A T-NES Attack

Algorithm 2 T-NES Attack

Input: original audio $\boldsymbol{X}$, ground truth label y, query budget q, ASR model F, CTC loss function $\mathcal{L}(\cdot)$, Character error rate $CER()$

Parameter: coordinates to perturb k, sliding window of length h and stride 1, random noise U, variance of NES search distribution σ, NES population size n, perturbation strength b, termination error threshold c

Output: successful AE $\boldsymbol{X}_{adv}$

```
Initialize iteration $t \leftarrow 0$
Initialize perturbed audio $\boldsymbol{X}_{\text{adv}} \leftarrow \boldsymbol{X} + U$
while $t < q$ and $CER(F(\boldsymbol{X}_{adv}), y) - CER(F(\boldsymbol{X}), y) < c$ do
    $k \leftarrow TDCoordinateSelection(\boldsymbol{X}_{\text{adv}}, \boldsymbol{X})$
    estimate gradients $\hat{\boldsymbol{g}} \leftarrow \frac{1}{\sigma n} \sum_{i=1}^{n} \boldsymbol{v}_i \mathcal{L}(\boldsymbol{X} + \sigma \boldsymbol{u}_i)$
    $\hat{\boldsymbol{g}} \leftarrow clip(\hat{\boldsymbol{g}}, -\boldsymbol{X} \cdot b, \boldsymbol{X} \cdot b)$
    $\boldsymbol{X}_{\text{adv}} \leftarrow \boldsymbol{X} + \hat{\boldsymbol{g}}$
end while
return $\boldsymbol{X}_{\text{adv}}$
```

B ASR Models

	LibriSpeech	TEDLIUM
DeepSpeech (LSTM)	2.98%	10.55%
DeepSpeech (GRU)	8%	(not used)
Wav2Letter	9.4%	(not used)

Table B.1: Testing CERs of different ASR models.

Since training ASR models is very time-consuming, we use publicly available pre-trained models of DeepSpeech2 (LSTM)[1] and Wav2Letter[2]. We train DeepSpeech2 (GRU)[3] on LibriSpeech by ourselves. The testing CERs of different models are summarized in Table B.1. We follow the provided instruction of each ASR model to extraction acoustic features. The inputs to both DeepSpeech2 (LSTM) and DeepSpeech2 (GRU) are spectrums with the size of 161. Wav2Letter is trained with Fbank with each frame size of 128.

C Attack Hyperparameters

For T-NES, we use a Gaussian search distribution with variance $\sigma = 0.0001$ and population size 2 [4]. For FD and ZOO-based attacks, we examine generated examples every 50 iterations, and other settings follow [Chen *et al.*, 2017] [5]. For GA, we follow [Taori *et al.*, 2019] and switch GA to FD after GA becomes less effective when closing to the target [6]. For HSJA and Time Prior NES, we follow the implementation in [Chen *et al.*, 2020a] [7] and [Andrew *et al.*, 2018] [8], respectively.

D Full Attack Results

We show the full attack results of perturbing 200 coordinates in Table D.1 and Table D.2. We show the attack results of perturbing 500 coordinates in Table D.3 and Table D.4. Overall, the results are consistent with perturbing 200 coordinates.

E Stealthiness Analysis Results

The PESQ and user study ratings are shown in Figure E.1. In the user study, the 5-degree rating for use study are annoying (1), slight annoying (2), acceptable (3), almost satisfied (4), satisfied (5). The results show that the quality of generated AE is acceptable for humans.

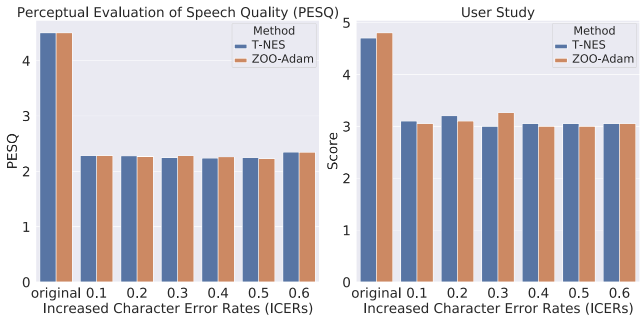

Figure E.1: Left: PESQ for AEs generated from T-NES and ZOO-Adam. Right: User study results for T-NES and ZOO-Adam. Better audio quality achieves higher PESQ and user study scores.

To visually compare the difference between the original audios and the AEs, we plot the spectrograms of one audio and its AEs in Figure E.2. AEs exhibit very similar patterns with the original audio. The perturbations bring more dark areas to the frequency domain (e.g., over 2048Hz), but such noises have less impact on audio quality as human is less sensitive to high frequency noises.

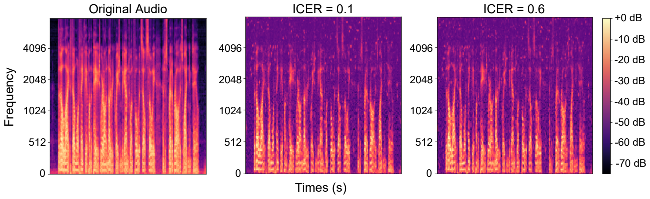

Figure E.2: Spectrograms of an original audio and AEs. X-axis and Y-axis represent the time and frequency domain, respectively. Darker color indicates lower magnitude.

[1] https://github.com/SeanNaren/deepspeech.pytorch

[2] https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition/wave2letter

[3] https://github.com/jiwidi/DeepSpeech-pytorch

[4] https://github.com/labsix/limited-blackbox-attacks

[5] https://github.com/IBM/ZOO-Attack

[6] https://github.com/rtaori/Black-Box-Audio

[7] https://github.com/cleverhans-lab/cleverhans/blob/master/cleverhans/torch/atta

[8] https://github.com/MadryLab/blackbox-bandits

Dataset	Method	Attack Success Rates Under Different Increased Character Error Rates (ICERs)					
		ICER=0.1	ICER=0.2	ICER=0.3	ICER=0.4	ICER=0.5	ICER=0.6
LibriSpeech	FD	**100% (117)**	**100% (207)**	85% (234)	70% (256)	60% (279)	35% (317)
	ZOO-Adam	95% (170)	90% (256)	75% (278)	60% (288)	50% (305)	50% (337)
	ZOO-Newton	90% (169)	80% (251)	50% (251)	40% (280)	35% (280)	30% (278)
	GA	100% (167)	95% (227)	80% (250)	70% (265)	70% (310)	50% (323)
	Time Prior NES	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)
	HSJA	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)
	T-NES	95% (158)	95% (253)	**85% (231)**	**80% (248)**	**75% (263)**	**70% (287)**
TEDLIUM	FD	70% (232)	15% (396)	5% (355)	5% (479)	0% (N/A)	0% (N/A)
	ZOO-Adam	90% (253)	65% (280)	45% (274)	40% (289)	40% (318)	35% (316)
	ZOO-Newton	75% (242)	50% (265)	35% (259)	30% (266)	30% (296)	30% (320)
	GA	95% (237)	45% (266)	30% (257)	30% (310)	25% (383)	15% (410)
	Time Prior NES	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)
	HSJA	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)	0% (N/A)
	T-NES	**100% (176)**	**90% (188)**	**75% (197)**	**75% (248)**	**70% (255)**	**70% (289)**

Table D.1: Performance of different attacks on the same *DeepSpeech2 (LSTM)* model but across different datasets including *LibriSpeech* and *TEDLIUM*. 200 coordinates are perturbed in one query. The cells are in the form of "Attack Success Rate% (#queries)".

Model	Method	Attack Success Rates Under Different Increased Character Error Rates (ICERs)					
		ICER=0.1	ICER=0.2	ICER=0.3	ICER=0.4	ICER=0.5	ICER=0.6
DeepSpeech2 (GRU)	FD	100% (6)	100% (11)	100% (22)	100% (41)	100% (71)	95% (88)
	ZOO-Adam	100% (9)	100% (17)	100% (33)	100% (64)	95% (81)	95% (124)
	ZOO-Newton	100% (9)	100% (17)	100% (34)	100% (59)	85% (70)	85% (130)
	GA	100% (145)	100% (152)	100% (164)	100% (175)	85% (210)	85% (220)
	T-NES	**100% (6)**	**100% (11)**	**100% (18)**	**100% (34)**	**95% (50)**	**95%(80)**
Wav2Letter	FD	80% (267)	55% (289)	40% (285)	30% (268)	25% (335)	10% (241)
	ZOO-Adam	100% (143)	100% (222)	85% (237)	80% (276)	65% (275)	50% (267)
	ZOO-Newton	100% (181)	90% (251)	70% (280)	40% (193)	40% (230)	35% (272)
	GA	10% (318)	5% (154)	5% (156)	5%(166)	5% (276)	0% (N/A)
	T-NES	**100% (152)**	**100% (170)**	**100% (187)**	**100% (200)**	**100% (217)**	**100% (240)**

Table D.2: Performance of different attacks on the same *LibriSpeech* dataset but across different ASR models including *DeepSpeech (GRU)* and *Wav2Letter*. 200 coordinates are perturbed in one query. The cells are in the form of "Attack Success Rate% (#queries)".

Dataset	Method	Attack Success Rates Under Different Increased Character Error Rates (ICERs)					
		ICER=0.1	ICER=0.2	ICER=0.3	ICER=0.4	ICER=0.5	ICER=0.6
LibriSpeech	FD	**100% (53)**	100% (93)	100% (123)	100% (154)	95% (184)	95% (231)
	ZOO-Adam	100% (71)	100% (112)	100% (143)	100% (171)	100% (194)	100% (218)
	ZOO-Newton	100% (67)	100% (119)	100% (161)	100% (202)	90% (204)	90% (237)
	GA	100% (146)	100% (178)	100% (202)	100% (223)	100% (249)	100% (273)
	T-NES	100% (60)	**100% (83)**	**100% (99)**	**100% (118)**	**100% (137)**	**190% (155)**
TEDLIUM	FD	100% (160)	95% (305)	40% (304)	30% (358)	20% (423)	15% (430)
	ZOO-Adam	100% (116)	100% (181)	100% (226)	100% (251)	100% (270)	100% (285)
	ZOO-Newton	100% (147)	95% (217)	85% (234)	85% (276)	80% (289)	75% (307)
	GA	100% (126)	100% (226)	90% (277)	80% (294)	65% (305)	60% (311)
	T-NES	**100% (113)**	**100% (175)**	**100% (221)**	**100% (232)**	**100% (252)**	**100% (267)**

Table D.3: Performance of different attacks on the same *DeepSpeech2 (LSTM)* model but across different datasets including *LibriSpeech* and *TEDLIUM*. 500 coordinates are perturbed in one query. The cells are in the form of "Attack Success Rate% (#queries)".

Dataset	Method	Attack Success Rates Under Different Increased Character Error Rates (ICERs)					
		ICER=0.1	ICER=0.2	ICER=0.3	ICER=0.4	ICER=0.5	ICER=0.6
DeepSpeech (GRU)	FD	**100% (3)**	100% (5)	100% (8)	100% (17)	100% (38)	100% (64)
	ZOO-Adam	100% (4)	100% (7)	100% (10)	100% (23)	95% (35)	95% (48)
	ZOO-Newton	100% (4)	100% (7)	100% (28)	100% (38)	95% (34)	90% (70)
	GA	100% (126)	100% (128)	100% (132)	100% (140)	19% (144)	95% (154)
	T-NES	100% (3)	**100% (4)**	**100% (9)**	**100% (16)**	**100% (39)**	**95% (50)**
Wav2letter	FD	100% (115)	100% (167)	100% (206)	85% (200)	80% (219)	80% (267)
	ZOO-Adam	100% (70)	100% (97.5)	100% (120)	100% (141)	100% (168)	100% (192)
	ZOO-Newton	100% (74)	100% (119)	100% (163)	100% (210)	95% (229)	85% (238)
	GA	25% (280)	20% (249)	15% (205)	15% (231)	15% (240)	15% (284)
	T-NES	**100% (62)**	**100% (67)**	**100% (73)**	**100% (79)**	**100% (85)**	**100% (95)**

Table D.4: Performance of different attacks on the same Dataset (*Librispeech*) but across different models including *DeepSpeech (GRU)* and *Wav2letter*. 500 coordinates are perturbed in one query. The cells are in the form of "Attack Success Rate% (#queries)".