Multi-Discriminator Sobolev Defense-GAN Against Adversarial Attacks for End-to-End Speech Systems

Mohammad Esmaeilpour, Student Member, IEEE, Patrick Cardinal, Member, IEEE, and Alessandro Lameiras Koerich, Member, IEEE

Abstract—This paper introduces a defense approach against end-to-end adversarial attacks developed for cutting-edge speechto-text systems. The proposed defense algorithm has four major steps. First, we represent speech signals with 2D spectrograms using the short-time Fourier transform. Second, we iteratively find a safe vector using a spectrogram subspace projection operation. This operation minimizes the chordal distance adjustment between spectrograms with an additional regularization term. Third, we synthesize a spectrogram with such a safe vector using a novel GAN architecture trained with Sobolev integral probability metric. To improve the model's performance in terms of stability and the total number of learned modes, we impose an additional constraint on the generator network. Finally, we reconstruct the signal from the synthesized spectrogram and the Griffin-Lim phase approximation technique. We evaluate the proposed defense approach against six strong white and blackbox adversarial attacks benchmarked on DeepSpeech, Kaldi, and Lingvo models. Our experimental results show that our algorithm outperforms other state-of-the-art defense algorithms both in terms of accuracy and signal quality.

Index Terms—Speech adversarial attack, spectrogram, short time Fourier transform, generative adversarial networks, Sobolev integral probability metric, Schur decomposition, chordal distance, adversarial defense.

I. INTRODUCTION

THERE is a large volume of publications on applying deep learning algorithms for audio and speech classification (i.e., transcription), which report high recognition accuracy [1]–[3]. During the last decade, the primary focus has been designing new architectures, for instance, variants of convolution [4], recurrent [5], and attention configurations [6] to improve classification accuracy and model generalizability. However, it has been proven that these advanced models might undergo extreme vulnerability against carefully crafted adversarial signals both in 1D and 2D representation (spectrogram) domains [7], [8].

The major focus of this paper is in response to this vulnerability issue. We have developed an adversarial defense approach against varieties of end-to-end speech-to-text attack algorithms. Toward this end, we firstly review the state-ofthe-art adversarial attacks in Section II. We also provide details about the background of the defense approaches in Section III. Section IV introduces the proposed adversarial defense algorithm followed by comprehensive experimental results in Section V. In summary, we make the following contributions in this paper:

- (i) introducing a novel adversarial defense approach based on a multi-discriminator generative adversarial network (GAN) in the restricted Sobolev space [9];
- (ii) establishing simple yet effective architectures for both the generator and discriminator networks;
- (iii) developing an adjusted chordal distance with a complementary regularization term toward achieving a safe input vector for the generator model;
- (iv) characterizing a constraining technique for improving the stability of our generative model in adverse environmental scenarios;
- (v) experimentally proving the effectiveness of the proposed defense approach for white and black-box as well as targeted and non-targeted attack scenarios.

II. BACKGROUND: ADVERSARIAL ATTACK

An adversarial signal \vec{x}_{adv} carries inaudible perturbation δ , and it can fool the victim classifier (the transcription model) toward any target phrase $\hat{\mathbf{y}}$ defined by the adversary [7]. The actual value of δ is dependent on the length of $\hat{\mathbf{y}}$ (the number of characters) and the characteristics of the original carrier signal \vec{x}_{org} ($\vec{x}_{adv} = \vec{x}_{org} + \delta$) [7], [10]. For measuring the loudness (distortion) of this perturbation relative to the carrier signal, a logarithmic-scale metric has been proposed by Carlini and Wagner [7]:

$$l_{dB}(\vec{x}_{adv}) = l_{dB}(\delta) - l_{dB}(\vec{x}_{org}) \tag{1}$$

where $l(\cdot)$ denotes the loudness of the original 1D signal $\vec{x}_{org} \in \mathbb{R}^{n \times m}$ in dB, and *n* and *m* denote the length and number of channels, respectively. For $l_{dB}(\vec{x}_{adv}) < \epsilon$ where ϵ is a small threshold, \vec{x}_{adv} sounds almost seamless to \vec{x}_{org} according to the C&W attack for the speech-to-text model [7]:

$$\min |\delta|_2^2 + \sum_i c_i \mathcal{L}_i(\vec{x}_{org,i} + \delta_i, \pi_i) \quad \text{s.t.} \quad l_{dB}(\vec{x}_{adv}) < \epsilon$$
(2)

where c_i is a scaling coefficient for the connectionist temporal classification loss function $\mathcal{L}(\cdot)$ [11]. Additionally, π_i denotes string tokens without duplication, which should reduce to the character alignments $\hat{\mathbf{y}}_i$ ($\hat{\mathbf{y}}_i \neq \mathbf{y}_i$, where the latter refers to the ground truth character alignment) [7]. The C&W attack has been primarily developed for the speech-to-text DeepSpeech model [3], and the experiments have shown a complete collapse of this victim model against adversarial signals crafted through Eq. 2 [7].

M. Esmaeilpour, P. Cardinal, and A. L. Koerich are with the Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS), Université du Québec, Montréal, QC, Canada, e-mail: (mo-hammad.esmaeilpour.1@ens.etsmtl.ca, patrick.cardinal@etsmtl.ca, alessan-dro.koerich@etsmtl.ca).

The C&W attack splits the input signal into 50 frames per second, and it eventually yields a universal perturbation for the entire chunks in \vec{x}_{org} . This operation reduces the computational complexity of the attack algorithm compared to optimizing fine-grained δ_i for every chunk. However, it might negatively affect the robustness of \vec{x}_{adv} in a real-world environment. In other words, playing these speech chunks over the air and recording them by another microphone, involving environmental reverberating and signal echo, might easily remove the adversarial effect (δ) [12]. Several algorithms crafting more resilient adversarial signals in natural environments have been proposed in response to this issue. These algorithms are based on psychoacoustic loss function [13], feature vector analysis [14], and employing a set of filters (band-pass, impulse response, and white Gaussian noise) [15]. However, these approaches have been evaluated within static environments with predefined room setups, which might reduce these algorithms' generalizability in more challenging scenarios [12]. Inspired by Athalye et al. [16], which introduces the expectation over transformation (EOT) to the attack optimization formulation for regularizing the cost function (similar to Eq. 2), many other EOT variants have been proposed for the speech domain [10], [12], [17]. These regularizations help craft more robust adversarial signals for non-static environments, which fit in both white and black-box attack scenarios.

The EOT proposed by Qin *et al.* [10] is based on an acoustic room simulator, which generates artificial utterances and environmental reverberations. This algorithm is known as Robust Attack and encodes the EOT regularization into the loss function of a speech-to-text model as [10]:

$$\ell(\vec{x}_{org,i}, \delta_i, \mathbf{y}_i) = \mathbb{E}_{t \sim \tau} \left[\ell_{net} \left(\mathbf{y}_i, \hat{\mathbf{y}}_i \right) + \alpha \ell_m(\vec{x}_{org,i}, \delta_i) \right]$$
(3)

where α is a static scaling factor, $\ell_{net}(\cdot)$ denotes the cross entropy loss and $\ell_m(\cdot)$ indicates the loss function for masking threshold (ϵ). In fact, $\ell_m(\cdot)$ constrains over the normalized power spectral density function of \vec{x}_{org} and contributes to the imperceptibility of the adversarial signal [10]. Additionally, τ refers to the transformation set including room reverberation settings. This attack has been tested on the Lingvo speech-totext system [2] and could achieve a very high fooling rate on this advanced system.

The Imperio attack proposes another variant of EOT, which implements simulated room impulse response (RIR) filters, taking advantage of a simple deep neural network (DNN) architecture [12]. Additionally, this attack embeds psychoacoustic thresholding for reducing adversarial distortion similar to Qin *et al.* [10] (see Eq. 4 in [12]).

$$\vec{x}_{adv} = \arg\max_{\vec{x}_i} \mathbb{E}_{h \sim H_{\text{dim}}} \left[P(\hat{\mathbf{y}}_i | \vec{x}_{i,h}) \right]$$
(4)

where $h \in H_{\text{dim}}$ denotes a RIR filter and dim indicates the dimension of the filter set. The Imperio is an iterative algorithm and minimizes the adversarial perturbation δ via approximating the $\nabla_{\vec{x}_{org}} = \partial \ell_{net}(\mathbf{y}, \hat{\mathbf{y}}) / \partial f^*(\vec{x}_{org})$ where $f^*(\cdot)$ denotes the post activation function. In each iteration and according to the distribution of H_{dim} , an adversarial candidate $\vec{x}_{adv} = \vec{x}_i + \kappa \nabla_{\vec{x}_i}$ with the learning rate κ should satisfy $\hat{\mathbf{y}}_i \neq \mathbf{y}_i$. This procedure continues up to reach the predefined audible threshold ϵ . This attack was evaluated on the Kaldi speech-to-text system [18], which employs both DNN and hidden Markov model (HMM) configurations for real-time speech transcription. It has been shown that under various environmental settings, including lecture, meeting, and office rooms, the Imperio attack has considerably turned down the transcription performance of the Kaldi system [12].

The EOT regularization in the Metamorph adversarial attack [17] is similar to the RIR filtration in the Imperio algorithm with one major difference: it implements channel impulse response (CIR) to characterize potential over the air distortions on δ . This attack algorithm employs M pairs of microphone-speaker transmission in different distances (similar to H_{dim}) to encompass a wide range of reverberations in yielding minimal perturbation:

$$\arg\min_{\delta} \alpha_t l_{dB}(\vec{x}_{adv}) + \frac{1}{M} \mathcal{L}(\vec{x}_{org} + \delta_i, \pi_i)$$
(5)

where α_t denotes a trade-off scalar between the fooling rate of the model and the signal quality. Similar to the C&W attack, the Metamorph attack was evaluated on the DeepSpeech victim model. The experiments showed an attack success rate of around 90% and low Mel-cepstral distortion for this whitebox algorithm [17].

Since integrating the EOT regularization into the adversarial optimization problem requires access to the victim model's cost function, the black-box attacks can not directly incorporate it into their formulations. For addressing this issue, a surrogate technique has been proposed and called the overthe-line approach, which provides multiple varieties of the adversarial signals to the victim model before playbacks over the air [14]. This operation helps the adversary to capture the environmental scene distribution without directly simulate it through reverberation filters. However, the performance of this approach is directly dependent on the comprehensiveness of the over-the-line adversarial signals. More straightforward yet effective black-box adversarial attacks, which do not incorporate EOT regularization with competitive performance on the DeepSpeech system, are the genetic algorithm attack (GAA) [19] and multi-objective optimization attack (MOOA) [20]. These algorithms were tested for targeted and non-targeted attacks and achieved high fooling rates.

While all the aforementioned adversarial attacks pose major security concerns against cutting-edge speech-to-text models, namely DeepSpeech, Kaldi, and Lingvo, there are few investigations on defense algorithms. The following section reviews the state-of-the-art defense approaches developed for counteracting white and black-box adversarial attacks.

III. BACKGROUND: ADVERSARIAL DEFENSE

Developing defense approaches against robust adversarial attack algorithms can be very challenging due to several reasons. Firstly, standard speech signals have high dimensionality (e.g., 8 kHz), and even running effective compression techniques [21] for potentially discarding adversarial perturbations can be time-consuming in real-time speech-to-text transcription. Secondly, speech signals often have various channels for quality enhancement purposes [22]. Thus an adversary

can optimize δ for such channel(s), which human auditory systems are less sensitive to them and more effectively fool the victim model [23]. Thirdly, usually, speech signals carry environmental and microphone-speaker noises, which makes distinguishing a noisy signal from an adversarial very difficult even after band-pass filtering [24]. In the following, we briefly review a couple of multiscale approaches that have been able to tackle these challenges to some extent.

Inspired by Das *et al.* [25], a compression-based approach has been introduced for removing the potential adversarial perturbation on the speech signals [21]. This algorithm implements both adaptive multi-rate and MPEG audio layer-3 encoding for such an aim. Reported results showed the effectiveness of this approach in adverse scenarios for shortlength signals [21]. Furthermore, for sophisticated adversarial signals, which have been precisely optimized through running the Robust Attack [10], this defense scheme failed to remove adversarial perturbations [26].

Autoencoder-based defense GAN (A-GAN) [27] is structurally similar to the compression approach mentioned above. Instead of low-level signal filtering, it implements high-level feature transformation. The intuition behind this approach is transforming the signal into a similar recording using an autoencoder. The proposed autoencoder implements a complex architecture for reconstructing feature vectors to remove potential adversarial perturbation δ . Extensive experiments of A-GAN on DeepSpeech and Lingvo systems have been reported by Esmaeilpour *et al.* [26].

Since it has been proven that adversarial subspace is distinct from original and noisy signals [28], a defense GAN based on this fact has been developed by Esmaeilpour *et al.* [26]. Unlike the compression approach and A-GAN approaches, this defense algorithm employs neither low nor high-level transformations for discarding adversarial perturbations directly on the signal. Instead, it uses a class-conditional GAN for computing a refined latent variable z_i for the generator network via:

$$\nabla_{\mathbf{z}_i} \|\gamma [G(\mathbf{z}_i), \mathbf{x}_i]\|_2^2 \tag{6}$$

where $\mathbf{z}_i \in \mathbb{R}^{d_z}$ with dimension d_z is the random variable from $p_z \sim \mathcal{N}(0, 0.4I)$ and $G(\cdot)$ with distribution p_g denotes the generator network. Additionally, $\gamma[\cdot]$ is the chordal distant adjustment function between the input spectrogram \mathbf{x}_i and $G(\mathbf{z}_i)$. Eq. 6 is iterative and finds the optimal latent variable \mathbf{z}_i^* , which not only forces $G(\mathbf{z}_i^*)$ to lie in the original signal subspace, but also generate a spectrogram very similar to \mathbf{x}_i .

The effectiveness of this class-conditional defense GAN (CC-DGAN) has been evaluated against the C&W attack, the Robust Attack, and the GAA for both DeepSpeech and Lingvo systems [26]. However, it might fail for long-length signals (above six seconds) due to the generator network's instability in around 10k iterations. For addressing this issue, we propose two techniques: (i) introducing a multi-discriminator GAN to provide more informative gradients to the generator network; (ii) implementing such a GAN in the restricted Sobolev space [9] and training the generator network according to the Sobolev function class with a bounded dominant measure. Since a special case of this restricted space is proportional to the 2D Fourier transform representation (spectrogram) [9], we

can train our generative model in a much lower dimensionality compared to 1D speech signals. In the following section, we explain these steps as part of the proposed defense scheme.

IV. PROPOSED ADVERSARIAL DEFENSE METHOD: SOBOLEV DEFENSE GAN (SOBOLEV-DGAN)

The proposed adversarial defense approach against speech attacks has four steps, as depicted in Fig. 1: (i) signal representation (conversion from 1D vector to 2D matrix) using short-time Fourier transform (STFT) [29]; (ii) chordal distance adjustment with a complementary regularization term for projecting the given input spectrogram onto the original subspace (the process shown in the green color); (iii) spectrogram synthesis using a Sobolev GAN and an optimal safe vector \mathbf{z}_i^* (yellow block in Fig. 1); (iv) inverse STFT (i-STFT) for reconstructing the speech signal.

A. Spectrogram: 2D Representation of 1D Speech Signal

There are several standard transformations in the audio and speech processing domains for representing a signal into a 2D spectrogram, such as continuous or discrete wavelet transform, Mel-frequency cepstral coefficients, and STFT. All these transformations have some advantages over each other, and they have been widely used for unsupervised, weakly supervised, and supervised learning tasks. Moreover, the highest recognition accuracies have been often reported for the models trained on these representations over 1D signals [3], [31]. This is presumably due to the lower dimensionality of spectrograms and the inherent ability of these transformations in extracting more distinctive learning features compared to 1D signals [32].

This paper uses the STFT to generate spectrograms from the given speech signals since it is more closely related to the Sobolev integral probability metric (IPM) [33], which we employ to train our generator network. This metric correlates well with the Fourier coefficients encoded in the STFT spectrograms and likely helps extract more distinctive features. The theoretical approach for crafting an STFT spectrogram is as follows.

For a given discrete signal a[n] with length n (sampled from a 1D speech signal \vec{x} in the time domain), we can define the Fourier transform using a Hann function $A[\cdot]$ as [29]:

STFT
$$\{a[n]\}(k,\omega) = \sum_{n=-\infty}^{\infty} a[n]A[n-k]e^{-j\omega n}$$
 (7)

where k is the shifting scale $(k \ll n)$ and ω indicates the frequency coefficients. For capturing more features from a[n], this operation applies on the overlapping signal chunks (i.e., 50 ms) according to a predefined sampling rate (e.g., 16 kHz). For generating the spectrogram, we need to compute the power spectrum of Eq. 7 as:

$$\operatorname{SP}_{\operatorname{STFT}}\left\{a[n]\right\}(k,\omega) = \left|\sum_{n=-\infty}^{\infty} a[n]A[n-k]e^{-j\omega n}\right|^2 \quad (8)$$

where it generates a 2D matrix for a given speech signal \vec{x}_i . In the next subsection, we explain the second step of the proposed defense approach, which finds a refined \mathbf{z}_i^* from the combination of a random $\mathbf{z}_i \in \mathbb{R}^{d_z}$ and the original input spectrogram (\mathbf{x}_i) .



Fig. 1: An overview of the proposed defense GAN approach. The 1D speech signal (\vec{x}_i) is converted to a STFT spectrogram (\mathbf{x}_i) . Moreover, $\gamma[\cdot]$ denotes the chordal distance adjustment required for making \mathbf{x}_i in the same subspace of the synthesized spectrogram $G(\mathbf{z}_i)$ ($\mathbf{z}_i \in \mathbb{R}^{d_z}$ is the latent random variable). The output speech signal (\vec{x}_i) is reconstructed using the i-STFT operation and the Griffin-Lim phase approximation approach [30]. Additionally, rank(\mathbf{x}_i) refers to the input spectrogram's rank according to its eigenvalues computed in the Schur decomposition domain.

B. Chordal Distance Adjustment for Spectrogram Projection

Generally, there are two categories in developing defense approaches against adversarial attacks: running low or highlevel transformations for filtering the input signal aiming at discarding potential adversarial perturbation (as discussed in Section III); synthesizing a very similar signal to a given input vector without running any filtration operation [26], [34]. While most of the introduced algorithms fall into the first category, they are often less reliable since they obfuscate gradient vectors [35]. However, developing a synthesis-based defense algorithm is more challenging since it requires two key steps — a projection of the input space and a stable generative model. Since the proposed defense approach fits the second category, therefore we introduce novel techniques for these steps.

The main goal in this step is finding a safe $\mathbf{z}_i^* \in \mathbb{R}^{d_z}$ for the generator network according to two main conditions: $G(\mathbf{z}_i^*)$ should lie in the subspace of the original signal distribution represented by p_r (approximated by p_g); the synthesized spectrogram $G(\mathbf{z}_i^*)$ should be very similar to the spectrogram of the given 1D speech signal (\mathbf{x}_i) using the ℓ_2 distance metric. Toward this end, for every input spectrogram \mathbf{x}_i , we solve an optimization problem searching all possible $\mathbf{z}_i \in \mathbb{R}^{d_z}$ to find the \mathbf{z}_i^* that meets the conditions above. Fig. 2 shows an overview of this operation.

Inspired by Xingjun *et al.* [37], which proved that adversarial examples lie in distinct subspaces from original and noisy input samples, the chordal distance metric has been introduced for measuring interspaces among spectrogram manifolds [28]. This metric, defined in the Schur decomposition domain for the triplet of original, noisy, and adversarial spectrograms, can be written as [36]:

$$\operatorname{chord}(\lambda[G(\mathbf{z}_{i})], \lambda[\mathbf{x}_{i}]) \leq \frac{\epsilon}{\sqrt{\left[\left(\Phi^{\mathcal{H}}G(\mathbf{z}_{i})\Gamma\right) + \left(\Phi^{\mathcal{H}}\mathbf{x}_{i}\Gamma\right)\right]^{2}}}$$
(9)

where $\epsilon \leq 20$ dB is the maximum audible perturbation threshold, which can be defined (or optimized) by the adversary, $\lambda[\cdot]$ denotes the eigenvalue vector function class obtained with Schur decomposition. Γ , Φ , and $\Phi^{\mathcal{H}}$ (conjugate transpose of Φ) are random unit 2-norm operators, which satisfy [36]:

$$\mathbf{x}_i \Gamma = \lambda[\mathbf{x}_i] G(\mathbf{z}_i) \Gamma \quad \text{and} \quad \Phi^{\mathcal{H}} \mathbf{x}_i = \lambda[G(\mathbf{z}_i)] \Phi^{\mathcal{H}} G(\mathbf{z}_i)$$
(10)

For simplicity, we assume that these operators are static for all samples. Although this assumption simplifies the computation, it might result in ill-conditioned cases where an adjustment $\gamma[\cdot]$ is needed (chord(\cdot) + $\gamma[\cdot]$) [36]. It has been shown that this adjustment is relatively large for adversarial spectrograms compared to original and noisy samples [28]. Therefore, iteratively minimizing over $\gamma[\cdot]$ for chord($\lambda[G(\mathbf{z}_i)], \lambda[\mathbf{x}_i]$) considerably increases the chance of finding the safe \mathbf{z}_i^* that satisfies the conditions mentioned above [26], [28].

Since $\lambda[\cdot]$, defined in the Schur decomposition domain, is sorted (descending) and it is inductive (coefficient of both $\lambda[G(\mathbf{z}_i)]$ and $\lambda[\mathbf{x}_i]$ have upper bound [9]), according to Zorn lemma [9] there exists a relative maximal coefficient for both $G(\mathbf{z}_i)$ and \mathbf{x}_i in the Hahn–Banach analytic form. Thus, we define:

$$\underbrace{\gamma[\lambda[G(\mathbf{z}_j)], \lambda[\mathbf{x}_j]]}_{\gamma^*[\cdot]} \leq \gamma[\lambda[G(\mathbf{z}_i)], \lambda[\mathbf{x}_i]] \quad \text{for} \quad j \ll i \quad (11)$$

where j should be chosen according to the properties of the spectrograms. However, we empirically set $j \doteq \max(i) \cdot 0.25$ to make a reasonable trade-off between spectrogram quality and computational complexity (75% improvement). On the other hand, this operation might constitute ill-conditioned pencils (a pencil is a manifold in the closed-form of $\psi G(\mathbf{z}_i) - \mathbf{x}_i$ where $\psi \propto p_g$ [36]) by discarding (i - j) eigenvectors. To tackle this challenge, we add a complementary regularization term to the spectrogram subspace projection formulation:

$$\nabla_{\mathbf{z}_{i}} \left\| \gamma^{*} \left[G(\mathbf{z}_{i}), \mathbf{x}_{i} \right] \right\|_{F}^{2} + \underbrace{\nabla_{\mathbf{z}_{i}} \left\| \operatorname{span}(G(\mathbf{z}_{i}) - \mathbf{x}_{i}) \right\|_{1}}_{\text{regularization}} \quad (12)$$

where span(·) computes a linearly independent manifold in the Schur decomposition domain from the difference between the input and synthesized spectrograms [36]. The intuition behind this regularization term is tying $G(\mathbf{z}_i)$ as close as possible to \mathbf{x}_i and counteracting with the potential ill-conditioned pencils imposed from $\gamma^*[\cdot]$. Ill-conditioned cases often happen when $\gamma^*[\cdot]$ is minimized, but $G(\mathbf{z}_i)$ and \mathbf{x}_i are not similar.

Upon solving this optimization problem (Eq. 12), we achieve a candidate set $Z_{\mathfrak{V}} = \{\mathbf{z}_{\mathfrak{V},i}\}$ among all the possible $\mathbf{z}_i \in \mathbb{R}^{d_z}$. Finally, we find the most optimal vector from $Z_{\mathfrak{V}}$ via solving for:

$$\mathbf{z}_{i}^{*} := \arg\min_{\mathbf{z}_{i} \in \mathcal{Z}_{\mho}} \left\| \gamma^{*} \left[G(\mathbf{z}_{i}), \mathbf{x}_{i} \right] \right\|_{F}$$
(13)

where \mathbf{z}_i^* is presumably refined to provide a safe input vector for the generator model. We do not directly filter the spectrograms to remove adversarial perturbation δ . We find a reliable vector for a generative model to synthesize a similar spectrogram. However, the performance of all these operations



Fig. 2: Overview of the proposed spectrogram subspace projection using the chordal distance adjustment and a complementary regularization term. The subsampling process is implemented with the distribution $\mathcal{N}(0.5, 0.5I)$ (ratio of 0.5) for avoiding ill-conditioned pencils [36], and a dotted line shows the internal loop. Upon producing a candidate set of Z_{\Im} vectors from the given inputs, we select that \mathbf{z}_i which minimizes the adjusted chordal distance between the synthesized spectrogram $G(\mathbf{z}_i)$ and the input spectrogram \mathbf{x}_i .

is highly dependent on the generalizability and stability of the GAN model.

and the dominant probability density function $\mu(\cdot)$ has the following definition [9], [33]:

C. Spectrogram Synthesis Using a Sobolev-GAN

The generative model proposed for synthesizing spectrograms is based on the vanilla GAN [38] but with an integral probability metric defined in the Sobolev space [9], [33]. Since a specific case of such a space correlates with Fourier transform, we use this measure for training our GAN on STFT spectrograms. Moreover, we introduce novel architectures for both generator and discriminator networks. For improving the generalizability and the stability of the entire model, we propose imposing a constraint on the restricted Sobolev space and incorporating multiple discriminator networks.

The task of a generator network in a GAN configuration is minimizing the discrepancies between the synthesized (p_g) and real/original (p_r) sample distributions based on a specific measure [38]. The choice of such a measure is quite important since it contributes to the generalizability of the entire model (both generator and discriminator networks) [39]. During the last years, many improvements have been made in designing comprehensive distance measures on top of the φ -divergence [38] such as Wasserstein [39], Stein [40], Cramér [41], maximum mean discrepancy (MMD) [42], [43], and μ -Fisher IPM [44]. The function which measures this discrepancy is called critic, and it can be formulated (in the closed-form) as [45]:

$$\sup_{f \in \mathcal{F}} \left[\mathbb{E}_{G(\mathbf{x}_i) \sim p_g} f(G(\mathbf{x}_i)) - \mathbb{E}_{\mathbf{x}_{org} \sim p_r} f(\mathbf{x}_{org}) \right]$$
(14)

where \mathcal{F} refers to the function class, which is independent of p_g and p_r [46]. For improving the GAN stability during training, restriction often applies to the critic function following the characteristics of \mathcal{F} such as Lipschitz continuity $(||f||_{\text{Lip}} \leq 1)$ in Wasserstein-GAN [39] and kernel Hilbert unit ball $(||f||_{\text{Hil}} \leq 1)$ in MMD-GAN [43]. Moreover, these restrictions should be inline with the properties of the training sample modality. They might result in a weak or unstable generative model, especially for sequence generation (e.g., text and speech) [33].

The similarity measure used for training our GAN is the Sobolev IPM, adapted for sequence-to-sequence generation [33] such as chunks of speech signals. Formally, the function class in the Sobolev space with the zero boundary condition

$$\mathcal{F} = \left\{ \mathcal{X} \to \mathbb{R}^{d_z} \mid \mid \mathcal{X} \to L^{p_s}(\mathbb{R}^{d_z}), f \in W^{k_s, p_s}(\mathcal{X}, \mu), \\ \mathbb{E}_{\mathbf{x} \sim \mu} \left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|^2 \le 1, \mu \sim \mathbb{P}(p_r, p_g), k_s, p_s \in \mathbb{N} \right\}$$
(15)

where $\mathcal{X} \in \mathbb{R}^{d_z}$ is a compact open subset, $L^{(\cdot)}$ indicates the Lebesgue norm for $1 \leq p_s \leq \infty$, k_s denotes the order of the critic function, and \mathbb{P} is the probability density function. The special case of the function class \mathcal{F} is for $p_s = 2$ where it forms a Hilbert space $\mathcal{H}^{k_s} = W^{k_s,2}$ in connection with Fourier transform as follows [9]:

$$\mathcal{H}^{k_s}(\cdot) \simeq \sum \alpha_s \left| \hat{f}(x) \right|^2 < \infty, f \in L^2(\cdot)$$
 (16)

where α_s is a scalar, and $\hat{f}(x)$ refers to the Fourier series for $f(\cdot)$. Since a spectrogram is also a set of Fourier coefficients, $W^{k_s,2}$ provides a meaningful domain for capturing local distributions of SP_{STFT}. We also assume $k_s = 1$ and simplify the underlying Sobolev space as [33]:

$$W^{1,2}(\mathcal{X},\mu) = \left\{ f : \mathcal{X} \to \mathbb{R}^{d_z}, \int_{\mathcal{X}} \left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|^2 \mu(\mathbf{x}) d\mathbf{x} < \infty \right\}$$
(17)

where this restricted Sobolev space also constraints the critic function f into a unit ball $\mathbb{E}_{\mathbf{x}\sim\mu} \|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \leq 1$. There are numerous possible choices for defining the dominant measure $\mu(\cdot)$ according to this restricted space's properties. However, we initialize it to $\mu(\cdot) = 0.5 \cdot (p_r + p_g)$ which is the optimal case in training a GAN [33]. Based on these explanations and using Eq. 14, we can formulate the Sobolev GAN as [33]:

$$\min_{G_{\theta_g}} \left[\sup_{f_{\vartheta}, \frac{1}{N} \sum_{i=1}^{N} \|\nabla_{\mathbf{x}} f_{\vartheta}(\tilde{\mathbf{x}})\|^2 \leq 1} \right] \mathcal{E}(f_{\vartheta}, G_{\theta_g}) = \\
\frac{1}{N} \left(\sum_{i=1}^{N} f_{\vartheta}(G(\mathbf{z}_i)) - \sum_{i=1}^{N} f_{\vartheta}(\mathbf{x}_i) \right), \vartheta \geq 2$$
(18)

where the critic function f_{ϑ} follows the imposed constraint in Eq. 17, and ϑ is the degree of the critic function. Additionally, N refers to the total number of training samples, and θ_g denotes the weight vectors of the generator network. Moreover, for supporting the continuity and smoothness of f_{ϑ} , especially for higher-order ϑ , it is recommended to define [47]:

$$\tilde{\mathbf{x}}_i = \alpha_\vartheta \left(u \mathbf{x}_{r,i} + (1-u)G(\mathbf{z}_i) \right) \tag{19}$$

where $u \sim \mathcal{U}[0, 1]$ and α_{ϑ} is an empirical hyperparameter (we initialize it to $\alpha_{\vartheta} = 0.9$). This change of variable implicitly interpolates between p_r and p_g to enhance generator model stability [47]. However, this enhancement is also dependent on the configurations of both the generator network in optimizing for Eq. 18 and the discriminator network, which provides gradient vectors to G_{θ_g} .

Our proposed architecture for the generator network employs convolution and residual blocks due to their representation power in capturing continuous density functions of the input space [48], [49] such as spectrograms (see Fig. 3). The generator network contains a fully connected 1D vector layer equivalent to the total dimension of the spectrogram (128×128) , followed by batch normalization (BN) and rectified linear unit activation function (ReLU). This network's first hidden layers are two convolution blocks with the receptive field and stride of $5 \times 5 \times 1$. The second hidden layer contains three consecutive residual blocks where each of them has dilated convolution operation with aggregation. Inspired by Kumar *et al.* [50], the filter size of these blocks are identical. Finally, this network's output layer is a transposed convolution [51], which yields an RGB spectrogram.

Since the discriminator network provides gradients to the generator and has a crucial role in the entire model's stability [52], we empirically embedded five discriminators with identical architectures. However, we unloaded these networks from residual and long short-term memory (LSTM) blocks to avoid unnecessary complications. The filter sizes in these networks are different and, they escalate by a factor of two so that encompassing a broader range of spectrum distribution. Unlike the generator network, all the convolution layers in the discriminators deploy leaky ReLU (LReLU), as discussed in [53]. The general formulation for training these GANs is:

$$\min_{G} \max_{D_i} \mathbb{E}_{\mathbf{x} \sim p_r} \left[\log D_i(\mathbf{x}) \right] + \mathbb{E}_{\mathbf{z} \sim p_z} \left[\log \left(1 - D_i(G(\mathbf{z})) \right) \right]$$
(20)

where $\forall i = 1 : 5$ and $p_z \sim \mathcal{N}(0, I)$. The loss function of these networks is similar to the hinge objective function introduced in [52]. However, according to the Sobolev IPM:

$$\mathcal{L}_{S}(\vartheta, \theta_{g}, \theta_{d}, \varrho_{s}) = \mathcal{E}(f_{\vartheta}, G) + \varrho_{s}(1 - \Omega_{s}(f_{\vartheta}, G)) -\frac{\rho}{2}(\Omega_{s}(f_{\vartheta}, G) - 1)^{2}$$
(21)

and in this definition, $\Omega_s(\cdot)$ in the restricted Sobolev space $W^{1,2}(\mathcal{X},\mu)$ is differentiable and regarding Eq. 18, it is defined as [33]:

$$\frac{1}{2N} \left(\sum_{i=1}^{N} \left\| \nabla_{\mathbf{x}} f_{\vartheta}(\mathbf{x}_{i}) \right\|^{2} + \sum_{i=1}^{N} \left\| \nabla_{\mathbf{x}} f_{\vartheta}(G(\mathbf{z}_{i})) \right\|^{2} \right)$$
(22)

Moreover, ρ_s , θ_d , and $\rho > 0$ denote the Lagrange multiplier, the weight vectors of each discriminator network, and the penalty weight for providing higher smoothness in training, respectively [44]. One potential side effect of training the generator with multiple discriminators is the difficulty of making a trade-off between sample variety an quality. For tackling this challenge, we use orthogonal regularization (OR) for all the discriminator networks using a simple linear similarity measure [54]:

$$R_{\varpi} = \varpi \left\| \theta_d^{\top} \theta_d \odot (\mathbf{1} - I) \right\|_F^2$$
(23)

where empirically $\varpi \in (10^{-5}, 10^{-5}]$ is a small tuning coefficient, and **1** indicates a matrix with constant values of one [49]. This regularization forces the discriminator network to reduce dissimilarity among filters to learn more distinctive features. However, this might negatively affect the generator performance in capturing all the possible modes from the spectrogram, cause instability in a higher number of iterations, and generate oversmoothed samples [55]. In response to this issue, we propose a new constraint for the critic function f_{ϑ} as the following.

Proposition: There is an achievable upperbound (supremum) for the continuous (and partially differentiable) critic function $f_{\vartheta}(\cdot)$ in the restricted Sobolev space $W^{1,2}(\mathcal{X},\mu)$ with:

$$L^{\eta}(\mathcal{X}) = \left\{ f_{\vartheta} : \mathcal{X} \to \mathbb{R}, |f_{\vartheta}|^{\eta} \in L^{1}(\mathcal{X}) \right\}$$
(24)

where $||f||_{L^1} = ||f||_1$ and $1 \le \eta \le \infty$. This reduces the space definition in Eq. 17 to $\int_{\mathcal{X}} ||\nabla_{\mathbf{x}} f(\mathbf{x})||^2 \mu(\mathbf{x}) d\mathbf{x} \le c_{\Upsilon}$ where c_{Υ} is a positive static scalar.

Proof: According to the rigid constraint $\mathbb{E}_{\mathbf{x}\sim\mu} \|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \leq 1$ imposed on $W^{1,2}(\mathcal{X},\mu)$ in Eq. 17, it always supports $\|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \in L^{\eta}$ (the Lebesgue norm). If we bind $\mu(\mathbf{x}) \in L^{\eta'}$ where η' denotes the conjugate exponent of $\eta (1/\eta + 1/\eta' = 1)$, then using the Hölder's inequality [9], we can write:

$$\int_{\mathcal{X}} \left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|^2 \mu(\mathbf{x}) d\mathbf{x} \le \underbrace{\left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|_{\eta}^2 \left\| \mu(\mathbf{x}) \right\|_{\eta'}}_{c_{\Upsilon} < <\infty} \quad \Box \quad (25)$$

where c_{Υ} is dependent on the cumulative distribution of $\mu(\mathbf{x})$. This constraint forces the generator network to discard local sample distributions which lie far from the optimal generator distribution ($\lceil p_r + p_g \rceil / 2$). It also implicitly helps the discriminator network avoid shattering gradients vectors since the learning space bound to c_{Υ} .

For synthesizing a spectrogram similar to the given \mathbf{x}_i , the generator network maps the safe vector \mathbf{z}_i^* onto $\hat{\mathbf{x}}_i$ and then tunes the generated spectrogram with the \mathbf{x}_i 's rank [36] in the Schur decomposition domain. Even if this tuning is optional, it improves the quality of $\hat{\mathbf{x}}_i$ and reduces the potential dissimilarity between $G(\mathbf{z}_i^*)$ and \mathbf{x}_i .

The last step of the proposed adversarial defense approach is transforming the synthesized spectrograms into the time domain using the inverse STFT operation. This step is necessary only for end-to-end speech-to-text victim models upon adversary's discern.

Reconstructing an audio or speech signal from a spectrogram requires the associated phase vectors from the transformation function (e.g., STFT). There are two main approaches for such an aim: using original phase vectors and approximating phase vectors. Obviously, in the first approach, the reconstructed signals' quality will be very similar to the original counterparts since they share the same timing. However, original phase vectors might not always be accessible,



Fig. 3: Overview of the proposed GAN architecture (one generator and five discriminators $D_{i,\theta}$ for $\forall i = 1:5$) for spectrogram synthesis. Fully connected (FC), convolution (Conv.), dilated convolution (D-Conv.), transposed convolution (T-Conv.), and residual (Res.) layers are followed by weight normalization. The top and bottom parts of the layers refer to the input and output filters' dimensions, respectively. Moreover, ν_i for $\forall i = 1:5$ denotes the logits of the discriminator.

contrary to the second signal reconstruction approach. On the other hand, approximated phase vectors usually add audible noise to the reconstructed signal and degrade its quality. Therefore we opted for the second approach since accessing the original phase vectors might be prohibitive in some senses. Specifically, we use the recognized Griffin-Lim algorithm for the i-STFT procedure [30]. Since this may raise concerns about the quality of the reconstructed signals, we measure their peculiarity with some metrics.

V. EXPERIMENTAL RESULTS

In this section, we analyze the proposed defense scheme's performance from two points of view: the defense algorithm's success rate by measuring the word error rate and sentencelevel accuracy scores, and the quality of the signals from the synthesized spectrograms and the approximated phase vectors. The latter also includes comparing signals after filtration by various defense algorithms. This shows the impact of defense algorithms on speech signals.

Our benchmarking victim models are DeepSpeech, Kaldi, and Lingvo, which employ both the conventional and cuttingedge learning blocks, such as HMM, convolutional, recurrent, LSTM, and residual configurations. These models are trained on Mozilla common voice (MCV) [57] and LibriSpeech [58] comprehensive datasets, including numerous utterances. Moreover, they contain above 1,000 hours of recordings organized in short (≤ 6 sec) and long (> 6 sec) voice clips.

In all our experiments, we use a combination of strong white and black-box end-to-end adversarial attacks, as discussed in Section II. For every adversarial signal, regardless of EOT type, we assign ten targeted incorrect different phrases, including silence [7], and five non-targeted incorrect random phrases with different lengths to more effectively challenge defense approaches. Meanwhile, we take identical assumptions for those algorithms that require environmental settings such as CIR and RIR filter sets for fairness in comparison. Following a common practice in adversarial studies [7], [10], [19], we also craft adversarial signals for a group of randomly selected portions (with shuffling) of the datasets mentioned above. More specifically, we randomly choose 25k English-speaking samples from both MCV and LibriSpeech with an almost equal number of genders (male and female), accent (e.g., United States, England, etc.), and age (the majority between 19 to 39 regarding the dataset limitation). We assign almost 60% of these samples for training, tuning, and validating our generative model. Hence, the remaining portion will be used for developing adversarial signals using six attack algorithms.

Since we train our GAN on the spectrograms, we firstly convert speech signals into SP_{STFT} with a sampling rate of 22.05 kHz. Additionally, we set the total number of Melfrequency coefficients to 20 per frame with an overlapping ratio of 0.5 and the hop length of 512. The Hann window length is initialized to 2048 with reflect padding.

We discard checkpoints with unstable learning curves during training and opt to early stop when any signs of instability become present [49]. For all the architectures (the generator and five discriminators) we use Adam optimizer with a static learning rate of $2 \cdot 10^{-5}$ and hyperparameters $\beta_1 = 0$ and $\beta_2 = 0.9$. We empirically set the required number of steps for the generator network over the discriminators to two with a decay ratio of 0.99 on four NVIDIA GTX-1080-Ti and two 64-bit Intel Core-i7-7700 (3.6 GHz) with 8×11GB and 2×64GB memory, respectively.

For evaluating the performance of the proposed defense

algorithm against adversarial attacks, we also use the word error rate (WER) and sentence level accuracy (SLA) [10]. The first metric measures the summation of total phrase insertion, substitution, and deletion over the ground-truth phrases (y_i). The second metric measures the ratio of correctly transcripted phrases over the total number of test speech signals. To avoid bias in our analysis, we repeat each experiment 10 times and report the average WER and SLA for each defense algorithm. Table I summarizes the achieved results.

Table I shows that for most cases, the proposed defense approach (Sobolev-DGAN*) and its variant without employing the constraining proposition (Sobolev-DGAN) introduced in Section IV-C outperform other defense algorithms against six strong end-to-end speech attacks. Averaged over all the conducted experiments on the three victim speech-to-text models, Sobolev-DGANs have similar performance on white (C&W, Metamorph, Imperio, and Robust Attack) and blackbox (GAA and MOOA) attack algorithms. That indicates the independence of our defense algorithm to the adversarial attack scenarios. Moreover, the total number of required iterations (\mho) toward achieving the safe input vector \mathbf{z}_i^* for the C&W attack and the Robust Attacks is relatively more than others. That could be interpreted as the higher power of these attacks in yielding more destructive adversarial signals since they demand an additional cost for our defense algorithm to find the input vector. However, any discussion on the resiliency of adversarial attacks and their potentials in optimizing upscale examples is beyond this paper's scope.

Furthermore, Table I also proves the effectiveness of the proposed constraining technique for the critic function f_{ϑ} as discussed in Section IV-C. Except for the GAA, Sobolev-DGAN* has shown higher SLA than the Sobolev-DGAN on all the victim speech-to-text models.

For evaluating the potential negative impact of running defense algorithms on the crafted adversarial signals, we use four objective speech quality metrics: perceptual evaluation of speech quality (PESQ) [59], segmental signal to noise ratio (segSNR) [60], short-term objective intelligibility (STOI) [61], and log-likelihood ratio (LLR) [60]. The first metric is based on cognitive modeling, and the input filter set aligns with identifying noisy intervals (high-level quality analysis). The second metric is the enhanced version of the conventional signal-to-noise ratio in audible logarithmic scale for chunks of speech signals (low-level quality analysis). The third metric evaluates the ratio of band-pass local noise perceptibility to the entire signal chunks. Unfortunately, these metrics are not normalized in a scaled interval. However, there is a direct relationship between their magnitudes and signal quality. The fourth metric is associated with a logarithmic noise ratio relative to the ground-truth scaled between [0, 1]. Therefore, high-quality signals have lower LLR. As shown in Table I, for the most cases, averaged over ten times experiment repetitions, both the Sobolev-DGAN* and Sobolev-DGAN outperform others in keeping the quality of the signals after running the defense filtration.

In Section IV-C, we mentioned that $W^{k_s,2}$ provides a meaningful (and comprehensive) domain for capturing local distributions of spectrograms. To investigate this claim, Fig. 4 shows the relation between the Sobolev IPM and extracted local and global probability distributions from spectrograms compared to others. Toward this end, inspired by Mao *et al.* [51], we compare the mode collapse issue between the GANs trained with various IPMs as mentioned in Section IV-C. We have used an identical architecture for all generative models (generator and discriminators depicted in Fig. 3) for fairness in comparison. Additionally, we have used the same settings for these networks.



Fig. 4: Monitoring the average learned modes (per batch size of 2×512) by our GAN model during training on SP_{STFT} with different IPMs indicates potential collapse over the total number of iterations.

Fig. 4 shows that the average number of learned modes has an increasing behavior of up to 20k iterations for MMD and μ -Fisher IPMs. For Wasserstein and Cramér IPMs, this behavior reaches around 26k iterations. Among these, the Sobolev IPM keeps its incremental behavior up to 30k iteration with considerable bias (along the *y*-axis). That demonstrates the higher performance of f_{ϑ} in capturing the local distribution of spectrograms in the restricted Sobolev space compared to other IPM. However, it does not immune our generative model against the mode collapse issue. As depicted in Fig. 4, our GAN gradually starts losing sample modes after 31k iterations. For tackling this issue, we used OR, spectral normalization [52], and early stopped at checkpoints before the collapse.

Since there is a direct relationship between stability and generalizability of the GAN and our proposed defense algorithm, even a partially unstable generator network might result in absolute divergence in the chordal distance adjustment operation. In other words, if the GAN model is not comprehensive enough in terms of the number of learned modes, the process shown in Fig. 2 might never converge. This poses more concerns for long signals with too much environmental noise. Additionally, for multi-speaker speech signals, our proposed Sobolev-DGANs not only might not be able to learn enough modes but also might recover adversarial perturbation after the i-STFT procedure. We believe that employing more constraining conditions on both the generator and discriminators may improve model stability. Moreover, TABLE I: Comparison of the defense algorithms against strong white and black-box adversarial attacks for the DeepSpeech, Kaldi, and Lingvo victim speech-to-text models. Whereas WER and LLR, higher values for the SLA, PESQ, segSNR, and STOI metrics are better. The difference between Sobolev-DGAN* and Sobolev-DGAN is the latter does not incorporate the constraint proposition (Eq. 25) mentioned in Section IV-C. Outperforming results are shown in boldface.

Model	Attack	Defense	Average \mho	WER (%)	SLA (%)	PESQ	segSNR	STOI	LLR
DeepSpeech	C&W	Compression [21]	-	19.14 ± 2.36	49.26 ± 2.67	1.64	09.31	0.85	0.44
		A-GAN	-	26.32 ± 3.03	36.21 ± 0.12	1.15	06.95	0.87	0.41
		CC-DGAN	-	14.52 ± 1.16	61.23 ± 1.02	2.01	12.56	0.89	0.38
		Sobolev-DGAN	163	07.61 ± 0.47	76.15 ± 2.18	2.36	18.73	0.91	0.31
		Sobolev-DGAN*	159	04.21 ± 1.39	79.24 ± 1.17	2.71	19.91	0.95	0.30
	Metamorph	Compression [21]	-	21.54 ± 2.17	51.57 ± 1.91	1.55	10.34	0.76	0.48
		A-GAN	-	19.81 ± 3.72	58.39 ± 0.49	1.59	10.86	0.83	0.32
		CC-DGAN	-	11.89 ± 1.23	71.94 ± 1.56	1.96	11.08	0.85	0.35
		Sobolev-DGAN	039	09.37 ± 1.12	75.19 ± 2.18	2.17	14.76	0.88	0.34
		Sobolev-DGAN*	027	06.79 ± 0.19	80.34 ± 3.67	2.45	16.01	0.93	0.31
	GAA	Compression [21]	-	27.41 ± 3.61	43.71 ± 1.32	2.14	14.37	0.87	0.39
		A-GAN	-	29.49 ± 5.26	40.88 ± 5.37	1.66	12.53	0.88	0.37
		CC-DGAN	-	14.98 ± 3.56	69.46 ± 2.37	2.03	13.52	0.90	0.34
		Sobolev-DGAN	101	09.68 ± 2.73	73.98 ± 0.77	2.39	16.02	0.93	0.29
		Sobolev-DGAN*	097	05.01 ± 0.11	72.88 ± 4.28	2.38	18.91	0.94	0.30
	MOOA	Compression [21]	-	17.06 ± 0.19	55.16 ± 3.86	1.87	19.42	0.92	0.38
		A-GAN	-	18.74 ± 43.21	53.07 ± 3.06	1.85	14.63	0.87	0.41
		CC-DGAN	-	15.69 ± 1.97	61.11 ± 2.99	1.99	17.81	0.89	0.39
		Sobolev-DGAN	051	12.25 ± 2.84	68.84 ± 1.56	2.46	19.35	0.90	0.36
		Sobolev-DGAN*	049	04.23 ± 2.32	79.36 ± 2.16	2.30	18.06	0.91	0.39
Kaldi	Imperio	Compression [21]	-	16.29 ± 5.17	56.42 ± 6.11	2.42	15.79	0.83	0.32
		A-GAN	-	17.76 ± 0.16	54.28 ± 1.90	1.23	09.76	0.74	0.48
		CC-DGAN	-	10.19 ± 2.93	69.62 ± 2.63	1.84	16.53	0.78	0.45
		Sobolev-DGAN	093	06.78 ± 0.91	75.33 ± 2.97	1.96	13.98	0.81	0.41
		Sobolev-DGAN*	047	03.29 ± 1.14	82.37 ± 3.62	2.35	16.52	0.89	0.35
Lingvo	Robust Attack	Compression [21]	-	21.56 ± 4.15	55.11 ± 3.05	2.06	15.08	0.74	0.33
		A-GAN	-	17.90 ± 4.21	59.98 ± 1.38	2.17	14.43	0.72	0.34
		CC-DGAN	-	14.46 ± 0.35	64.16 ± 2.14	1.71	11.09	0.79	0.28
		Sobolev-DGAN	114	11.99 ± 2.76	69.33 ± 0.81	1.92	12.25	0.76	0.34
		Sobolev-DGAN*	136	05.86 ± 1.64	83.46 ± 2.27	1.96	17.07	0.81	0.22

conditioning the discriminator networks aligned with timedistributed filter sets can provide more distinctive features for the discriminator network to resolve the multi-speaker issue. We are determined to address these issues in future work.

VI. CONCLUSION

In this paper, we proposed a novel approach for defensing speech-to-text models against end-to-end adversarial attacks. Our approach is based on reconstructing signals from synthesized spectrograms and approximated phase vectors. For spectrogram synthesis, we use a multi-discriminator GAN defined in the restricted Sobolev space. Our GAN generator network requires a safe input vector achievable through an iterative spectrogram subspace projection operation using the adjusted chordal distance. To improve our implemented generative model's performance, we impose a constraint for the critic function that learns discrepancies between real and synthesized sample distributions. We evaluated our defense approach against six strong white and black-box adversarial attacks on advanced DeepSpeech, Kaldi, and Lingvo victim models. The proposed defense approach, averaged over the total number of experiments, outperformed other algorithms according to WER and SLA metrics. Furthermore, we used four objective quality metrics for measuring the impact of running defense algorithms on speech signals. For the majority of the cases, our defense approach demonstrated higher signal quality compared to other algorithms.

ACKNOWLEDGMENT

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN 2016-04855 and Grant RGPIN 2016-06628.

REFERENCES

- H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," in 21st Annual Conf Intl Speech Comm Assoc, 2020.
- [2] 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.
- [3] Mozilla, "Project DeepSpeech," https://github.com/mozilla/DeepSpeech, 2017, [Online; accessed 9-March-2021].
- [4] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2013, pp. 8614–8618.
- [5] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2013, pp. 6645–6649.
- [6] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "Endto-end attention-based large vocabulary speech recognition," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2016, pp. 4945–4949.
- [7] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Secur Privacy Works*, 2018, pp. 1–7.
- [8] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "A robust approach for securing audio classification against adversarial attacks," *IEEE Trans Inf Forensics Security*, vol. 15, pp. 2147–2159, 2020.
- [9] H. Brezis, Functional analysis, Sobolev spaces and partial differential equations. Springer, New York, NY, 2010.
- [10] 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.

- [11] 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.
- [12] 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.
- [13] J. Szurley and J. Z. Kolter, "Perceptual based adversarial audio attacks," arXiv preprint arXiv:1906.06355, 2019.
- [14] 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," in 26th Annual Netw Distrib Syst Secur Symp, 2019.
- [15] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," in 28th Intl J Conf Artif Intell, 2018, pp. 5334–5341.
- [16] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Intl Conf Mach Learn*, 2018, pp. 284–293.
- [17] 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.
- [18] 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 Recognit* Underst, 2011.
- [19] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," in *IEEE Secur Priv Works*, 2019, pp. 15–20.
- [20] S. Khare, R. Aralikatte, and S. Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," in 20th Annual Conf Intl Speech Comm Assoc, 2019, pp. 3208–3212.
- [21] N. Das, M. Shanbhogue, S. Chen, L. Chen, M. E. Kounavis, and D. H. Chau, "ADAGIO: interactive experimentation with adversarial attack and defense for audio," in *European Conf Mach Learn Knowl Discov Datab*, 2018, pp. 677–681.
- [22] A. M. Peinado and J. C. Segura, Speech Recognition over Digital Channels. John Wiley & Sons, Ltd, 2006.
- [23] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans Speech Audio Process*, vol. 7, no. 2, pp. 126–137, 1999.
- [24] Y. Hu and P. C. Loizou, "A comparative intelligibility study of singlemicrophone noise reduction algorithms," *Journal Acoust Soc America*, vol. 122, no. 3, pp. 1777–1786, 2007.
- [25] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," *arXiv preprint arXiv:1705.02900*, 2017.
- [26] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Class-conditional defense GAN against end-to-end speech attacks," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2021, p. to appear.
- [27] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," arXiv preprint arXiv:1811.11402, 2018.
- [28] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Detection of adversarial attacks and characterization of adversarial subspace," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2020, pp. 3097–3101.
- [29] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans Acoust, Speech, Signal Process*, vol. 32, no. 2, pp. 236–243, 1984.
- [30] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep griffin–lim iteration," in *IEEE Intl Conf Acoust, Speech, Signal Process*, 2019, pp. 61–65.
- [31] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans Audio, Speech, Language Process*, vol. 27, no. 12, pp. 2041–2053, 2019.
- [32] L. Deng and D. O'Shaughnessy, Speech processing: a dynamic and optimization-oriented approach. CRC Press, 2018.
- [33] Y. Mroueh, C. Li, T. Sercu, A. Raj, and Y. Cheng, "Sobolev GAN," in 6th Intl Conf Learn Repres, 2018.
- [34] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *Intl Conf Learn Repres*, 2018.
- [35] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in 35th Intl Conf Mach Learn, vol. 80, 2018, pp. 274–283.

- [36] C. F. Van Loan and G. H. Golub, *Matrix computations*. Johns Hopkins University Press, 1983.
- [37] 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.
- [38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv Neural Inf Process Syst*, 2014, pp. 2672–2680.
- [39] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *5th Intl Conf Learn Repres*, 2017.
- [40] Y. Feng, D. Wang, and Q. Liu, "Learning to draw samples with amortized stein variational gradient descent," in 33rd Conf Uncert Artif Intell, 2017.
- [41] 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.
- [42] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," in 31st Conf Uncert Artif Intell, 2015, pp. 258–267.
- [43] C. Li, W. Chang, Y. Cheng, Y. Yang, and B. Póczos, "MMD GAN: towards deeper understanding of moment matching network," in Adv Neural Inf Process Syst, 2017, pp. 2203–2213.
- [44] Y. Mroueh and T. Sercu, "Fisher GAN," in Adv Neural Inf Process Syst, 2017, pp. 2513–2523.
- [45] A. Müller, "Integral probability metrics and their generating classes of functions," Adv Appl Probab, pp. 429–443, 1997.
- [46] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet *et al.*, "On the empirical estimation of integral probability metrics," *Electr Journal Statis*, vol. 6, pp. 1550–1599, 2012.
- [47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Adv Neural Inf Process Syst*, 2017, pp. 5767–5777.
- [48] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in 4th Intl Conf Learn Repres, 2016.
- [49] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Intl Conf Learn Repres*, 2019.
- [50] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Adv Neural Inf Process Syst*, 2019, pp. 14910–14921.
- [51] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "On the effectiveness of least squares generative adversarial networks," *IEEE Trans Patt Anal Mach Intell*, vol. 41, no. 12, pp. 2947–2960, 2018.
- [52] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Intl Conf Learn Repres*, 2018.
- [53] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with leakyrelu for environmental sound classification," in 22nd Intl Conf Digital Signal Process, 2017, pp. 1–5.
- [54] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Cneural photo editing with introspective adversarial networks," in *Intl Conf Mach Learn*, 2017.
- [55] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Unsupervised feature learning for environmental sound classification using weighted cycleconsistent generative adversarial network," *Applied Soft Computing*, vol. 86, p. 105912, 2020.
- [56] K. M. Koerich, M. Esmaeilpour, S. Abdoli, A. S. Britto Jr., and A. L. Koerich, "Cross-representation transferability of adversarial attacks: From spectrograms to audio waveforms," in *IEEE Intl J Conf Neural Netw*, 2020, pp. 1–8.
- [57] Mozilla, "Mozilla common voice dataset," https://voice.mozilla.org, 2019, [Online; accessed 9-March-2021].
- [58] 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, Signal Process*, 2015, pp. 5206–5210.
- [59] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Intl Conf Acoust, Speech, Signal Process*, vol. 2, 2001, pp. 749–752.
- [60] 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.
- [61] 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, Language Process*, vol. 19, no. 7, pp. 2125– 2136, 2011.