TOWARDS ROBUST SPEECH-TO-TEXT ADVERSARIAL ATTACK

Mohammad Esmaeilpour, Patrick Cardinal, Alessandro Lameiras Koerich

École de Technologie Supérieure, Université du Québec, Département de Génie Logiciel et des TI
1100 Notre-Dame Ouest, Montréal, H3C 1K3, Québec, Canada

mohammad.esmaeilpour.1@ens.etsmtl.ca, {patrick.cardinal, alessandro.koerich}@etsmtl.ca

ABSTRACT

This paper introduces a novel adversarial algorithm for attacking the advanced speech-to-text transcription systems. Our proposed approach is based on developing an extension for the conventional distortion condition of the general adversarial optimization formulation using the Cramér integral probability metric. Minimizing over such a metric contributes to crafting signals very close to the subspace of legitimate speech recordings. That helps yield more robust adversarial signals against over-the-air playbacks without employing neither costly expectation over transformations nor static room impulse response simulations. Our approach considerably outperforms other targeted and non-targeted algorithms in terms of word error rate and sentence-level accuracy. Furthermore compared to seven other strong white and black-box adversarial attacks, our proposed approach is considerably more resilient against multiple consecutive over-the-air playbacks, corroborating its higher robustness in noisy environments.

Index Terms— Speech-to-text transcription, adversarial attack, Cramér integral probability metric, DeepSpeech, Kaldi, Lingvo.

1. INTRODUCTION

During the last years and especially after the characterization of adversarial attacks for computer vision applications [1], several investigations have been conducted on generalizing this threat to the audio recognition and speech transcription models [2, 3, 4]. It has been proven by Carlini and Wagner [3] that adversarial signals exist for both 1D and 2D representations, which can seriously debase the performance of cutting-edge speech-to-text models such as DeepSpeech [5], Kaldi [6], and Lingvo [7]. However, developing effective adversarial signals resilient to environmental noises and room settings is challenging [8, 9]. These settings include the position and characteristics of both the microphone and speaker and the room's geometry. Under various settings, simply playing the crafted adversarial signal over the air and recording it by another microphone most likely removes the obtained adversarial perturbation [3]. For addressing this issue, several expectation over transformation (EOT) operations have been introduced [4, 10, 11, 12]. These operations often employ room filter sets (e.g., channel impulse response [11]) as part of the adversarial optimization procedure to avoid bypassing the perturbation after playing over the air. However, developing EOT operations is dependent on some static room assumptions, which might negatively affect the generalizability of the filter sets [10, 13]. In a big picture, the optimization formulation toward crafting an adversarial signal for a speech-to-text model has two parts: (i) optimization term and (ii) the distortion condition (relative constraint) [3]:

$$\underbrace{\min_{\delta} \|\delta\|_F + \sum_i c_i \mathcal{L}_i(\vec{x}_{\text{adv}}, \hat{\mathbf{y}}_i)}_{\text{optimization term}} \quad \text{s.t.} \quad \underbrace{l_{\text{dB}}(\vec{x}_{\text{adv}})}_{\text{distortion condition}} < \epsilon \quad (1)$$

where δ is the adversarial perturbation achievable through this iterative procedure for the original input signal $\vec{x}_{\text{org}}$ to yield the adversarial signal $\vec{x}_{\text{adv}}$:

$$\vec{x}_{\text{adv}} \leftarrow \vec{x}_{\text{org}} + \delta \quad (2)$$

where c_i, ϵ, and $\hat{\mathbf{y}}_i$ are a scaling coefficient, an audible threshold, and the targeted incorrect phrase defined by the adversary, respectively. Furthermore, $\mathcal{L}(\cdot)$ denotes a loss function such as the connectionist temporal classification (CTC) loss [14, 3], the psychoacoustic loss function [9], the cross-entropy loss [4], etc. In this typical formulation, the distortion condition is usually known as the loudness metric $l_{\text{dB}}(\cdot)$ computed in the logarithmic dB-scale w.r.t. the human hearing range [3].

The EOT operations incorporated in the state-of-the-art adversarial attack algorithms often involve the optimization term in Eq. 1 [4, 10, 11]. Herein, we discuss extending the distortion condition in this equation to avoid implementing costly EOT-based operations applied on the optimization term, which also helps craft more robust adversarial signals. Toward this end, we review some strong adversarial attack approaches in Section 2. Then, we provide theoretical explanations on developing a relative constraint (the distortion condition) in Section 3. Finally, we analyze the achieved results from the conducted experiments on attacking speech-to-text models in Section 4. In summary, we make the following contributions in this paper: (i) developing an extension for the distortion condition of an adversarial attack formulation using the Cramér integral probability metric; (ii) introducing a white-box attack framework for crafting adversarial signals more robust against over-the-air playbacks; (iii) avoiding time-consuming room impulse response simulations and costly EOT operations in the adversarial optimization formulation (i.e., Eq. 1).

2. BACKGROUND ON ADVERSARIAL ATTACK

This section reviews cutting-edge white and black-box adversarial attacks against speech-to-text models. We focus on the EOT-based attacks since they are, to some extent, capable algorithms in crafting over-the-air resilient adversarial signals [4]. However, we start with the baseline EOT-free C&W attack [3] developed for the DeepSpeech speech-to-text system. This algorithm is based on Eq. 1 and

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grants 2016-04855 and 2016-06628. Supplementary material is available at this github-Repo.

introduces a simple yet effective distortion condition for a targeted attack scenario [3]:

$$l_{\text{dB}}(\vec{x}_{\text{adv}}) = l_{\text{dB}}(\delta) - l_{\text{dB}}(\vec{x}_{\text{org}}) \tag{3}$$

where $l_{\text{dB}}(\cdot)$ can be scaled by a factor of 20 to fit better the human audible range [3]. The C&W attack uses the CTC loss function with an assumption of optimizing $\min_\delta \|\delta\|_2^2$ for the string tokens π_i (without duplication), which eventually should reduce to $\hat{\mathbf{y}}_i$ (after greedy or beam search decoding [3]). Although such a distortion metric constraints the C&W algorithm to craft an adversarial signal almost seamlessly to the original sample $\vec{x}_{\text{org}}$, it does not impose a strict condition to generate an over-the-air resilient adversarial signal. Presumably, this is due to making a reasonable trade-off between adversarial signal quality and attaining small magnitude for the adversarial perturbation δ.

The EOT operation introduced by Qin *et al.* [4] uses an acoustic room simulator followed by speech reverberation filtrations for crafting resilient adversarial signals in adverse scenarios (i.e., multiple over-the-air playbacks). This algorithm is known as the Robust Attack, and it fits in the targeted adversarial category incorporating a variety of room settings for improving its performance. The optimization procedure of this attack subject to $\|\delta\| < \epsilon$ is [4]:

$$\min_\delta \mathbb{E}_{t\sim\tau}\left[\ell_{net}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) + c_i \ell_m(\vec{x}_{\text{org},i}, \delta_i)\right] \tag{4}$$

where τ is an EOT filter set predefined (computed according to the room setting) by the adversary and $\mathbf{y}_i \neq \hat{\mathbf{y}}_i$, where the latter refers to the ground truth phrase associated with $\vec{x}_{\text{org}}$. Moreover, $\ell_{net}(\cdot)$ and $\ell_m(\cdot)$ denote the cross-entropy and the masking threshold loss functions, respectively. The Robust Attack has been tested on the Lingvo speech-to-text system, and it has demonstrated a high capacity for crafting resilient over-the-air adversarial signals.

Yakura *et al.* [8] introduced a similar EOT operation, which employs band-pass filtration according to the human cut-off hearing range on top of the simulated room impulse response (RIR) filter set. Moreover, this attack implements the white Gaussian noise (WGN) filtration to simulate environmental noises effectively as:

$$\min_\delta \mathbb{E}_{t\ \in\tau,\omega\sim\mathcal{N}(0,\sigma^2)}\left[\mathcal{L}(\text{mfcc}(\vec{x}_{\text{adv}}), \hat{\mathbf{y}}_i) + \alpha_k \|\delta\|\right] \tag{5}$$

where:

$$\vec{x}_{\text{adv}} = [\vec{x}_{\text{org}} + \Omega(\delta)] \circledast t + \omega \tag{6}$$

Additionally, ω, mfcc, and α_k denote the WGN filter drawn from a normal distribution with variance σ^2, the Mel-frequency cepstral coefficient transform [15], and a scaling hyperparameter defined by the adversary, respectively. Additionally, $\Omega(\cdot) \in [1, 4]$ kHz refers to the band-pass filtration operation, and $\circledast$ is the convolution operator. Herein, $\mathcal{L}(\cdot)$ stands for the CTC loss function, which was adapted to the DeepSpeech victim model. The reported experiments demonstrated that Yakura's attack outperforms the C&W in various environmental scenes [8] at the cost of higher computational complexity for computing the τ filter set.

One reliable approach that implements the RIR simulation with a relatively lower computational cost is the Imperio attack [10]. This algorithm employs a deep neural network (DNN) to simulate the RIR filter set and the psychoacoustic thresholding (pst) for crafting over-the-air resilient adversarial signals (Eq. 7 [10]).

$$\vec{x}_{\text{adv}} = \underbrace{\arg\max_{\vec{x}_i} \mathbb{E}_{t\sim\tau_d}\left[P(\hat{\mathbf{y}}_i|\vec{x}_{i,t})\right]}_{\vec{x}_{\text{org}} + \kappa\left[\partial \ell_{net}(\mathbf{y},\hat{\mathbf{y}})/\partial f^*(\vec{x}_{\text{org}})\right]} \tag{7}$$

where d, κ, and $f^*(\cdot)$ denote the dimension of the filter set, the learning rate, and the post-activation function of the DNN model mentioned above, respectively. The EOT operation incorporated into the Imperio attack is dynamic and fits well for various settings including meeting, lecture, and office rooms. The distortion condition in this attack is $\delta \leq pst$ and should be tuned for every incorrect phrase $\hat{\mathbf{y}}$. Imperio has been tested on the Kaldi system. Such an attack has considerably reduced this advanced speech-to-text model's performance even after over-the-air playback.

Since the robustness of an over-the-air adversarial signal can also depend on both speaker and microphone characteristics, the channel impulse response (CIR) filter set is developed as part of the EOT operation in the Metamorph adversarial attack [11]. The general formulation of this attack is:

$$\min_\delta \alpha_t l_{\text{dB}}(\vec{x}_{\text{adv}}) + \frac{1}{M}\mathcal{L}(\vec{x}_{\text{org}} + \delta_i, \pi_i) \quad \text{s.t.} \quad \|\delta\| < \epsilon \tag{8}$$

where α_t is the balancing coefficient between the quality of the crafted adversarial signal and the overall success rate of the attack algorithm on the victim model. Additionally, M indicates the number of microphone-speaker positions in an enclosed environment. These hyperparameters have a key role in crafting robust adversarial signals, which the adversary should precisely locate. The effectiveness of the Metamorph adversarial attack has been proven for the DeepSpeech system at the cost of employing various CIR filter sets [11].

Developing EOT operations for the black-box adversarial attack is extremely challenging since the adversary does not have access to the victim model and its associated settings. In response to this limitation, an over-the-line technique has been developed to surrogate the over-the-air EOT operations [12]. However, this technique requires numerous experiments to capture local and global environmental scene distributions. Regarding this concern, there are two EOT-free black-box adversarial attacks with competitive performance to the over-the-line approach in attacking the DeepSpeech system: (i) the genetic algorithm attack (GAA) [16] and (ii) the multi-objective optimization attack (MOOA) [17]. Furthermore, all these algorithms are often used in targeted attack scenarios, as discussed in [13].

3. PROPOSED DISTORTION CONDITION & ADVERSARIAL ATTACK FORMULATION

This section introduces an extension for the distortion condition of the adversarial attack formulation (Eq. 1) for end-to-end speech-to-text systems in targeted and non-targeted scenarios. This condition fits well for the optimization formulation of the white-box adversarial attack scenario. Our motivation for developing such a distortion condition is threefold: improving the robustness of the adversarial speech signals after over-the-air playbacks, avoiding costly EOT operations, and keeping the quality of the crafted adversarial signal as close as possible to the ground truth input signals. We firstly introduce an integral probability metric (IPM) to measure discrepancies between the adversarial and original signals. Then, we build our distortion condition for adversarial attacks based on this IPM. Finally, we explain all the required details in the following subsections.

3.1. Cramér Integral Probability Metric (Cramér-IPM)

One of the standard statistical approaches in measuring the dissimilarity between two probability distributions regardless of the total number of their independent variables is using an IPM [18, 19]. Formally, an IPM is a measure for approximating the discrepancies be-

tween two (generalizable to higher orders) probability density functions $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$ as [18, 20]:

$$\sup_{f\in\mathcal{F}} [\mathbb{E}_{\vec{x}_i\sim\mathbb{P}} f(\vec{x}_i) - \mathbb{E}_{\vec{x}_i\sim\mathbb{Q}} f(\vec{x}_i)] \tag{9}$$

where $f(\cdot)$ is the critic function that analytically compares the dissimilarity between $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$. Furthermore, $\mathcal{F}$ denotes the possible function class for the critic function. It is entirely independent of both probability distributions mentioned above [21]. Mathematically, there are many choices for the function class. However, we opted for Cramér ($\mathcal{F}_{Cr}$) due to its simplicity, differentiability, and generalizability [22, 23]. The statistical definition for $\mathcal{F}_{Cr}$ in the closed-form is [23, 24, 25]:

$$\mathcal{F}_{Cr} = \left\{ f_\vartheta : \mathcal{X} \to \mathbb{R}, \mathbb{E}_{\vec{x}_i\sim\mathbb{P}} \left(D^{(1)} f_\vartheta(\vec{x}_i) \leq 1 \right) \right\} \tag{10}$$

where $D^{(1)}$ indicates the first-order derivation operator and the critic function f_ϑ is smooth with the zero boundary condition [26]. Moreover, $\vec{x}_i \in \mathbb{R}^{n\times m}$ is an m-channel signal with the length n, and $\mathcal{X}$ is a compact subset in $\mathbb{R}$. According to this definition, $\mathcal{F}_{Cr}$ restricts the derivative of $f_\vartheta(\cdot)$ within a unit ball to enforce its continuity for higher degrees of ϑ [23, 25].

Assuming the probability distribution functions for the original and adversarial signals are represented by $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$. Therefore, minimizing over Eq. 9 using the $\mathcal{F}_{Cr}$ reduces dissimilarities between random pairs of $\vec{x}_{\text{adv}}$ and $\vec{x}_{\text{org}}$. However, the convergence of such a minimization procedure is highly dependent on the availability of f_ϑ. One possible approach for finding this critic function could be training a neural network (mainly in the generative model frameworks [23, 27]). Nevertheless, it imposes unnecessary complications and computational overhead to the adversarial optimization formulation. To tackle this issue, we empirically approximate f_ϑ with the joint cumulative distribution function (CDF) [28] of $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$ as:

$$f_{\mathbb{PQ}}(\cdot) \simeq \sum_{i=1}^{n_t} \mathbb{P}(\vec{x}_{i,\text{org}}) + \mu \cdot \mathbb{Q}(\vec{x}_c), \quad \mu \sim \mathcal{U}[-1, 1] \tag{11}$$

where $\vec{x}_c$ is a candidate for the adversarial signal $\vec{x}_{\text{adv}}$ achieved through optimizing for Eq. 1 and eventually $\vec{x}_c \xrightarrow{\mu} \vec{x}_{\text{adv}}$. Furthermore, n_t refers to the total number of original samples and μ is a uniform scaling probability prior to avoid dominating $\mathbb{P}(\vec{x}_{i,org}), \forall i$ over $\mathbb{Q}(\vec{x}_c)$. Using the critic function $f_{\mathbb{PQ}}(\cdot)$ in Eq. 10 provides a meaningful space for measuring discrepancies between original and adversarial distributions (see similar note in [29]). Thus, minimizing over Eq. 9 maps $\vec{x}_c$ onto the original signal manifold and yield a more robust adversarial signal as shown in Section 4.

3.2. Distortion Condition Using the Cramér-IPM

This section introduces the proposed distortion condition based on the Cramér-IPM with the critic function $f_{\mathbb{PQ}}(\cdot)$. In fact, we extend the relative constraint of Eq. 1 to:

$$\min_{\delta, f_{\mathbb{PQ}}\in\mathcal{F}_{Cr}} \|\mathbb{E}_{\vec{x}_i\sim\mathbb{P}} f_{\mathbb{PQ}}(\vec{x}_i) - \mathbb{E}_{\vec{x}_c\sim\mathbb{Q}} f_{\mathbb{PQ}}(\vec{x}_c)\| \tag{12}$$

where:

$$l_{\text{dB}}(\vec{x}_c) < \epsilon \quad \text{and} \quad \vec{x}_{\text{adv}} = \arg\min \vec{x}_c \tag{13}$$

The intuition behind exploiting this condition is finding the best signal $\vec{x}_c$, which sounds similar to $\vec{x}_{\text{org}}$ according to the loudness metric $l_{\text{dB}}(\cdot)$ and lies closer to the original signal manifold. Since $\mathbb{E}_{\vec{x}_i\sim\mathbb{P}} f_{\mathbb{PQ}}(\vec{x}_i)$ incorporates the CDF of original and adversarial signals containing background and room noises, it implicitly learns the impulse responses available in the speech dataset. This also possibly makes bypassing δ very challenging after over-the-air playbacks. From a statistical perspective, the proposed distortion condition forces an attack optimization formulation to craft an adversarial signal marginally close to the original signals' distribution. This is for counteracting with adversarial defense algorithms, which measure the distance between distribution manifolds to detect adversarial signals [13]. These defense approaches are inspired by Ma *et al.* [30], which proves that the subspace of adversarial signals is distinct from original and noisy samples [2]. In other words, it is possible to measure the distance between subspaces using metrics defined in orthogonal decomposition forms (e.g., chordal distance in Schur decomposition space [2].) Based on this finding, variants of defense algorithms have been developed and they have shown a outstanding performance against strong white and black-box adversarial attacks [13]. Therefore, incorporating our proposed distortion condition into the attack optimization formulation (i.e., Eq. 1) helps to yield a more robust adversarial signal.

Algorithm 1 Robust adversarial attack with distortion condition using the Cramér-IPM. This attack is primarily a targeted attack. However it is conveniently generalizable to a non-targeted framework via randomly selecting a wrong output phrase ($\hat{\mathbf{y}}_i$) for any given input signal.

Require: $\vec{x}_{\text{org}}, \mathbf{y}, \hat{\mathbf{y}}, \epsilon$ ▷ input signal, original phrase, incorrect target phrase, hearing threshold
Ensure: $\vec{x}_{\text{adv}}$ ▷ adversarial speech signal
$\vec{x}_c \leftarrow \vec{x}_{\text{org}}$ ▷ initializing
initialize μ ▷ random latent variable
while $\hat{\mathbf{y}} = \mathbf{y}$ **do** ▷ the goal is reaching to $\hat{\mathbf{y}} \neq \mathbf{y}$
  $\delta \leftarrow \min_\delta \|\delta\|_F + \sum_i c_i \mathcal{L}_i(\vec{x}_c, \hat{\mathbf{y}}_i)$
  $\vec{x}_c \leftarrow \vec{x}_c + \delta$ ▷ candidate adversarial signal
  while $l_{\text{dB}}(\vec{x}_c) > \epsilon$ **do** ▷ up to reach $l_{\text{dB}}(\vec{x}_c) < \epsilon$
    draw a random $\mu \sim \mathcal{U}[-1, 1]$
    $\delta \leftarrow \min_{\delta, f_{\mathbb{PQ}}} \|\mathbb{E}_{\vec{x}_i\sim\mathbb{P}} f_{\mathbb{PQ}}(\vec{x}_i) - \mathbb{E}_{\vec{x}_c\sim\mathbb{Q}} f_{\mathbb{PQ}}(\vec{x}_c)\|$
    $\vec{x}_c \leftarrow \vec{x}_c + \delta$ ▷ update the candidate signal
$\vec{x}_{\text{adv}} \leftarrow \vec{x}_c$ ▷ crafted adversarial signal

The general overview of our proposed attack algorithm is shown in Algorithm 1. Regarding this pseudocode, we do not employ any EOT operations in our optimization formulation since Eq. 12 implicitly captures local and global distributions of the signals available in the comprehensive speech datasets.

4. EXPERIMENTS

We have implemented Algorithm 1 to attack DeepSpeech (Mozilla's standard implementation), Kaldi, and Lingvo speech-to-text models. Although the proposed algorithm resembles a targeted adversarial attack and requires defining an incorrect target phrase ($\hat{\mathbf{y}}_i$), it is generalizable to the non-targeted scenarios with the assumption of choosing a random phrase for $\hat{\mathbf{y}}_i$ other than the ground-truth ($\mathbf{y}_i$). Regarding the standard practice in evaluating adversarial attacks that craft adversarial signals only for a portion of the given speech datasets [3, 4, 10, 11, 13], we have also randomly selected 1000 samples from Mozilla common voice (MCV) [31] and LibriSpeech [32] to evaluate the performance of the proposed attack. These two comprehensive datasets contain utterances from different genders, accents, and ages in short and long speech recordings. We

Table 1. Performance comparison of adversarial attack algorithms both in white and black box as well as targeted and non-targeted scenarios. A targeted attack aims at a specific $\hat{\mathbf{y}}$, while a non-targeted attack aims at any incorrect phase. n_{ota} stands for the total rounds of robustness against consecutive over-the-air playbacks. This is conducted via various simulations over different room settings with multiple speaker-microphone configurations [8]. These results are averaged over 10 different experiments in order to avoid bias in evaluations.

Model	Attack	WER (%)	SLA (%)	segSNR	STOI	LLR	Type	EOT	n_{ota}
DeepSpeech	C&W [3]	78.94 ± 2.01	30.74 ± 3.16	$\mathbf{21.34}$	0.86	0.35	T	–	0
	Yakura's attack [8]	80.28 ± 3.14	35.49 ± 0.28	19.57	0.82	0.38	T	✓	3
	Metamorph [11]	72.48 ± 1.06	45.84 ± 4.71	17.66	0.84	0.36	T	✓	1
	GAA [16]	65.80 ± 2.55	48.35 ± 3.38	17.02	0.79	0.31	T	–	1
	MOOA [17]	68.06 ± 2.71	47.01 ± 1.42	18.46	0.81	0.42	T or NT	–	1
	Proposed	$\mathbf{88.19 \pm 3.15}$	$\mathbf{21.69 \pm 3.09}$	18.88	$\mathbf{0.88}$	$\mathbf{0.29}$	T or NT	–	**4**
Kaldi	Imperio [10]	69.34 ± 0.47	31.49 ± 1.36	$\mathbf{24.71}$	0.91	0.28	T	✓	2
	Proposed	$\mathbf{83.51 \pm 1.44}$	$\mathbf{25.86 \pm 1.94}$	23.16	$\mathbf{0.93}$	$\mathbf{0.27}$	T or NT	–	**3**
Lingvo	Robust Attack [4]	84.37 ± 2.07	28.21 ± 2.31	19.44	$\mathbf{0.85}$	$\mathbf{0.41}$	T	✓	3
	Proposed	$\mathbf{89.73 \pm 1.75}$	$\mathbf{22.78 \pm 2.62}$	$\mathbf{21.58}$	0.82	0.43	T or NT	–	**5**

Recall: WER: word error rate. SLA: sentence level accuracy. segSNR: segmental signal to noise ratio. LLR: log-likelihood ratio. STOI: short-term objective intelligibility. EOT: expectation over transformation. T: targeted attack. NT: non-targeted attack.

assign ten incorrect targeted and non-targeted phrases ($\hat{\mathbf{y}}_i$) toward crafting $\vec{x}_{\text{adv}}$ for every selected signal $\vec{x}_{\text{org}}$.

Since the implementations of the benchmarking speech-to-text models are different, for attacking these systems, we use the CTC loss function ($\mathcal{L}(\cdot)$) for DeepSpeech and the cross-entropy loss with masking threshold ($\ell_{net}(\cdot)$, $\ell_m(\cdot)$) for the Lingvo and Kaldi systems as described in [4] and [10]. The rest of the settings, such as simulated speaker-microphone positions, room dimensions, the definition of ϵ, and beam search decoding for output phrases, follow the instructions described in [3]. We make the same assumptions in all experiments for a fair comparison to the Robust Attack, Yakura's attack, Imperio, GAA, MOOA, and Metamorph. We implement all the attack algorithms on two machines with four Nvidia GTX-1080-Ti and two 64-bit Intel Core-i7-7700 (3.6 GHz, Gen. 10) processors with 8×11 GB and 2×64 GB memory, respectively.

We compare the adversarial attack algorithms' performance from two points of view: (i) attack success rate and (ii) adversarial signal quality. For addressing the first view, we measure the word error rate (WER) and sentence level accuracy (SLA) metrics as they have been characterized for such an aim [4, 33]:

$$\text{WER} = \frac{(S+D+I)}{N} \times 100 \quad \text{and} \quad \text{SLA} = \frac{n_c}{n_{tot}} \times 100 \quad (14)$$

where S, D, I, and N denote the total number of substitutions, deletions, insertions, and reference phrases ($\mathbf{y}_i$), respectively. Furthermore, n_c indicates the number of crafted adversarial signals attaining the correct phrase after passing through the transcription system (the victim speech-to-text model). Herein, the total number of phrases are denoted by n_{tot}.

For addressing the second view, we use three quality metrics: segmental signal to noise ratio (segSNR) [34], short-term objective intelligibility (STOI) [35], and log-likelihood ratio (LLR) [34]. The first two metrics compute the absolute quality of the crafted adversarial signals relative to the available ground-truth speeches ($\vec{x}_{\text{org}}$). These two metrics measure how natural the crafted $\vec{x}_{\text{adv}}$ sounds relative to $\vec{x}_{\text{org}}$. Higher values for segSNR and STOI metrics interpret as the closer quality of $\vec{x}_{\text{adv}}$ to the original signals. Since these two metrics are not necessarily bounded, comparing adversarial signals' quality may not be tangible enough. We use the LLR, which ranges between zero and one, in response to this potential concern. Furthermore, there is an inverse relationship between the magnitude of this metric and the quality of the signals. In other words, for adversarial signals close to their associated $\vec{x}_{\text{org}}$, the LLR is relatively low.

Table 1 summarizes our achieved results. As shown in this table, our proposed attack algorithm outperforms the other algorithms regarding WER and SLA metrics. However, it partially fails against C&W, Imperio, and the Robust Attack regarding the quality of the crafted adversarial signals. Table 1 also demonstrates that the proposed attack algorithm's robustness is higher than others after multiple consecutive over-the-air playbacks (n_{ota}). Since playing adversarial signals over the air often results in losing adversarial perturbation [3], higher n_{ota} indicates better robustness.

Regarding the computational cost in runtime, the EOT-based attacks, namely Yakura's, Metamorph, Imperio, and the Robust Attack, are more costly than the proposed algorithm with the relative ratio of 2.23, 1.97, 2.07, and 2.49, respectively. These values are averaged over ten different experiments on the three benchmarking victim speech-to-text models. Further discussion about the computational cost is available in the supplementary material.

5. CONCLUSION

This paper introduced a new adversarial algorithm for effectively attacking the cutting-edge DeepSpeech, Kaldi, and Lingvo speech-to-text systems. Our proposed approach incorporates a novel extension for the relative constraint of the adversarial optimization formulation to improve the crafted signals' robustness after multiple over-the-air playbacks. This extension minimizes the Cramér-IPM between the probability distributions of the original and adversarial signals. This minimization operation projects a candidate adversarial signal onto the original speech recordings' subspace to counteract potential defense approaches that measure the distance between subspaces. We experimentally demonstrated that the proposed white-box attack algorithm outperforms other advanced algorithms in terms of attack success rate according to WER and SLA metrics.

Moreover, the crafted adversarial signals' average quality via our proposed attack is competitive to other algorithms using objective quality metrics of segSNR, STOI, and LLR. Our approach is EOT-free, and it has shown considerably higher robustness against consecutive over-the-air playbacks than other costly EOT-based adversarial algorithms. However, we could not achieve more than four playbacks averaged over the three victim models. We are determined to address this issue in our future works by developing more constraints on the critic function of the Cramér function class such as imposing marginal distribution operators.

6. REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd Intl Conf Learn Repres*, 2014.

[2] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Detection of adversarial attacks and characterization of adversarial subspace," in *IEEE Intl Conf Acoust, Speech and Signal Process*, 2020, pp. 3097–3101.

[3] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Secur Privacy Workss*, 2018, pp. 1–7.

[4] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Intl Conf Mach Learn*, 2019, pp. 5231–5240.

[5] Mozilla Implementation, "Mozilla. project deepspeech," https://github.com/mozilla/DeepSpeech, 2017.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE Works Autom Speech Recog Underst*, 2011.

[7] J. Shen, P. Nguyen, Y. Wu, Z. Chen, M. X. Chen, Y. Jia, A. Kannan, T. Sainath, Y. Cao, C.-C. Chiu, et al., "Lingvo: a modular and scalable framework for sequence-to-sequence modeling," *arXiv preprint arXiv:1902.08295*, 2019.

[8] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," in *28th Intl J Conf Artif Intell*, 2018, pp. 5334–5341.

[9] J. Szurley and J. Z. Kolter, "Perceptual based adversarial audio attacks," *arXiv preprint arXiv:1906.06355*, 2019.

[10] L. Schönherr, T. Eisenhofer, S. Zeiler, T. Holz, and D. Kolossa, "Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems," in *Annual Comp Secur Appl Conf*, 2020, pp. 843–855.

[11] T. Chen, L. Shangguan, Z. Li, and K. Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *Netw Distrib Syst Secur Symp*, 2020.

[12] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. B. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," *arXiv preprint arXiv:1904.05734*, 2019.

[13] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Class-conditional defense GAN against end-to-end speech attacks," in *IEEE Intl Conf Acoust, Speech and Signal Process*, 2021, pp. 2565–2569.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *23rd Intl Conf Mach Learn*, 2006, pp. 369–376.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans Acoust, Speech, Signal Process*, vol. 28, no. 4, pp. 357–366, 1980.

[16] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," in *IEEE Security and Privacy Works*, 2019, pp. 15–20.

[17] Shreya Khare, Rahul Aralikatte, and Senthil Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," *Interspeech Conf., Graz, Austria, 15-19 September 2019*, pp. 3208–3212.

[18] A. Müller, "Integral probability metrics and their generating classes of functions," *Adv Appl Probability*, pp. 429–443, 1997.

[19] Y. Dodge and D. Commenges, *The Oxford dictionary of statistical terms*, Oxford University Press on Demand, 2006.

[20] J. Dedecker and F. Merlevède, "The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in l_p," *ESAIM: Probability and Statistics*, vol. 11, pp. 102–114, 2007.

[21] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. Lanckriet, et al., "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

[22] Gábor J Székely, "E-statistics: The energy of statistical samples," *Bowling Green State Univ, Dept Mathematics and Statistics Tech Rep*, vol. 3, no. 05, pp. 1–18, 2003.

[23] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The cramer distance as a solution to biased wasserstein gradients," *CoRR*, vol. abs/1705.10743, 2017.

[24] M. L. Rizzo and G. J. Székely, "Energy distance," *Wiley Interdisc Reviews: Comput Stat*, vol. 8, no. 1, pp. 27–38, 2016.

[25] Harald Cramér, "On the composition of elementary errors: First paper: Mathematical deductions," *Scandinavian Actuarial Journal*, vol. 1928, no. 1, pp. 13–74, 1928.

[26] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," *Journal Statis Plann Inference*, vol. 143, no. 8, pp. 1249–1272, 2013.

[27] T. Salimans, H. Zhang, A. Radford, and D. N. Metaxas, "Improving gans using optimal transport," in *6th Intl Conf Learn Repres*.

[28] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*, Cambridge University Press, 2020.

[29] Y. Mroueh, C. L. Li, T. Sercu, A. Raj, and Y. Cheng, "Sobolev GAN," in *6th Intl Conf Learn Repres*, 2018.

[30] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," in *6th Intl Conf Learn Repres*, 2018.

[31] Mozilla: commonvoice.mozilla.org, "Mozilla common voice dataset," https://voice.mozilla.org/en/datasets, 2019.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE Intl Conf Acoust, Speech and Signal Process*, 2015, pp. 5206–5210.

[33] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data," in *Intl Conf Rec Adv Nat Lang Process*, 2013, pp. 198–206.

[34] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2019, pp. 106–110.

[35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans Audio, Speech, Lang Process*, vol. 19, no. 7, pp. 2125–2136, 2011.