Mel frequency spectral domain defenses against adversarial attacks on speech recognition systems

Nicholas Mehlman,a) Anirudh Sreeram, Raghuveer Peri, and Shrikanth Narayanan
Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California 90089, USA
nmehlman@usc.edu, asreeram@usc.edu, rperi@usc.edu, shri@usc.edu

Abstract: Automatic speech recognition (ASR) systems are vulnerable to adversarial attacks due to their reliance on machine learning models. Many of the defenses explored for defending ASR systems simply adapt defense approaches developed for the image domain. This paper explores speech-specific defenses in the feature domain and introduces a defense method called mel domain noise flooding (MDNF). MDNF injects additive noise to the mel spectrogram speech representation prior to re-synthesizing the audio signal input to ASR. The defense is evaluated against strong white-box threat models and shows competitive robustness.

1. Introduction

Recently, there has been growing urgency to defend speech technology systems such as speaker recognition (Jati et al., 2021) and automatic speech recognition (ASR) against adversarial attacks, in which a malicious actor attempts to cause the model to mis-transcribe an utterance by introducing small carefully crafted perturbations—often with limited impact on human perception—to the benign audio. A vast majority of existing defenses in speech processing are inspired by similar work in computer vision, for example, randomized smoothing (RS) as defense was originally proposed for images by Cohen et al. (2019) but has been since applied to ASR (e.g., Zelasko et al., 2021).

While these “image-inspired” methods often provide acceptable adversarial robustness in the audio domain, they fail to take full advantage of the unique properties of speech signals. In particular, information conveyed in the speech signal is highly structured in both time and frequency, with a sophisticated underlying production and processing mechanism that distinguishes it from generic images. Therefore, defenses that have been shown to be effective in the image domain, such as RS, may not be optimal for ASR applications. For example, Hussain et al. (2021) has demonstrated enhanced robustness of techniques that leverage speech-specific representations such as mel spectrograms and linear-predictive coefficients (LPCs).

Building along these aforementioned lines, this paper explores novel ways to leverage mel domain speech representation within an adversarial defense framework. We introduce a novel noise flooding technique termed as “mel domain noise flooding” (MDNF) that injects white Gaussian noise (WGN) into the log mel spectrogram representation of speech prior to re-synthesis of the time domain signal.1 Robustness is further enhanced by shaping the variance of the noise to match the frequency distribution of adversarial perturbations. While prior work (e.g., Zelasko et al., 2021) has demonstrated the defensive capability of using speech re-synthesized from a mel frequency representation, we show that MDNF substantially improves the adversarial performance of a relatively weak re-synthesis defense and retains the robustness of an already well-performing one. We also show that the proposed defense outperforms RS across a range of attack threat models.

2. Background

2.1 Adversarial attacks

Adversarial attacks can be broadly separated into two categories based on their level of knowledge of the ASR model’s internal state (Ren et al., 2020). Black-box attacks are completely agnostic to the inner workings of the model. White-box attacks, on the other hand, have complete knowledge of the ASR’s inner workings and, in particular, can also extract loss gradients (with respect to the input audio) to construct perturbations. Adversarial attacks can also be classified based on their objective (Ren et al., 2020). Untargeted attacks simply attempt to “fool” the ASR system, i.e., cause it to produce an
incorrect transcription, while targeted attacks try to produce a specific erroneous output. For a more detailed description of the different attacks, we refer the reader to Ren et al. (2020).

One popular white-box attack is the fast gradient sign method (FGSM) first introduced by Goodfellow et al. (2015). Here, the adversary adds a one-shot perturbation constructed using the sign of the model’s gradients. A more sophisticated approach is the projected gradient descent (PGD) attack from Madry et al. (2019). Similar to FGSM, the perturbations are constructed using the sign of the gradients, but instead of a one-step process, the PGD attack is generated iteratively. As shown in Eq. (1), at iteration $k$, the perturbed signal $x_k$ is projected onto an $\epsilon$-ball (typically using the $\ell_2$ or $\ell_\infty$ norm) around the original input to enforce an adversarial perturbation that is minimally noticeable (Madry et al., 2019; Zelasko et al., 2021),

$$x_{k+1} = \text{Proj}_\epsilon(x_k + \epsilon \text{sgn}(\nabla_x (\mathcal{L}(\phi(x_k), y_b)))).$$  (1)

Here, $\text{Proj}_\epsilon$ represents projection onto the $\epsilon$-ball, and $\mathcal{L}(\phi(x_k), y_b)$ is the loss function between the ASR transcription of the perturbed signal $x_k$ and the benign transcription $y_b$ (our experiments used the CTC loss). $\epsilon$ is a hyper-parameter that controls the relative strength of the attack (Madry et al., 2019). An alternative version of PGD uses a minimum signal-to-noise ratio (SNR) as a bound for the perturbation magnitude.

Another common attack is the targeted Carlini–Wagner (CW) attack. The CW attack is based on a constrained optimization problem that searches for a perturbation $\delta$ with $||\delta|| < \tau$, which produces the target mis-transcription (Carlini and Wagner, 2018). By repeatedly solving this problem with gradually smaller values of $\tau$, the attack enforces a minimally perceptible modification that still obtains the desired result.

### 2.2 Defenses against adversarial attacks

One common approach to making a neural network model (also integral to ASR) more robust to adversarial attacks is adversarial training (AT) (Szegedy et al., 2014) or integrating adversarial examples as a part of the model’s training data. This method builds upon other commonly employed methods for data augmentation, such as adding Gaussian noise to an audio signal or applying random rotations to an image, and has been used to develop ASR systems robust to adversarial attacks (Sun et al., 2018). However, such improved robustness generally comes at the cost of degraded performance on benign samples (Jati et al., 2021). Additionally, since AT is implemented during ASR training, it is not viable to use in pre-existing ASR systems.

Another popular defense (with certifiable guarantees) is RS, which involves the introduction of WGN to the input of the ASR system (Cohen et al., 2019; Zelasko et al., 2021). Predictions are made by averaging the model’s logits over several different random perturbations. Provided the ASR model is robust to additive noise, introducing these stochastic perturbations to the input can counteract the carefully crafted adversarial perturbations constructed by the attacker (Cohen et al., 2019; Zelasko et al., 2021). Cohen et al. (2019) prove that this procedure provides certifiable robustness to attacks constrained in the $\ell_2$ norm. Despite the good performance, RS increases computation time during inference due to the need for multiple forward passes per sample.

A variety of more sophisticated defenses utilize re-synthesis of the speech signal as a way to “discard” adversarial content. In this category of defenses, recently developed systems use generative adversarial networks (GANs) to produce a facsimile of the original input signal that retains the primary content of the speech (Yamamoto et al., 2020). GANs consist of a generator that seeks to produce realistic data and a discriminator that attempts to distinguish legitimate data samples from the synthetic samples produced by the generator. Zelasko et al. (2021) use the mel domain in conjunction with the generator portion of the WaveGAN vocoder introduced in Yamamoto et al. (2020). This approach takes the mel spectrogram as a conditioning input and attempts to re-synthesize the audio in a single pass. Their results demonstrate that introducing the vocoder as a pre-processor to the ASR system improves the model’s performance under the PGD, FGSM, and imperceptible attacks (Zelasko et al., 2021).

#### 2.3 Mel domain representation of speech signals

The mel domain represents a speech signal in a time-frequency format specific to the structure of human auditory processing. Mel spectrograms are classic speech signal representations commonly used as features for speech recognition and other speech applications (Davis and Mermelstein, 1980). The frequency response of the mel filter bank used to derive this representation is inspired by speech processing in the “critical bands” of the human auditory system (Rabiner and Schäfer, 2011) operationalized by a series of variable-width bandpass filters whose bandwidth increases with increasing frequency, where the human ear has been shown to have a coarser resolution. While the signal conversion to the mel domain is lossy and significantly reduces the dimensionality compared to a raw audio waveform, it is shown to retain information relevant to many speech related tasks, such as speech and speaker recognition.

### 3. Proposed defenses

We present two defenses that utilize the mel frequency representation of speech signals. First, we discuss the mel domain transform as a standalone defense, and then we show how stochastic noise flooding of the mel coefficients can further improve adversarial robustness.
As shown by Hussain et al. (2021), the process of re-synthesizing a speech signal from a mel spectrogram can act as a viable defense. Since the mel representation of a speech signal has limited time and frequency resolution, the process of converting a speech signal to a mel spectrogram is inherently lossy. However, given the speech-centric nature of this representation, it is reasonable to expect that information relevant to the linguistic content of the utterance will be retained (Hussain et al., 2021). The same does not apply to the adversarial perturbations, which are perceptually closer to random noise than to clean speech. Thus, the mel transform is likely to impede the propagation of adversarial information without irreparably degrading the speech itself (Hussain et al., 2021).

To re-synthesize the time domain speech, we use two different GAN-based generators: the WaveGAN (Yamamoto et al., 2020), which has been explored as a defense in Zelasko et al. (2021), and the melGAN, presented in Kumar et al. (2019).

To further improve the performance of this basic mel re-synthesis defense, we modify the above procedure to introduce mel domain noise flooding, as shown in Fig. 1. Prior to passing the spectrogram to the GAN for re-synthesis, we introduce WGN to the mel coefficients, thereby randomly perturbing the mel representation. The variance of the additive noise is a hyper-parameter that defines a trade-off between defense robustness and benign performance. A higher noise energy presents a greater challenge to an attacking adversary but also degrades the quality of the GAN’s reconstruction.

The most straightforward implementation would be to add WGN with equal variance across all mel frequency bins. However, an empirical analysis of the difference between adversarial and benign samples (produced using an untargeted PGD attack) revealed that the perturbations introduced are not uniformly distributed across the mel frequency bins. Instead, they appear to skew disproportionately toward lower frequencies as shown in Fig. 2. We conjecture that the non-uniform frequency distribution of the adversarial perturbations is reflective of the sensitivity of the ASR model. This finding inspired us to shape the additive noise along the frequency axis such that the variance in a given mel bin is directly

Fig. 1. Block diagram of the MDNF defense. Additive noise is introduced to the mel spectrogram. The shown curve is an empirically derived distribution of adversarial perturbations across the mel bins. Noise-variance shaping applies this curve to the noise across the frequency axis for each time step.

Fig. 2. Average mel domain distribution of adversarial perturbations (normalized). The curve was computed from 100 samples from an untargeted $\ell_2$ PGD attack by evaluating the difference between adversarial and benign spectrograms. We use this curve to shape the variance of the additive noise along the frequency dimension.
proportional to the relative strength of the attack as reflected by the curve shown in Fig. 2. As shown in Fig. 1, the noise is shaped prior to being added to the clean speech mel spectrogram.

It should be noted that our attack is still differentiable; therefore, the adversary is still able to leverage end-to-end (i.e., from ASR output to waveform) gradients for the defended model in its attack.

4. Experiments

Our defenses were tested within the Armory\textsuperscript{2} framework using the DeepSpeech\textsuperscript{2} ASR model (Amodei et al., 2016). We compare the performance of our proposed defenses with a competitive baseline, RS (Cohen et al., 2019), across a range of targeted and untargeted threat models. Our implementation of RS uses additive WGN at 15 dB SNR and averages the character probabilities over five forward passes.

Based on empirical observations, we found that the defense performed sub-optimally even under benign conditions when used with the pre-trained DeepSpeech\textsuperscript{2} model. This is likely because the re-synthesized signals generated by the WaveGAN and melGAN have audible artifacts, creating a mismatch with the signal expected by the ASR model. To address this discrepancy, we fine-tuned the DeepSpeech\textsuperscript{2} ASR on the re-synthesized samples (this was performed separately for the WaveGAN and melGAN, producing two fine-tuned ASRs). Based on preliminary experiments, we found that a mixture of 70\% re-synthesized and 30\% clean samples with WGN added to all samples produced a model with an acceptable benign ASR performance close to the original ASR model without re-synthesis.

The shaping curve shown in Fig. 2 was produced empirically using 100 pairs of adversarial and benign samples. By considering the difference between these pairs, we were able to isolate the adversarial perturbations and, thus, determine their average mel domain distribution.

4.1 Dataset

Our experiments used the Librispeech dataset, which consists of approximately 1000 h of wideband speech utterances sourced from audio books (Panayotov et al., 2015). For fine-tuning of the ASR model the “train clean 100” split (100 h) was used. As described previously, we used a mixture of 70\% re-synthesized and 30\% clean utterances, with WGN added to all samples. We performed initial evaluations and tuning of the noise-flooding amplitude on the first 500 samples of the “dev clean” set and used the first 500 samples of the “test clean” split in the final evaluation.

4.2 Adversarial attacks (threat models)

We test four different defenses: melGAN only (MG), WaveGAN only (WG), melGAN with MDNF (MG-MDNF), and WaveGAN with MNDF (WG-MDNF) in addition to the RS baseline. Defenses were evaluated against both targeted and untargeted PGD attacks. We performed experiments with various norms and $\epsilon$/SNR values to investigate the robustness of our methods to different conditions. All PGD attacks used 100 max iterations. To distinguish between the different untargeted PGD scenarios, we will use the notation “norm/epsilon.” Targeted SNR-bounded PGD scenarios will be referred to by the SNR in dB. In addition, we also evaluate the defenses against the CW attack with initial $\epsilon = 0.01$ and 400 iterations with learning rate $1e-4$. We evaluate the defenses using word error rate (WER) between the ground-truth transcription and the transcription produced by the ASR in each scenario. Targeted attacks are additionally evaluated by their “target WER,” which reflects the effectiveness of the adversary in forcing the desired transcription. High values indicate that the adversary is failing to achieve its objective.

Since MDNF is a stochastic method, we implemented an adaptive attack called expectation over transformation (EOT) with the PGD attack in which the adversary averages ASR gradients across multiple inference calls (Athalye et al., 2017). EOT allows the attacker to estimate the expected value of the gradients under the random perturbations introduced by the defense and has been shown to produce robust adversarial examples even in the presence of random transforms.

5. Results

Table 1 shows the results for the melGAN-based defenses against various untargeted PGD attacks, as well as benign performance (denoted by “Ben.”), which considers the case where the ASR system is not under attack. The MG adds some degree of robustness over the undefended model (defense is None), improving WER on the $\ell_2/1.5$ attack by about 3.5\% (see rows 1 and 3). However, the addition of noise flooding (i.e., MG-MDNF, row 4) improves performance substantially, reducing the WER by almost 32\% over MG. MG-MDNF also outperforms the RS baseline in the stronger (i.e., larger $\epsilon$) attack scenarios. For the weaker attack scenario ($\epsilon = 0.5$), MG-MDNF produces slightly worse adversarial performance compared to RS (WER goes up by 1\%). For these experiments, the adversarial perturbations are substantially smaller in magnitude than the noise added during RS. We hypothesize that RS is better able to overwhelm the adversarial perturbations for the high SNR (small $\epsilon$) attacks and, thus, performs slightly better relative to MDNF.

We observe a similar trend in the $\ell_\infty$-norm PGD attack scenarios, where the MG-MDNF method outperforms the RS baseline by a large margin in the high $\epsilon$ scenario, while still retaining competitive performance in the low $\epsilon$ threat model. These results suggest that the proposed MG-MDNF defense shows strong robustness, particularly against high $\epsilon$. 

\[ \text{ARTICLE} \]

\[ \text{asa.scitation.org/journal/jel} \]
attacks. This is despite the fact the MG alone produces only a small improvement in WER over the no-defense scenario, implying that the bulk of the performance gains derive from the noise flooding.

The improved robustness to attack comes at the cost of an increase in benign WER of about 5% (column 1) compared to the undefended scenario. While this increase is not negligible, we believe it represents a reasonable trade-off given the large improvement in adversarial WER. Furthermore, the user has the ability to vary the amount of noise flooding introduced, thus, allowing them to establish a balance between benign and adversarial performance that best fits their particular use case.

We also tested the robustness of the MG-MDNF defense to an adaptive adversary using EOT. We found that by averaging loss gradients over five calls, the adversarial WER on the untargeted PGD increased from 52.4 to 63.3. A similar effect was observed for the $\ell_1/0.01$ attack. Therefore, MG-MDNF still provides a fairly strong defense even against an EOT adversary.

The results for targeted PGD and CW attacks are shown in Table 2. In these experiments, the MG defense produces a greater improvement in WER over the undefended model than it does for the untargeted attacks (Table 1). For example, the WER for the 20 dB PGD attack is reduced by around 14.5% over the undefended model (see rows 1 and 3). Once again, however, MG-MDNF (row 4) does markedly better than either the undefended or MG scenarios, further reducing the undefended WER by 48.5%. As with the untargeted attacks, MG-MDNF performs better than the RS baseline for higher $\epsilon$ (stronger) attacks and about the same for the smaller $\epsilon$ scenarios. Additionally, MG-MDNF substantially improves robustness against CW attacks, implying that the defense generalizes well to a range of threat models.

Table 3 shows the performance of the WG and WG-MDNF defenses against a variety of untargeted PGD attacks. Unlike the melGAN, the WaveGAN appears to provide a robust defense even without noise flooding. For the $\ell_2/1.5$ threat model, the adversarial WER with the WaveGAN alone is 39.2% (row 2). However, the benign performance of the WG defense is slightly worse than that of the MG defense. The addition of noise flooding (WG-MDNF) produces minimal change in adversarial performance, with a minor (1.6%) increase in benign WER (column 1). This may be because the WaveGAN takes a global noise vector (in addition to the mel spectrogram) as input. Therefore, the gradients seen by the adversary are slightly different on each iteration, leading to inherent stochasticity even without noise flooding. In summary, we find that noise flooding substantially improves the adversarial robustness of a relatively poor-performing mel-transform defense (i.e., the melGAN). When applied to an already strong defense, such as the WaveGAN, it largely retains the benign and adversarial performance of the WG defense.

As discussed previously, we found that adversarial performance improved when the noise was shaped along the frequency axis in proportion to the empirical distribution of the adversarial perturbations. Without this noise shaping, i.e., with equal variance noise added to each bin, the MG-MDNF defense produced an adversarial WER of 64.0% compared to
52.4% when using the curve shown in Fig. 1 (\(\ell_2/1.5\) threat model). This particular curve was generated using an \(\ell_2\) PGD attack and seems to perform well across various threat models, including the CW attack. However, our experiments suggest that this shaping curve may not be globally optimal: For the \(\ell_\infty/0.01\) attack, we observed that the unshaped WGN (equal variance) with melGAN significantly outperforms the shaped noise (25.8 vs 65.6 adversarial WER).

We also ran experiments using a short-time Fourier transform (STFT) in place of a mel domain transform. We found that the STFT alone provided essentially no adversarial robustness, but with noise flooding on the frequency domain coefficients, the adversarial WER improved by 26.26% (PGD 1.5 threat model). This supports the hypothesis that the additive noise in the transform domain can act to boost the robustness of a weak defense. However, the absolute defended WER was still worse than either of the MDNF defenses. Therefore, the unique properties of the mel representations, i.e., their lossy nature and perceptual basis, are likely also key factors in our defense’s overall performance.

We found that increasing the number of attack iterations (e.g., max iterations = 250, 500) produced only a minor change, suggesting that this is not a substantive limiting factor to the adversary’s success.

6. Conclusion

This paper introduces MDNF for defense against adversarial attacks on ASR systems, which leverages additive noise in the mel frequency domain followed by GAN-based re-synthesis of the time domain samples. While prior work (Zelasko et al., 2021) has shown that the mel conversion process can itself act as a defense, we find that not all generator models provide the same degree of robustness. We show that MDNF can substantially improve upon the defensive capability of a poorer performing generator, while having minimal impact on one with strong baseline robustness. Therefore, noise flooding provides a quick and easy way to enhance defensive performance without the computational costs required to train a more robust generator from scratch. We demonstrate this result across a variety of untargeted and targeted PGD attacks as well as the targeted CW attack. We additionally find that MDNF is competitive with RS and substantially outperforms RS for stronger, i.e., lower SNR, attacks.

Acknowledgments

This work was supported by DARPA (Grant No. HR00112020009), The USC Amazon Center for Secure and Trusted Machine Learning, and an Annenberg Fellowship (to N.M.).

References and links

1We adopt the term “noise flooding” from Rajaratnam and Kalita (2018).

2https://github.com/twosixlabs/armory.

3https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/art/attacks/evasion/adversarial_asr.py.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Qian, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. (2016). “Deep speech 2: End-to-end speech recognition in English and Mandarin,” in Proceedings of the 33rd International Conference on Machine Learning, June 20–22, New York, pp. 173–182.

Atheya, A., Engstrom, L., Iiyas, A., and Kwok, K. (2017). “Synthesizing robust adversarial examples,” arXiv:1707.07397.

Carlini, N., and Wagner, D. A. (2018). “Audio adversarial examples: Targeted attacks on speech-to-text,” arXiv:1801.01944.

Cohen, J., Rosenfeld, E., and Kolter, Z. (2019). “Certified adversarial robustness via randomized smoothing,” in Proceedings of the 36th International Conference on Machine Learning, June 9–15, Long Beach, CA, pp. 1310–1320.

Davis, S., and Mermelstein, P. (1980). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust. Speech Signal Process. 28(4), 357–366.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). “Explaining and harnessing adversarial examples,” arXiv:1412.6572.
Hussain, S., Neekhara, P., Dubnov, S., McAuley, J., and Koushanfar, F. (2021). “Waveguard: Understanding and mitigating audio adversarial examples,” arXiv:2103.03344.

Jati, A., Hsu, C.-C., Pal, M., Peri, R., AbdAlmageed, W., and Narayanan, S. (2021). “Adversarial attack and defense strategies for deep speaker recognition systems,” Comput. Speech Lang. 68, 101199.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebiisson, A., Bengio, Y., and Courville, A. (2019). “Melgan: Generative adversarial networks for conditional waveform synthesis,” arXiv:1910.06711.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). “Towards deep learning models resistant to adversarial attacks,” arXiv:1706.06083.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). “Librispeech: An ASR corpus based on public domain audio books,” in Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 19–24, South Brisbane, Australia, pp. 5206–5210.

Rabiner, L. R., and Schafer, R. W. (2011). Theory and Applications of Digital Speech Processing (Pearson, New York).

Rajaratnam, K., and Kalita, J. (2018). “Noise flooding for detecting audio adversarial examples against automatic speech recognition,” in 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (IEEE, New York).

Ren, K., Zheng, T., Qin, Z., and Liu, X. (2020). “Adversarial attacks and defenses in deep learning,” Engineering 6(3), 346–360.

Sun, S., Yeh, C.-F., Ostendorf, M., Hwang, M.-Y., and Xie, L. (2018). “Training augmentation with adversarial examples for robust speech recognition,” arXiv:1806.02782.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). “Intriguing properties of neural networks,” arXiv:1312.6199.

Yamamoto, R., Song, E., and Kim, J.-M. (2020). “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” arXiv:1910.11480.

Zelasko, P., Joshi, S., Shao, Y., Villalba, J., Trmal, J., Dehak, N., and Khudanpur, S. (2021). “Adversarial attacks and defenses for speech recognition systems,” arXiv:2103.17122.