Characterizing Speech Adversarial Examples Using Self-Attention U-NET Enhancement

Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, Chin-Hui Lee

45th IEEE International Conference on Acoustics, Speech and Signal Processing

October 17, 2020

Introduction

- Speech enhancement based defense technique.
- U-Net with self-attention.
- Tested against gradient and gradient-free attacks.

Attack Types

- Two types of attacks are used for generating perturbations in this experiment.
  - Gradient based attack. Yakura and Sakuma (2019)
  - Evolutionary optimization. Khare et al. (2019)

Attack

- Yakura and Sakuma (2019)
  - Incorporates transformations caused by playback and recording into generation process.
  - Band pass filter, impulse response and white Gaussian noise.

$$\underset{\boldsymbol{v}}{\operatorname{argmin}} \mathbb{E}_{h\sim\mathcal{H},\boldsymbol{w}\sim\mathcal{N}(0,\sigma^2)}[\operatorname{Loss}(MFCC(\tilde{\boldsymbol{x}}),\boldsymbol{y}) + \epsilon\|\boldsymbol{v}\|]$$
$$\text{where } \tilde{\boldsymbol{x}} = \operatorname{Conv}(\boldsymbol{x} + \underset{1000\sim4000\text{Hz}}{\operatorname{BPF}}(\boldsymbol{v})) + \boldsymbol{w}$$

- Khare et al. (2019)
  - Proposed evolutionary-based method focus on maximizing a fitness function

U-Net with Self-Attention

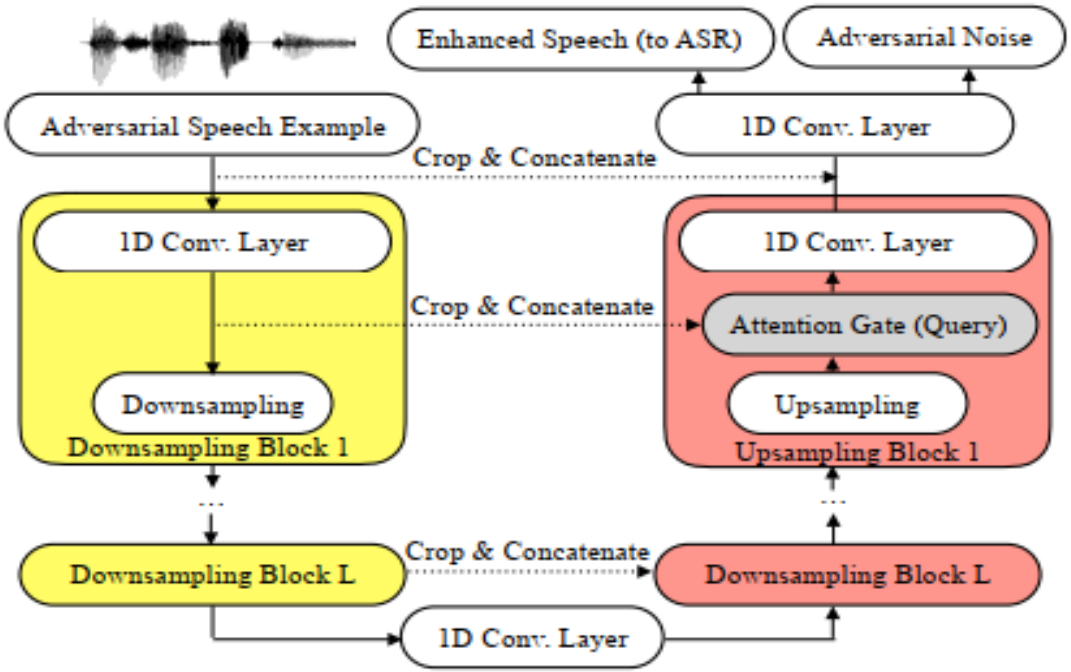


Fig. 1: The proposed a self-attention U-Net (U-Net$_{At}$) structure for improving the adversarial robustness by processing before ASR.

Dataset & Evaluation Metric

- LibriSpeech
  - Clean data.
  - Mixed with noise source from DEMAND dataset for noisy data.
- Speech Enhancement Evaluation Metric
  - Perceptual Evaluation of Speech Quality (PESQ)
  - Speech Transmission Index (STI)
  - Short Term Objective Intelligibility (STOI)
  - Signal to Noise Ration (SNR)
- Robustness Evaluation Metric:
  - Rate of Success Attack (ROSA)
  - Word Error Rate (WER)

Speech Enhancement

Metric	Noisy$_D$	DNN	U-Net$_W$	U-Net$_{At}$
PESQ	1.97	2.62	2.86	**2.88**
STI	0.65	0.73	**0.81**	**0.81**
STOI	0.82	0.90	**0.93**	0.92
SNR	-1.63	7.67	9.83	**9.85**

Table 1: We evaluate the untreated noisy signal (Noist$_D$) in [24], and the enhanced signals based on DNN, wave U-Net (as U-Net$_W$), and self -attention U-Net (as U-Net$_{At}$). The experimental results show that the U-Net based methods attain higher scores compared with DNN-based methods on the noisy speech.

Metric	Noisy$_{adv}$	DNN	U-Net$_W$	U-Net$_{At}$
PESQ	**1.31**	1.21	1.16	1.18
STI	**0.67**	0.66	0.62	0.64
STOI	**0.84**	0.81	0.80	0.81
SNR	-1.52	7.23	7.43	**7.68**

Table 2: We repeat the experiments of speech enhancement in Table 1 where the over-the-air adversarial speech examples (Noisy$_{adv}$) [6] were imposed to the ASR loss function. The evaluation results show that although all of the SNR scores increase by the speech enhancement, the other metric indexes are even lower than the Noisy$_{adv}$ before conducting speech enhancement.

Speech Enhancement

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1-\alpha) J\left(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)\right)$$

Metric	Noisy_{adv}	DNN_{T}	$\text{U-Net}_{T,W}$	$\text{U-Net}_{T,At}$
PESQ	1.31	2.55	2.72	**2.78**
STI	0.67	0.69	0.72	**0.75**
STOI	0.84	0.86	0.88	**0.90**
SNR	-1.52	7.45	7.67	**7.92**

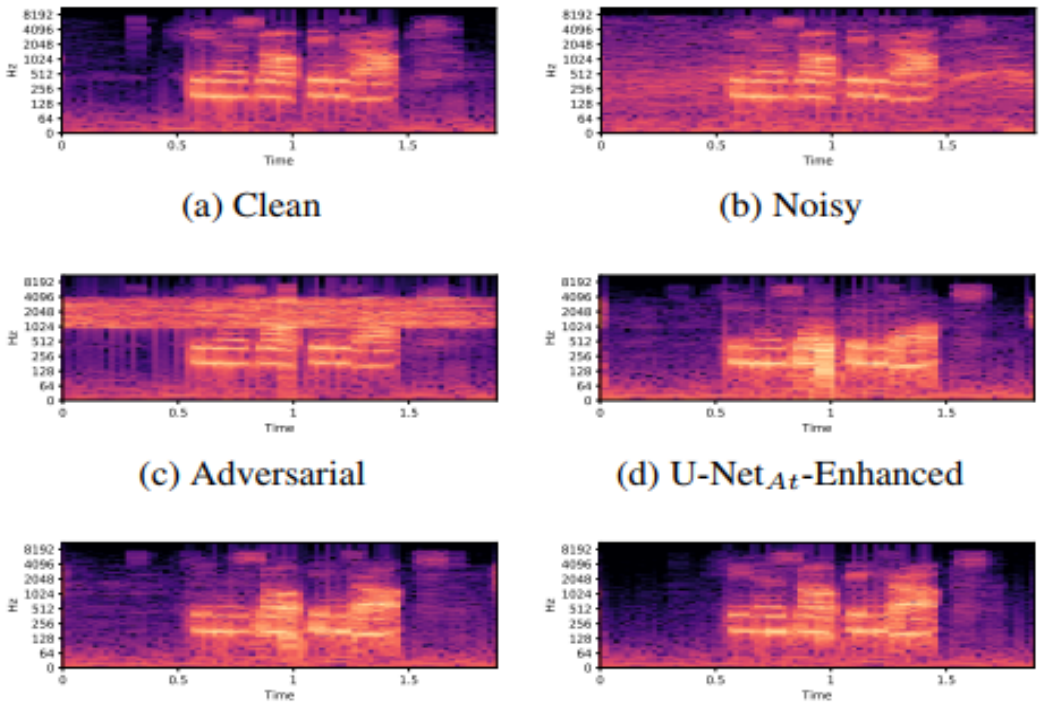
Table 3: To further improve the model generalization capability, we add adversarial examples into the dataset for the adversarial training. We observe that the model obtains a further improvement in terms of the intellectual speech quality based on the methods of the adversarial training on DNN (DNN_{T}), wave U-Net ($\text{U-Net}_{T,W}$), and self-attention U-Net ($\text{U-Net}_{T,At}$) with slightly improved SNRs.

Robustness

Method	Grad_{adv}	DNN	U-Net_{W}	U-Net_{At}
1: Baseline-ROSA	90.23	84.86	83.43	**83.11**
2: SE_{AdvT}-ROSA	27.34	22.21	19.29	**18.23**
1: Baseline-WER	85.90	72.97	67.46	**66.12**
2: SE_{AdvT}-WER	19.37	18.92	17.64	**17.15**
Method	Evo_{adv}	DNN	U-Net_{W}	U-Net_{At}
3: Baseline-ROSA	91.21	85.67	82.03	**79.35**
4: SE_{AdvT}-ROSA	20.47	18.45	17.81	**16.14**
3: Baseline-WER	87.90	83.12	79.20	**71.12**
4: SE_{AdvT}-WER	19.45	19.42	18.44	**17.42**

Table 4: To improve the adversarial robustness of the ASR, we modify the input with the U-Net_{At}-based speech enhancement with adversarial training (SE_{AdvT}). The use of gradient-based adversarial examples (Grad_{adv} in [6]) can improve the performance in terms of the rate of succeed attack (ROSA) and the word error rate (WER %). Compared with Grad_{adv}, the evolutionary-optimized [5] adversarial examples (Evo_{adv} in [5]) exhibits a higher error rate without SE_{AdvT}, but WER under SE_{AdvT} can obtain more gains than the Grad_{adv}. Targeted text output is **"open-the-door"**.

Visualization



(a) Clean (b) Noisy

(c) Adversarial (d) U-Net$_{At}$-Enhanced

(e) AdvT-U-Net$_{W}$-Enhanced (f) AdvT-U-Net$_{At}$-Enhanced

Fig. 2: The log-power spectrogram of (a) clean; (b) noisy; (c) adversarial; (d) pre-trained U-Net$_{At}$ enhanced adversarial examples; (e) DNN enhancement results adopted adversarial training, and (f) proposed U-Net$_{At}$ using adversarial training (AdvT.)

Bibliography

Khare, S., Aralikatte, R., and Mani, S. (2019). Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization. In Kubin, G. and Kacic, Z., editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3208–3212. ISCA.

Yakura, H. and Sakuma, J. (2019). Robust audio adversarial example for a physical attack. In Kraus, S., editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5334–5341. ijcai.org.