Vaccinating SER to Neutralize Adversarial Attacks with Self-Supervised Augmentation Strategy

Bo-Hao Su, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

borrissu@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Speech emotion recognition (SER) is being actively developed in multiple real-world application scenarios, and users tend to become intimately connected to these services. However, most existing SER models are vulnerable against a growing diverse set of adversarial attacks. The degraded performances can lead to dreadful user experiences. In this work, we propose a self-supervised augmentation defense (SSAD) strategy to learn a single purify network acts as a general front-end to neutralize adversarial distortions without knowing the types of attack beforehand. We show that our approach can robustly defend against two different gradient-based attacks at various intensities on the well-known IEMOCAP. Further, by examining metrics of protection efficacy and recovery rate, our approach shows a consistent protection behavior to prevent adverse outcomes and is capable to recover samples that are wrongly-predicted before purification.

Index Terms: speech emotion recognition, adversarial attacks, self-supervised learning, augmentation

1. Introduction

Speech emotion recognition (SER) is being actively deployed in real-world settings, as a result, developing robust SER has naturally become an important research topic. In the past, researchers have mostly worked on mitigating performance drops due to unwanted and unaccounted variations, e.g., small scale training corpus [1], distinct contextualized settings [2], wide ranges of cross-corpus scenarios [3, 4], and even semantic mismatch in emotion labeling [5], etc. Exemplary works include the use of very-deep structures [6, 7], generative models [8, 9, 10], adversarial learning [11], and transfer learning approach [12]. Most of these contemporary methods rely on deep learning of various network complexities to handle issues of non-robust variations to achieve reliability [13].

As these SER models begin to be deployed in diverse scenarios, *intentional* malicious attacks become an emerging robustness issue for model usability [13]. Deep learning-based models are vulnerable to attacks of gradient distortions. These gradient distortions would fatally degrade the model's performance while being spotless, e.g., in domains of anti-spoofing [14], speaker verification [15], and image classification [16, 17]. There has not been similar research in SER that addresses the non-robust issue due to adversarial distortions only until recently [18, 19] as it has become obvious that SERs are vulnerable to these adversarial attacks as well [20]. For instance, Saurabh was one of the first works that propose to impose a regularization term derived from the adversarial training to smooth the model prediction [21]. More recently, another straightforward and effective defense mechanism is to expose the SER model to adversaries at training, e.g., Ren et al. [18] proposed to train a defense model by augmenting the training dataset with attack samples along with a feature similarity loss to safeguard the performance of the trained SER under adversarial attacks.

However, these current SER defend strategies only guarantee protection against a specific (seen) type of adversarial attack at a pre-determined intensity. While achieving promising performances, considering the growing variants of adversarial attacks, where each can come at a different intensity, this combinatorics makes training a defense model for every case impractical. Hence, developing a method that operates without knowing the downstream SER model and the type of attacks beforehand is more desirable. In this work, we propose to use a self-supervised augmentation defense (SSAD) algorithm that learns to neutralize the gradient distortions of speech representation against adversarial attacks without knowing the attack types. First, we augment self-supervised variants of gradient distorted samples, and by learning a purify network to sanitize these distortions, we can use this single purify network as a general-purpose front-end to neutralize adversaries. This concept is similar to developing vaccines, i.e., by engineering variants of viruses (gradient distortions) and invoking corresponding protection (purify network), one simply needs to be vaccinated to be immune to the (gradient-based) virus.

Specifically, at the training stage, we generate adversaries using a self-supervised augmentation (SSA) procedure at each iteration, then optimize the purify network by removing these gradient-based distortions. At the inference stage, we only need to apply the purify network to sanitize the input sample which allows it to be applied generally to different downstream SER models of the same input. According to our experiments, our proposed SSAD achieves consistently robust results under two different types of attack with versatile intensities whereas the recent SOTA model [18] performs well only for the seen type of attack at low attack intensities. Moreover, we further provide an analysis of our SSAD by examining metrics of protection efficacy (reduced risk of adverse outcome) [22] and recovery rate (correcting the pre-purified prediction). In summary, our method maintains consistent SER performance under different adversarial attacks without prior knowledge. It not only shows a significant reduction in the risk of the adverse outcome but also has the ability to correct a significant portion of the originally-wronged prediction.

2. Methodology

In this section, we will describe our proposed SSAD defense strategy. The overall framework is shown in Fig. 1. Here, we divide it into two stages which are the training and the inference stage. To better understand the following details, we first define several major symbols used. The x, x^*_{adv} represent original and training adversarial samples respectively, and F_ϕ stands for a well-trained feature extractor, and P, D denotes the purify network and the discriminator respectively.

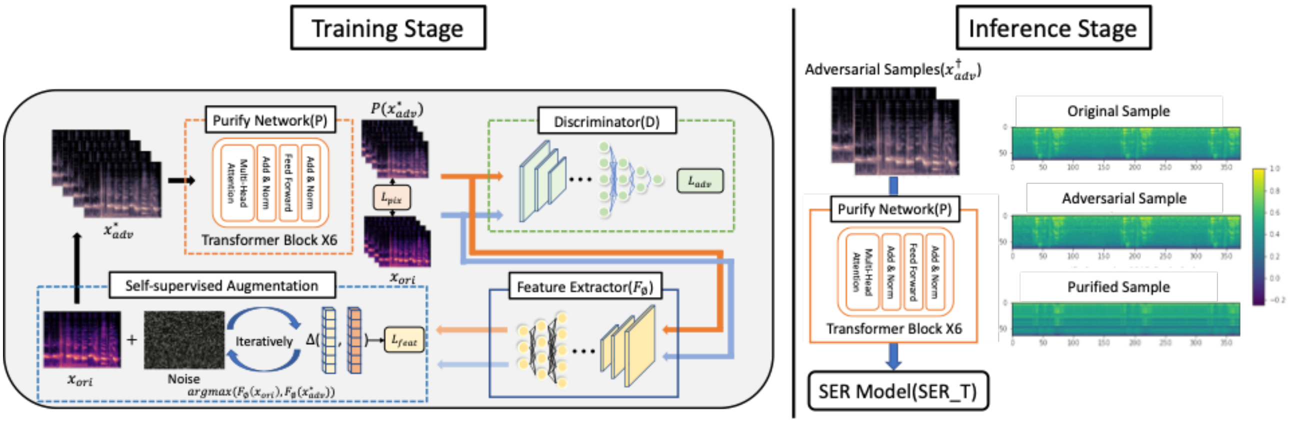

Figure 1: *An overall scheme of our self-supervised augmentation defense strategy (SSAD) depicted at training and inference stages. During training, SSAD is composed of Purify network (P), Discriminator network (D), well-trained feature extractor (F_ϕ) and SSA mechanism, where dash line and solid line represent that parameters are trainable and frozen respectively.*

2.1. Training Stage

2.1.1. Self-Supervised Augmentation (SSA)

Conventional logit-based attacks target the discriminative boundary of a inference model. In this work, our aim is to generate augmented samples without knowing the recognition model. The use of self-supervised method that generates gradient distorted samples has been suggested as an effective mechanism in image classification [23]. We adopt a similar concept to generate adversarial samples relying solely on exploring the representation space. In this work, we use VGG-16 [24] as the Mel-spectrogram representation extractor, F_ϕ. Instead of augmenting the dataset by generating a specific type of attack adversaries to the target inference model, our method encourages broad gradient exploration in the representation space by imposing a process of maximizing the distortion between adversarial and original samples according to the well-trained feature extractor. This exploration of representation space is formulated by the following perturbation procedure.

$$argmax(F_\phi(x_{ori}), F_\phi(x^*_{adv})), s.t. ||x_{ori} - x^*_{adv}|| < \delta \quad (1)$$

where δ is the perturbation budget.

Therefore, the adversarial samples are iteratively generated according to the gradient distortion ∇F_ϕ between x_{ori} and x^*_{adv}. The overall process is defined as:

$$x^*_{adv} = x_{ori} + s \cdot sign(\nabla F_\phi) \quad (2)$$

$$x^*_{adv} = clip(x^*_{adv}, x_{ori} - \delta, x_{ori} + \delta) \quad (3)$$

where s is step size, δ is the perturbation budget. The clip function would map the x^*_{adv} into the range of $[x_{ori} - \delta, x_{ori} + \delta]$. Note that in our training procedure, all the adversaries are regenerated for each iteration simultaneously to ensure the model is optimized iteratively.

2.1.2. Purify Network

After the SSA, the purify network is learned to sanitize these adversaries, i.e., taking the Mel-spectrogram of the adversarial samples to the cleaned one. The purification occurs in the Mel-spectrogram instead of representation space (e.g., VGGs) has advantages that it enables the network to act as a general front end, i.e., it can be applied to any of those SER models with Mel-spectrograms as input. Furthermore, this method does not depend on the specifics of the target SER model. This relaxes the constraint that we need to customize the purify network to handle differences in the number of hidden dimensions used in the target SER model.

Therefore, with paired input of original and adversarial samples, the purify network learns to clean the adversarial inputs. The pixel-wise clean-up between clean and dirty input is applied here. The objective function is defined as:

$$L_{pix} = ||x_{ori} - P(x^*_{adv})||_2 \quad (4)$$

Further, we impose a feature loss constraint to ensure that not only the Mel-spectrogram is cleaned but also the representation is matched from the well-trained feature extractor perspective. The loss is defined as below:

$$L_{feat} = |F_\phi(x_{ori}) - F_\phi(P(x^*_{adv}))| \quad (5)$$

2.1.3. Discriminator

Besides the clean-up and feature loss (described in section 2.1.2), we additionally impose a discriminator to distinguish the purified Mel-spectrogram and the original one. The discriminator encourages the purified sample to be indistinguishable from the real original one that acts as another constraint to properly retain the original sample information; the loss is defined as:

$$L_{adv} = log(D(x_{ori})) + log(1 - D(P(x^*_{adv}))) \quad (6)$$

The overall training objective function (L_{SSAD}) is defined as:

$$L_{SSAD} = \alpha L_{adv} + \beta L_{pix} + \gamma L_{feat} \quad (7)$$

where L_{adv}, L_{pix}, L_{feat} correspond to losses mentioned above, and α, β, γ are tunable weights between each loss.

2.2. Inference Stage

At the inference stage, we consider a SER application scenario where there is an existing recognition model and attackers would hack the model with adversarial samples to sabotage the performance. Under such circumstances, before passing the samples directly through the target SER model, one would vaccinate the sample by running it through the learned purify network as the sanitizer to clean the samples and pass the clean input, e.g., Mel-spectrogram, to the target recognition model. The procedure is formulated as below:

$$x_{clean} = P(x^\dagger_{adv}; \theta) \quad (8)$$

Table 1: *Results are presented in UAR on the target emotion recognition model by using the defense baseline model and our proposed SSAD, and the seen and unseen adversarial type are FGSM and PGD, respectively. The intensity (ε, k) corresponds to the parameter setting of FGSM and PGD as well, and the 'train w.' stands for the baseline models trained with a specific intensity. Not the pre-trained target model has an UAR of 51.06%.*

		Seen Adversarial Type					Unseen Adversarial Type				
	train w.	different intensity (ϵ)					different intensity (k)				
	ϵ	.003	.006	.009	.012	.015	3	6	9	12	15
Adv.[18]	.003	**49.03%**	**45.59%**	42.51%	39.43%	36.73%	32.53%	24.09%	21.72%	21.43%	21.31%
	.006	42.03%	40.28%	38.12%	36.10%	34.29%	31.02%	25.76%	24.71%	24.18%	24.18%
	.009	42.95%	40.40%	38.50%	36.47%	34.98%	32.08%	24.37%	23.02%	22.63%	22.41%
	.012	45.36%	43.44%	41.71%	39.77%	37.86%	35.28%	29.98%	28.57%	28.38%	28.18%
	.015	43.39%	41.78%	40.34%	38.63%	37.34%	35.07%	29.72%	28.53%	28.21%	28.12%
SSAD	-	45.94%	44.82%	**43.54%**	**42.88%**	**40.46%**	**39.10%**	**34.91%**	**33.21%**	**33.54%**	**34.19%**

$$y = SER\_T(x_{clean}; \phi) \tag{9}$$

where P, SER_T represent the trained purify network and target SER model respectively, and θ, ϕ are the corresponding model parameters, and $x^{\dagger}_{adv}$ is the adversarially attacked testing sample, x_{clean} is the sanitized sample, y is the prediction from the target SER model. The overall purify procedure is illustrated in Figure 1, and we also plot the spectrogram of an original, its corresponding adversarial sample, and the sample after being purified as an example.

3. Experimental Results

3.1. Database - IEMOCAP

The IEMOCAP dataset [25] is a well-known SER benchmark that contains five dyadic spoken interaction sessions in total and two actors (one male and one female) are included in each session. We use the four major categorical emotion utterances for our experiment, which contains 5531 utterances in total that includes happy(excited), sad, angry, neutral. In our experiment, the session independent cross-validation scheme is applied, i.e., one speaker in one session would be taken as a testing set and the other is a validation set to decide the early stopping point to optimize the performance of the model.

3.2. Experimental Setup

3.2.1. SER Model and Mel-spectrogram Input

In this work, our target SER model is pre-trained using VGG16-based CNN structure that achieves an UAR 51.06% in the 4 class recognition task. Log Mel-spectrogram is chosen as input that is computed by torchaudio toolkit. To properly compare to the previous work, the settings are set the same as [18]: the window size of the spectrogram is 512 units with an overlap of 256 units, 64-bin Mel filter bank is applied; for those sentences that are longer than 6 seconds (about 24%), the Mel-spectrogram is extracted from the middle parts. All log Mel-spectrograms are fixed to the dimension of (64, 373).

3.2.2. SSAD Setup

We use the transformer layers as our backbone model to implement the purify network. Specifically, we apply 8 multi-headed attention of 6 layers of transformer blocks cascaded, and the input size is set as 64. As for the discriminator, it is composed of 5 convolutional layers with residual connections, followed by an average pooling, and 4 fully connected layers of parameters set as (4096, 1024, 64, 1). The well-trained feature extractor for log Mel-spectrogram is VGG-16 that is further fine-tuned on the IEMOCAP without seeing the testing set in each cross-validation fold.

During the training stage, Adam optimizer is utilized with learning rate 1e-3, and the early stopping depends on the loss of the validation set with a patience setting of 5. The weight setting of α, β, γ are all set to 1. For the adversarial samples augmentation settings, the step size (s) and the perturbation budget (δ) are 0.01 and 0.06.

3.2.3. Baseline Model and Adversarial Attacks

As for the baseline defense method, we compare it to the most recent similarity-based model [18]. This approach trains the SER on an augmented dataset of adversarial samples that are generated by model-based gradient distortions with an additional feature similarity criterion. Since this baseline approach requires a pre-defined attack and intensity, we evaluate the method by training multiple defense models with different intensities of adversarial attacks following the procedure in [18].

For the attacks at the inference, we evaluate defense methods under two common gradient-based adversarial attacks. They are briefly described below:

- Fast Gradient Sign Method (FGSM) [26]

  FGSM compute the gradient trend of the target model and poison the original sample accordingly. In specifics, the attack is defined as:

  $$x^{\dagger}_{adv} = x + \epsilon \cdot sign(\nabla_x L(\theta, x, y)) \tag{10}$$

  where the x, y, $x^{\dagger}_{adv}$ are original input, its label, and its testing adversary respectively. Then ϵ is a disturbing parameter that represents the intensity of the attack, and L is the loss used in the target inference model with parameter θ.

- Projected Gradient Descent (PGD) [27]

  PGD can be seen as an advanced gradient-based attack algorithm. It searches for the optimal gradient smoothly through multiple steps and compute a more targeted and precise noise. The attack is defined as below:

  $$x^{\dagger k+1}_{adv} = \prod (x^{\dagger k}_{adv} + \epsilon \cdot sign(\nabla_x L(\theta, x, y))) \tag{11}$$

  the index k which stands for the step number of PGD and represents the intensity of the attack.

3.3. Result and Analysis

We first evaluate the robustness of the defense method in the setting when experiencing the two types of attack and present UAR

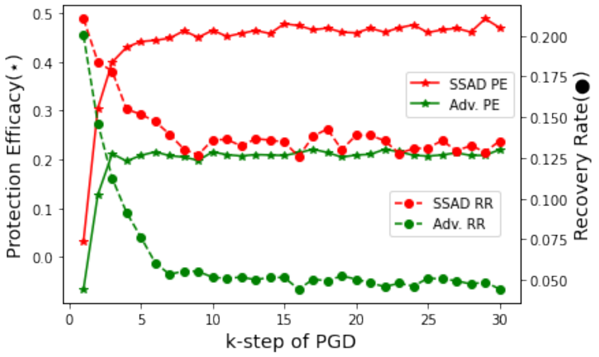

Figure 2: *Protection efficacy (∗) and recovery rate (·) curve of our proposed SSAD (red) and the baseline model (green).*

as metrics. We present two cases since the baseline method (denoted as Adv.) requires knowledge on the type of attacks. We treat FGSM as the "seen" attack and PGD as the "unseen" attack; note that there is no such distinction for our proposed SSAD. We then examine two additional metrics, protection efficacy and recovery rate, to bring additional insights into the defense mechanism.

3.3.1. Seen versus Unseen Type of Adversarial Attacks

In Table 1, we investigate the SER performances by examining combinations of seen and unseen attacks with a range of different intensity levels. The baseline models (Adv.) are trained with different intensities of the seen type (FGSM) listed in the second column. Note that in our proposed SSAD, all the adversarial attacks are unseen while training a purify network.

For the seen adversarial attack, we observe that the baseline method better maintains the performance only under low-intensity adversarial attacks (ϵ-0.003, 0.006, UAR of 49.03% and 45.59%). However, as the attack intensity increases (ϵ-0.009∼0.015), our proposed SSAD outperforms the baseline method and maintains the highest performance even when the Adv. method is trained in the perfectly matched condition (seen type with the same intensity). Specifically, our proposed SSAD achieves an average of 43.53% UAR over a range of intensity attacks which surpasses the best average of baseline models (train w. ϵ-0.003) 1.33% in absolute points (p-value$<$0.05). It's worth mentioning that the performance of our proposed SSAD is quite stable with only σ = 1.86% when compared to the best baseline model that has a high σ = 4.71%.

For the unseen adversarial type, our proposed SSAD achieves an average 34.99% UAR (σ = 2.14%), while the best average performance of baseline models results in 30.08% UAR (σ = 2.68%). A larger drop of performances between unseen and seen adversarial attacks for the baseline models shows that training with a single type of attack can lead to over-fitting. In contrast, our proposed SSAD utilizes self-supervised purify strategy that neutralizes a range set of gradient distortions is shown to provide a better generalization of defense capacity. Furthermore, SSAD only requires a single purify network instead of running many variants of defense models.

3.3.2. Protection Efficacy and Recovery Rate Analysis

To better understand the defense mechanism, we utilize two additional metrics to evaluate our defense model, i.e., protection efficacy and recovery rate. These two metrics are defined as:

$$\textbf{Protection Efficacy (PE)} = \frac{R_{\text{no-protect}} - R_{\text{protected}}}{R_{\text{no-protect}}} \tag{12}$$

where R indicates the risk of adverse outcome, i.e., the percentage of samples that result in the wrong prediction after being attacked when they are originally correctly predicted by SER_T. $R_{\text{no-protect}}$ indicates this percentage without any defense mechanism applied, where $R_{\text{protected}}$ indicated otherwise. PE is a common metric used in assessing vaccination efficacy [22], which is interpreted as the reduced risk of adverse outcome after applying defense in this context.

$$\textbf{Recovery Rate (RR)} = \frac{N_{\text{wrong} \rightarrow \text{correct}}}{N_{\text{wrong}}} \tag{13}$$

where N indicates the number of samples, and RR computes the percentage of the samples that are originally predicted wrongly by SER_T but correctly after applying a defense method. Note that for these two metrics, the higher is better.

In this analysis, we example these two metrics for the SSAD and the baseline Adv under a range of PGD attacks (k=1∼30). The result is presented in Figure 2. We see that SSAD (red solid line) holds a better protection efficacy (avg=0.440, σ=0.083) under various intensities of PGD attacks. The baseline models (green solid line) hold a significantly suboptimal protection efficacy (avg=0.198, σ=0.052). Our proposed SSAD achieves about 45% average reduced risk in turning the original correct samples to the wrong prediction whereas the current best approach obtains only about 20%.

When considering the recovery rate (dashed lines), our proposed SSAD obtains an average of 0.19 (σ=0.032) which is higher than the baseline models which have an average of 0.15 (σ=0.054). It is quite intriguing to see that these defense models not only protect the original SER, but also actively correct the original SER (e.g., SSAD corrects about 20%). This may be due to the usage of the "augmentation-as-defense" strategy, where the expansion on the representation-gradient spaces may indirectly add needed variability to the emotion corpus that is originally missing, i.e., a similar finding in previous works on using other augmentation techniques for mitigating limited corpus variability [28]. These two metrics further demonstrate two aspects of a defense model: the protection efficacy and the recovery rate of our proposed SSAD.

4. Conclusions and Future Works

With a proliferation of SER applications in our daily life, the robustness of the model against malicious attacks becomes an important issue. In this work, we propose a self-supervised augmentation defense (SSAD) model for preventing versatile adversarial attacks. Unlike recent SOTA similarity-based adversarial model training that works only for a specific attack, SSAD shows the generalizability across various intensities of adversarial attacks using a single purify network. Furthermore, the defense by augmentation through exploration in distorted gradient spaces not only provides a robust and better protection efficacy but also shows improvement in indirectly enhancing the target SER model.

In our future effort, we would investigate multiple SER datasets and other downstream speech tasks to evaluate the effectiveness of this purify network. Another direction would include black-box adversarial attacks to generate chaos samples without knowing the information of target recognition models.

5. References

[1] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition." in *INTERSPEECH*, 2019, pp. 2828–2832.

[2] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[3] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5734–5738.

[4] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE signal processing letters*, vol. 23, no. 5, pp. 585–589, 2016.

[5] G.-Y. Chao, Y.-S. Lin, C.-M. Chang, and C.-C. Lee, "Enforcing semantic consistency for cross corpus valence regression from speech using adversarial discrepancy learning." in *INTERSPEECH*, 2019, pp. 1681–1685.

[6] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[7] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Focal loss based residual convolutional neural network for speech emotion recognition," *arXiv preprint arXiv:1906.05682*, 2019.

[8] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 854–860.

[9] 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.

[10] B.-H. Su and C.-C. Lee, "A conditional cycle emotion gan for cross corpus speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 351–357.

[11] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2019.

[12] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition." *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, 2019.

[13] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C. Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, 2021.

[14] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *Proc. Interspeech 2019*, pp. 1033–1037, 2019.

[15] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial regularization for end-to-end robust speaker verification." in *Interspeech*, 2019, pp. 4010–4014.

[16] S. A. Fezza, Y. Bakhti, W. Hamidouche, and O. Déforges, "Perceptual evaluation of adversarial attacks for cnn-based image classification," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–6.

[17] A. Subramanya, V. Pillai, and H. Pirsiavash, "Fooling network interpretation in image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2020–2029.

[18] Z. Ren, A. Baird, J. Han, Z. Zhang, and B. Schuller, "Generating and protecting against adversarial attacks for deep speech-based emotion recognition models," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7184–7188.

[19] Z. Ren, J. Han, N. Cummins, and B. W. Schuller, "Enhancing transferability of black-box adversarial attacks via lifelong learning for speech emotion recognition models." in *INTERSPEECH*, 2020, pp. 496–500.

[20] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *arXiv preprint arXiv:1711.03280*, 2017.

[21] S. Sahu, R. Gupta, G. Sivaraman, and C. Espy-Wilson, "Smoothing model predictions using adversarial training procedures for speech based emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4934–4938.

[22] W. A. Orenstein, R. H. Bernier, T. J. Dondero, A. R. Hinman, J. S. Marks, K. J. Bart, and B. Sirotkin, "Field evaluation of vaccine efficacy." *Bulletin of the World Health Organization*, vol. 63, no. 6, p. 1055, 1985.

[23] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[26] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[28] B. Su and C. Lee, "Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-gan," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–1, jan 2022.