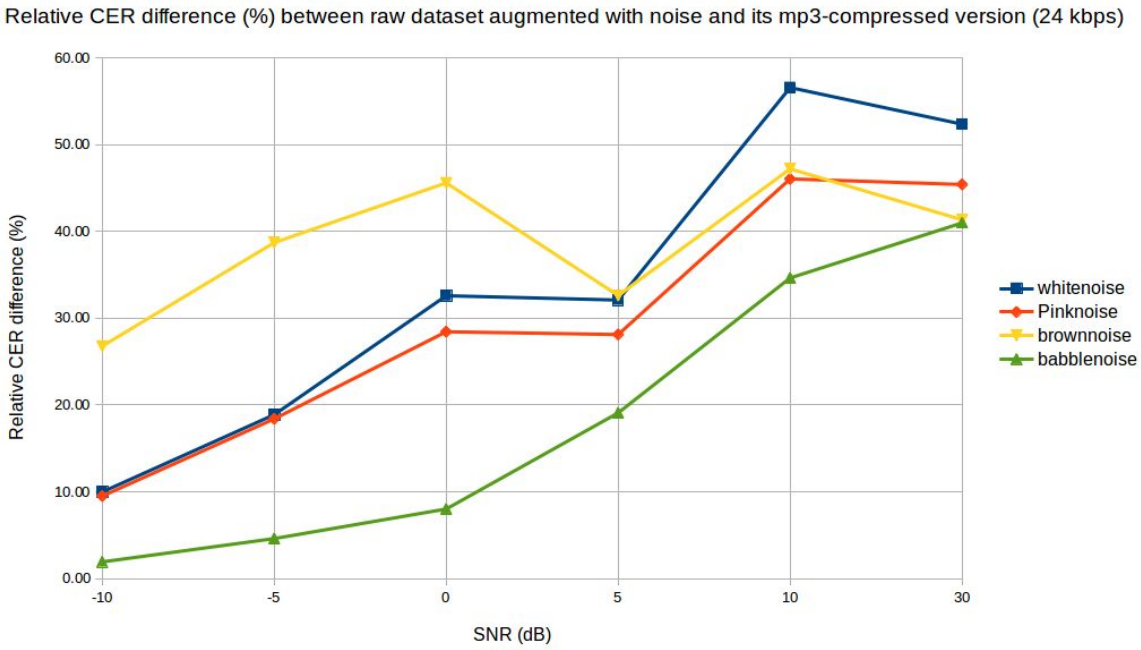
CER [%]	SNR [dB]					
	-10	-5	0	5	10	30
white noise	78.2	66.2	53.7	41.9	32.7	19.1
pink noise	82.1	67.4	51.7	38.1	29.1	18.5
brown noise	47.8	34.1	26.1	21.9	19.7	17.9
babble noise	93.6	89	77.4	53.4	35.8	18.3

24 kbps-mp3 data augmented with noise

CER [%]	SNR [dB]					
	-10	-5	0	5	10	30
white noise	86	78.7	71.2	61.7	51.2	29.1
pink noise	89.9	79.8	66.4	53	42.5	26.9
brown noise	60.6	47.3	38	32.5	29	25.3
babble noise	95.4	93.1	83.6	66	48.2	25.8

Noise Augmentation results

- Results for **ESPnet model #1** (trained on uncompressed data)



Takeaways

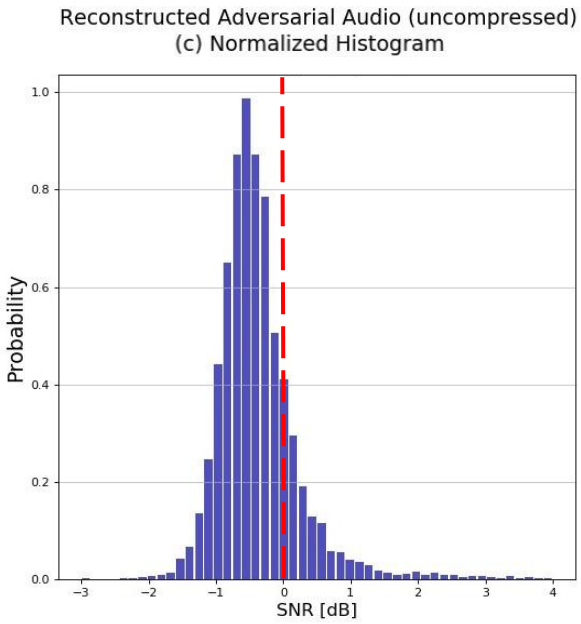
- only positive values => **compressing** audio samples augmented with *non-adversarial noise* **increases the errors**
- **Reverse trend** compared to how **compressed adversarial noise** behaved

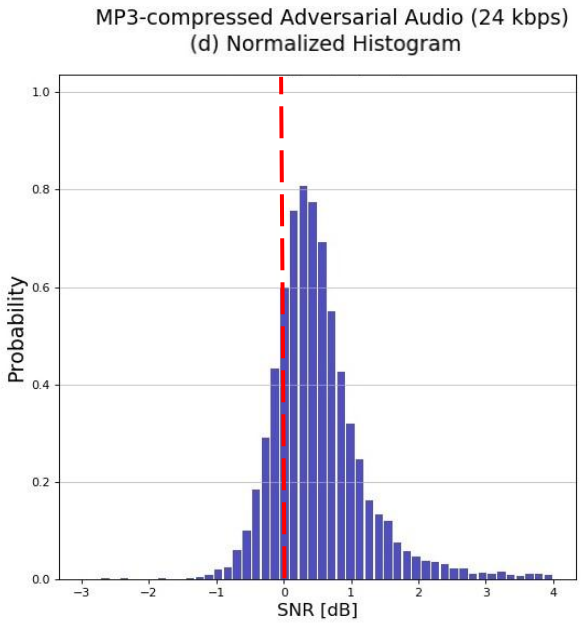
Estimation of Signal-to-Noise Ratio (SNR) of Adversarial Inputs

Q: Does MP3 compression have any **effect on the SNR** of adversarial samples?

$$\mathrm{SNR}_{\mathrm{adv}} = 10 \log_{10} \frac{\text{signal power}}{\text{adversarial noise power}}$$

$$= 10 \log_{10} \frac{\text{power of the reconstructed speech (non-adversarial)}}{\text{power of the adversarial reconstructed noise}}$$





Kolmogorov-Smirnov (KS) test:
p_val = 0.02

Takeaway:

MP3 compression *increases* the SNR of adversarial samples => reduces the adversarial noise

Conclusions

- We built **untargeted** adversarial samples with FGSM with and end-to-end ASR model
- Compressing adversarial samples *partially* **reduces** the transcription error rates
- A model trained on uncompressed data is **more robust** to compressed adversarial examples originating from *already* compressed data
- ASR systems are **not robust to** compressed inputs augmented with **non-adversarial noise** (higher error rates)
- MP3 **compression significantly increases the SNR** of adversarial samples => adversarial noise is reduced

Final takeaway

- mp3-compression does **reduce the effectiveness** of adversarial samples, but original transcription is still not fully recovered

Future directions

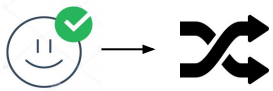
- How would mp3-compression behave for **more complex attacks** paradigms ?
  - targeted
  - black-box
  - over-the-air

Thank you!

Back Up

ASR results for cross-testing

- **Cross-test:** introduce a **mismatch** between the audio format of the train and test datasets
- **CER [%]** scores reported (only for the *new* Librosa features)



CER [%]		Test data			
		uncompr. (raw)	128 kbps-mp3	64 kbps-mp3	24 kbps-mp3
ESPnet model (train data)	#1 (uncompressed)	17.8	**6**	**6**	13.3
	#2 (128 kbps-mp3)	6.4	18.8	18.8	25
	#3 (64 kbps-mp3)	**5.5**	18.6	18.6	24.7
	#4 (24 kbps-mp3)	7.8	20.2	20.2	20.2

Takeaway: train-test audio format mismatch seems beneficial for more accurate speech recognition

ASR results for cross-testing with (compressed) Adversarial Input

- **Cross-testing:** introduce a **mismatch** between the audio format of the train and adversarial test datasets
- Results only for **ESPnet #3** - model trained on **64 kbps mp3 data**

CER [%]		Adversarial test sets	
		Reconstructed adv. audio	24 kbps-mp3
Original test sets	uncompressed	49.5	**47.5**
	128 kbps-mp3	57.6	53.4
	64 kbps-mp3	63.1	56.5
	24 kbps-mp3	61	54.5

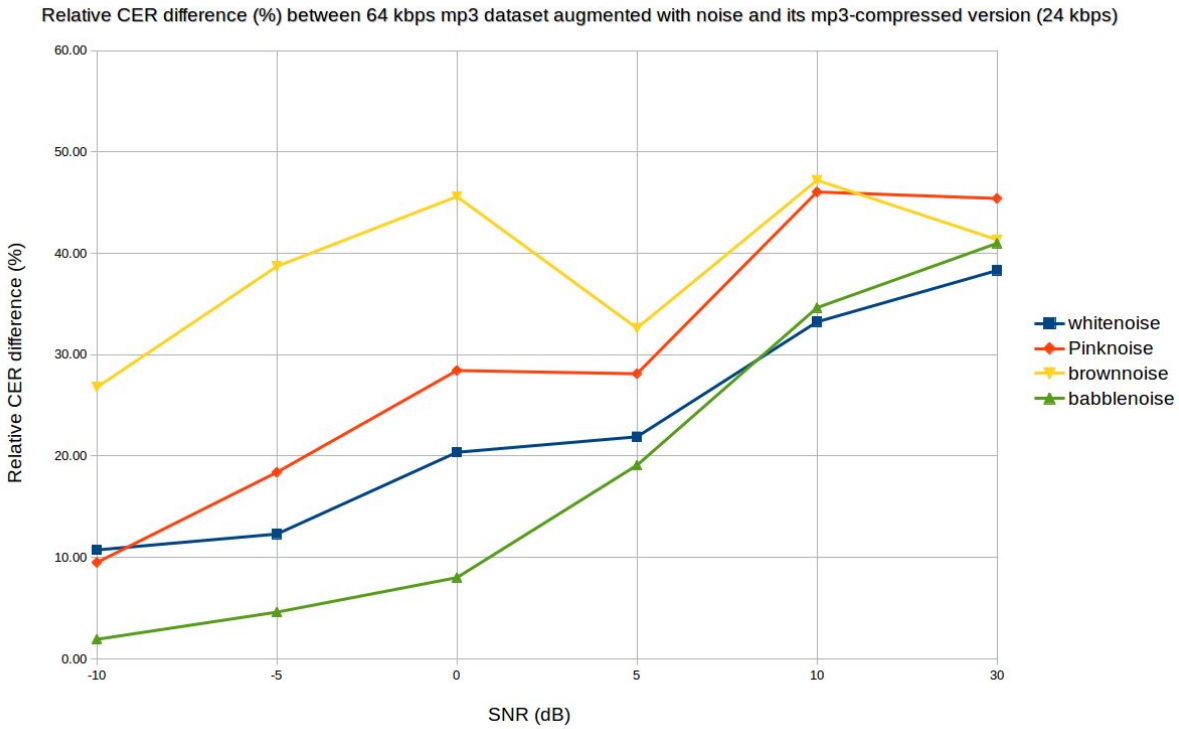
-22.13 % relative CER reduction

Takeaway:

- compressed adversarial examples originating from **uncompressed data**, further reduce the error rate of a model trained on 64kbps compressed data

Noise Augmentation results

- Results for **ESPnet model #3** (trained on 64 kbps-mp3 data)



Takeaway

- only positive values => **compressing** audio samples augmented with *non-adversarial noise* **increases the errors**
- *Reverse trend* compared to how compressed adversarial noise behaved

ASR performance to compressed Adversarial Input

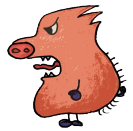
- Mismatch scenario

Inputs given for recognition

Original data type for crafting adv. input

CER [%]	clean audio features	clean audio reconstructed	Adv. features	Adv. audio reconstructed	Compressed Adv. Audio (24 kbps)
uncompressed Voxforge			70.5	62.5	
128 kbps mp3 -> wav			72.3	64.2	
64 kbps mp3 -> wav			71.8	63.2	
24 kbps mp3 -> wav			69	60.9	

- **Takeaway:** mp3-compression reduces error rates to AdvEx, but not to a significant extent
- **Possible solution:** re-encode the AdvEx with mp3



Next steps

	ASR model trained on raw audio	ASR model trained on compr. audio
Raw & clean test set	✔	✔
Compr. & clean test set	✔	✔
Raw & AdvEx test set	✔	TO DO
Compr. & AdvEx test set	TO DO	✔

- **Cross-testing**: feed adversarial uncompressed inputs to ASR models trained on compressed data and vice-versa
- use **psychoacoustic hiding principle** to add **imperceptible perturbation** - under the frequency masking threshold of the original audio