GENERATING AND PROTECTING AGAINST ADVERSARIAL ATTACKS FOR DEEP SPEECH-BASED EMOTION RECOGNITION MODELS

Zhao Ren[1]*, Alice Baird*[1]*, Jing Han*[1]*, Zixing Zhang*[2]*, Björn Schuller*[1,2]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] GLAM – Group on Language, Audio & Music, Imperial College London, UK

zhao.ren@informatik.uni-augsburg.de

ABSTRACT

The development of deep learning models for speech emotion recognition has become a popular area of research. Adversarially generated data can cause false predictions, and in an endeavor to ensure model robustness, defense methods against such attacks should be addressed. With this in mind, in this study, we aim to train deep models to defending against non-targeted white-box adversarial attacks. Adversarial data is first generated from the real data using the fast gradient sign method. Then in the research field of speech emotion recognition, adversarial-based training is employed as a method for protecting against adversarial attack. We then train deep convolutional models with both real and adversarial data, and compare the performances of two adversarial training procedures – namely, vanilla adversarial training, and similarity-based adversarial training. In our experiments, through the use of adversarial data augmentation, both of the considered adversarial training procedures can improve the performance when validated on the real data. Additionally, the similarity-based adversarial training learns a more robust model when working with adversarial data. Finally, the considered VGG-16 model performs the best across all models, for both real and generated data.

Index Terms— Speech Emotion Recognition, Adversarial Attacks, Adversarial Training, Convolutional Neural Network

1. INTRODUCTION

Emotion recognition has become a popular research topic in recent years, particularly as improving interaction between human and machine is an essential part of Artificial Intelligence (AI) research. As a result, systems with integrated speech-based emotion recognition have found many real-life applications, including in Human-robot-interaction (HRI) [1], educational settings [2], and as a diagnosis tool for conditions such as depression [3]. Computational approaches for emotion recognition can be achieved more robustly through multimodal approaches [4, 5]; however, speech alone has shown to be a valuable modality for such a task, due to the array of information transmitted via the speech signal [6].

More recently, deep learning-based methods have been successful for speech-based emotion recognition [7]. However, improving the robustness of deep learning models for real-life implementation is now an important factor in AI research [8]. One aspect of concern for robust development of real-world models is, they are vulnerable to external attacks [8], particularly adversarial attacks [9]. A very minimal, and well designed perturbation of the input – in some cases only a single pixel in an image [10] – can make a deep model fail to predict a class correctly. This is of particular concern to fields wanting to integrate AI for use with sensitive data sources, such as the governmental, finance, or health domains [11], and tampering with speech emotion data sources could lead to destructive and misinterpreted interactions [12]. For instance, an adversarial example attacking an emotion recognition for mental disease diagnosis, could lead to an inaccurate diagnosis or treatment plan.

With this in mind, training a robust model to defend against attacks is necessary. Adversarial-based training for defense (i. e., adversarial training) is one of the state-of-the-art methods to protect against attacks, and is achieved by training a model both on the original input data and adversarially generated (i. e., fake) data. Adversarial training has shown promise for improving model robustness for a range of applications [13, 14], and in this study, we are exploring this method for speech-based emotion recognition focusing specifically on adversarial attacks [9].

There are two main contributions in this paper. First, we make use of adversarial training, not only improving the performance by augmenting the training data, but also defending against the adversarial attacks as it helps the trained model to converge on the adversarially generated data [15]. Further, based on the similarity of high-level features extracted from the original input and fake data, we propose a similarity-based adversarial training approach to improve the robustness to adversarial attacks.

2. RELATED WORK

In the past decade, deep learning topologies have been successfully applied to speech-based emotion recognition tasks [6]. Networks such as Recurrent Neural Networks (RNNs) are able to learn the temporal-based features from emotional speech [16, 17], whilst Convolutional Neural Networks (CNNs) have shown great success for the prediction of emotional classes based on spectrogram image inputs [7, 18]. As well as this, enhancement methods using deep residual networks have also been implemented, as a means of training more robust speech-based emotion recognition models [19].

In regards to the development of more robust deep models, when applied as an input for a deep learning model, adversarially generated data can be a challenge, forcing the model to produce classification errors [9]. Targeted and non-targeted adversarial data were proposed to confuse deep learning models. An approach for generating an adversarial audio example with a targeted label was proposed for speech-to-text systems in [20]. Non-targeted adversarial data are misclassified by a deep learning model without a targeted label [21].

This work was partially supported by the European Union's Horizon H2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 766287 (TAPAS), and the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

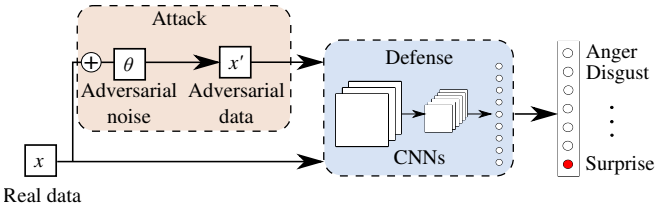

Fig. 1. The framework of our proposed adversarial training approach on emotional speech samples.

During the generating procedure, the adversarial attacks can be classified into white-box and black-box attacks [22]. White-box attacks are obtained when the adversary knows the whole parameters of the deep model; black-box attacks are generated when the adversarial knows only the output of the deep model. In this study, we focus on non-targeted white-box attacks to fail the deep model.

Recently, and with prominence, adversarial attacks have been applied in the computer vision community. For example, adversarial data was generated to fake deep image classification and captioning models [9, 14]. As well as this, CNN models for semantic segmentation and object detection have been shown to be susceptible to attacks by adversarial data [13]. However, to the best of the authors' knowledge, there are no research studies focusing on defending against adversarial attacks in the field of speech-based emotion recognition. An end-to-end scheme was proposed to generate fake emotional speech data in [12], and the validity of using generative networks for emotional speech data augmentation was presented in [23]. However, the corresponding approaches to defend against such adversarially generated attacks was not investigated further by these authors. Hence in this work, we not only propose an adversarial-based data augmentation method, but for the first time, also present two adversarial training approaches for defending speech emotion recognition models against adversarial-based attacks.

3. METHODOLOGY

In this section, we outline both our attack and defense strategies for deep speech emotion recognition models, as shown in Fig. 1. We first introduce our attack method for generating the adversarial (i. e., fake) data, which is used also for data augmentation. This is followed by a description of the adversarial-based training approaches which will be applied to defend against the adversarial attacks. Finally, the employed structures of the deep models will be introduced.

3.1. Adversarial-based Augmentation and Attacks

An adversarial attack is known in the literature as a method to fool a neural network into misclassifying an instance, typically through adding perturbations (i. e., additional noise) to the data [9]. In our case, let us define the input data as $\boldsymbol{x}$, the targets as y, and the learnt parameters of a deep model (e. g., a Covolutional Neural Network (CNN)) as $\boldsymbol{w}$. We first simplify the deep model into a linear function: $y = \boldsymbol{w}\boldsymbol{x}$, and add a slight perturbation to the input data $\boldsymbol{x}$, defining the new input data as $\boldsymbol{x'} = \boldsymbol{x} + \boldsymbol{\theta}$. Then, the function $y = \boldsymbol{w}\boldsymbol{x}$ can be updated to $\boldsymbol{w}\boldsymbol{x'} = \boldsymbol{w}\boldsymbol{x} + \boldsymbol{w}\boldsymbol{\theta}$. With the model going deeper (i. e., additional layers), the model could produce a wrong prediction $\boldsymbol{w}\boldsymbol{x'}$ from $\boldsymbol{w}\boldsymbol{x}$, although $\boldsymbol{\theta}$ is very small. Similarly, non-linear deep models will also be affected by the perturbations. Hence, the generated data $\boldsymbol{x'}$ (i. e., adversarial data) can attempt to fool the deep model through adding adversarial noise to the original real data $\boldsymbol{x}$, as shown in Fig. 1.

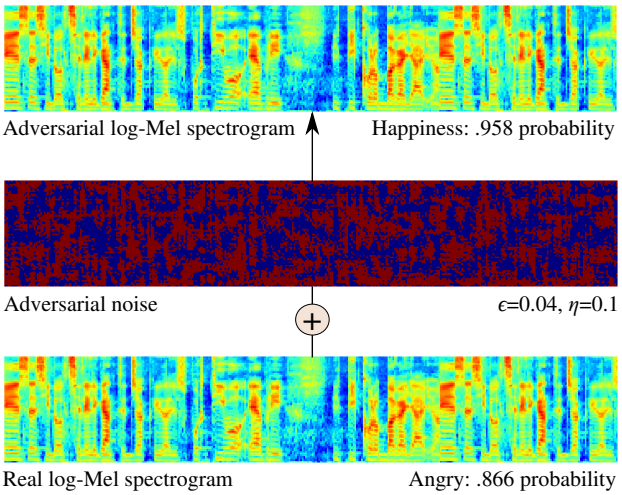

Fig. 2. An adversarial log-Mel spectrogram image generated from the log-Mel spectrogram of the speech sample *NP_m_27_ang07b.wav* from the DEMoS Database, which will be described in Section 4.1.

To generate the adversarial data, we employ the Fast Gradient Sign Method (FGSM) [9], which computes the gradient as the adversarial noise. The loss value during the training procedure is defined by $L(\mathbf{w}, \mathbf{x}, y)$, and the gradient $\nabla_{\boldsymbol{x}} L(\boldsymbol{w}, \boldsymbol{x}, y)$ can be obtained using back propagation. The adversarial data can be computed by

$$\boldsymbol{x'} = \boldsymbol{x} + \epsilon * sign(\nabla_{\boldsymbol{x}} L(\boldsymbol{w}, \boldsymbol{x}, y)), \quad (1)$$

$$\boldsymbol{x'} = clip(\boldsymbol{x'}, \boldsymbol{x} - \eta, \boldsymbol{x} + \eta), \quad (2)$$

where ϵ is a constant perturbation factor, and η is a constant parameter to clip $\boldsymbol{x'}$ into an interval $[\boldsymbol{x} - \eta, \boldsymbol{x} + \eta]$. The generated fake data can help for data augmentation while training the models, and attack a pre-trained model during the validation. As shown in Fig. 2, a deep model can produce the real log-Mel spectrogram image a correct prediction, which is *angry* with a probability of .866. However, it predicts *happiness* with a probability of .958 after adding a slight adversarial noise to the original image in our example.

3.2. Adversarial-based Training for Defense

To defend against the adversarial attacks, we implement two adversarial training architectures – vanilla, and our novel similarity-based adversarial training contribution. Adversarial training aims to train on both real and fake data. Training using the fake data allows for more robust classification results, as fundamentally the training set is larger (a necessity for deep networks) and in turn the augmented data regularises the parameters against more fine-grained differences, allowing for a more robust class prediction.

3.2.1. Vanilla Adversarial Training

Different from the loss function while only training on the real data, the loss function of our vanilla adversarial training considers the loss values of both real and fake data. The loss function is defined by

$$\hat{L}(\boldsymbol{w}, \boldsymbol{x}, y) = \alpha * L(\boldsymbol{w}, \boldsymbol{x}, y) + (1 - \alpha) * L(\boldsymbol{w}, \boldsymbol{x'}, y), \quad (3)$$

where α is a constant parameter to adapt the weights of the loss values on the real and fake data. Hence, this vanilla adversarial training procedure is mainly achieved by minimising the two loss functions $L(\boldsymbol{w}, \boldsymbol{x}, y)$ and $L(\boldsymbol{w}, \boldsymbol{x'}, y)$.

3.2.2. Similarity-based Adversarial Training

As the adversarial data is generated by adding noise to each real input data, a real instance and its corresponding adversarial data can be viewed as a pair, labeled with the same emotion. Therefore, we can assume that, the feature vectors from the final fully connected layer should be close in each of these pairs. With this in mind, we define the following similarity-based adversarial training procedure,

$$\begin{aligned}\hat{L}(\boldsymbol{w}, \boldsymbol{x}, y) = \beta * L(\boldsymbol{w}, \boldsymbol{x}, y) + \gamma * L(\boldsymbol{w}, \boldsymbol{x'}, y) \\ + (1 - \beta - \gamma) * \|\boldsymbol{v} - \boldsymbol{v'}\|_n,\end{aligned} \tag{4}$$

where β and γ are constant parameters, and v and v' are the feature vectors in the final fully connected layer from the real and adversarial data respectively. Here, L2 Loss ($n = 2$) is applied to measure the distance between two feature vectors.

3.3. Deep Convolutional Neural Networks

As we found the use of log-Mel spectrogram images extracted from audio waves to be successful in our previous work [24], we make use of log-Mel spectrogram images as input of the deep models in this study. Herein, three CNN architectures are employed due to CNN models' strong capability to extract high-level features from log-Mel spectrogram images [24]. The implemented CNN models contain a conventional CNN model with four convolutional layers (named as CNN-4), a ResNet model, and a VGG model.

The CNN-4 model contains, four convolutional layers $[64, 128, 256, 512]$, a global max pooling layer, a fully connected layer, and a softmax layer for the final classification. The four convolutional layers have a kernel with a size of $(5, 5)$, and each convolutional layer is followed by a local max pooling layer with a kernel size of $(2, 2)$. The global max pooling has shown better performance than flattening in our previous study [24], as it can extract smaller feature vectors from feature maps for classification.

Besides the CNN-4 model, another two state-of-the-art CNN models, ResNet and VGG, are employed for a comparison. ResNet [25] contains the Inception architecture which requires relatively low computational cost. ResNet has shown promise in the tasks of image processing [25]. ResNet has a series of structures with different numbers of layers. One of them, ResNet-50 [26], is utilised, since our work is focusing on adversarial attacks and training. Moreover, VGG has shown good performance on processing spectrogram images for audio classification tasks due to its deep architecture [27, 28]. Hence, we also train a VGG-16 model for this speech-based emotion recognition task.

4. EXPERIMENTAL RESULTS

4.1. Database

For this study, we utilise the Database of Elicited Mood in Speech (DEMoS) [29], which is an Italian emotional speech corpus. DEMoS was collected from 68 speakers (23 females, 45 males) with 9 365 emotional and 332 neutral speech samples in total. The neutral speech samples are not considered in our study, as *neutral* is a minority class. The 9 365 speech samples are annotated with seven classes of emotion shown in Table 1, of which all are used in our experiments. The emotions of DEMoS were induced by an arousal-valence progression [29]. To avoid speaker dependency during training, partitioning of the data (train, development, and test) was made speaker-independently with consideration to gender and emotional class balancing. The data distribution in the three partitions is described in Table 1.

Table 1. Speaker independent partitions, Train, (Dev)elopment, Test created from DEMoS, including the distribution of the 7-classes as well as gender, (F)emale and (M)ale.

#	Train	Dev.	Test	$\sum$	Gender (F:M)
Speakers	24	22	22	68	23: 45
Anger	492	472	513	1477	516: 961
Disgust	525	556	597	1678	596:1082
Fear	380	383	393	1156	415: 741
Guilt	351	366	412	1129	400: 729
Happiness	447	434	514	1395	524: 871
Sadness	493	486	551	1530	532: 998
Surprise	336	327	337	1000	349: 651
$\sum$	3024	3024	3317	9365	3332:6033

4.2. Experimental Setup

For parameter optimisation during our experiments, we train the CNN models on the training set, and test them on the development set; the CNN models for validation of the test set are trained from the combined training and development set. First, the speech files are resampled from 44.1 kHz to 16 kHz, as the data of 16 kHz can lead to faster progressing, and the data with these two sampling rates have similar results in our early experiments. Then, we extract log-Mel spectrogram images with a window size of 512 units, an overlap with a length of 256 units, and 64 mel bins. To unify the time length of log-Mel spectrogram images, we broadcast the spectrogram images which have shorter time lengths than the longest one, leading to a set of log-Mel spectrogram images with a size of $(373, 64)$. Further, the log-Mel spectrogram images are fed as the input of CNN models. During the training procedure, the 'Adam' optimiser is utilised with a learning rate of .001. After every 100 training iterations, the learning rate is reduced to 90 % percent of its value at the current iteration step, aiming to improve the stabilisation of the training models. Finally, the training procedure is stopped at the 10 000-th iteration.

For the adversarial training, the adversarial data is fed into the training model from the 1 000-th training iteration, and set the hyperparameter as $\alpha = 0.5$ for the vanilla adversarial training. For the similarity-based adversarial training, the hyperparameters are set as $\beta = \gamma = 0.4$. Due to class imbalance, all of the CNN models in our study are evaluated by Unweighted Average Recall (UAR).

4.3. Results and Discussion

To first verify the validity of our proposed adversarial attacks, we train the CNN models on the real training data (i. e., baseline (named as single training)), and then test these models on the adversarial development/test data, as shown in Fig. 3. The performances when $= 0.00$ are the results when testing on the real development/test data. We can see that, all of the six models perform well on the real data at around the UAR of 0.8, while the UAR values are decreasing when the value of ϵ increases. This shows that the adversarial data can be applied to attack the CNN models successfully.

Moreover, the results of the proposed vanilla and similarity-based adversarial training are shown in Table 2. As for the real data, the performance are mostly improved because of data augmentation using the adversarial data. When inferring on the fake data (i. e., adversarial attacks), both of the two proposed training approaches perform well on the adversarial data using the three CNN architectures, although their performance is slightly worse than the performances on the real data. While ϵ is increasing, the performances are becoming worse on the adversarial data. It might imply that a bigger value of

Table 2. Performance comparison of the three CNN topologies, showing results for the training strategies of single, (van)illa (adv)ersarial, and (sim)ilarity-based adversarial training. The CNN models are validated on both (dev)elopment and test set of the real data and fake (i. e., adveresarial) data in DEMoS Corpus.

		CNN-4				**ResNet-50**				**VGG-16**			
	UAR	Real		Fake		Real		Fake		Real		Fake	
NN		*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*
Single Training	.00	.826	.836	–	–	.719	.813	–	–	.798	.836	–	–
Van. Adv. Training	.02	.825	.856	.744	.800	.699	.817	.620	.774	.850	.847	.794	.806
Van. Adv. Training	.04	.817	.871	.657	.749	.813	.850	.685	.755	.849	.855	.743	.783
Van. Adv. Training	.06	.854	.869	.576	.671	.774	.839	.595	.672	.871	.878	.741	.770
Van. Adv. Training	.08	.853	.858	.520	.540	.813	.855	.607	.710	**.875**	**.867**	.709	.756
Van. Adv. Training	.10	.866	.880	.457	.570	.823	.845	.602	.678	.842	.870	.638	.716
Sim. Adv. Training	.02	.844	.797	.827	.784	.743	.798	.732	.771	.847	.823	**.839**	**.821**
Sim. Adv. Training	.04	.822	.825	.772	.769	.708	.798	.652	.759	.814	.842	.806	.820
Sim. Adv. Training	.06	.775	.824	.675	.731	.723	.788	.630	.728	.786	.839	.753	.815
Sim. Adv. Training	.08	.739	.806	.610	.674	.631	.803	.450	.714	.792	.653	.730	.550
Sim. Adv. Training	.10	.727	.752	.526	.585	.407	.734	.316	.498	.804	.833	.716	.769

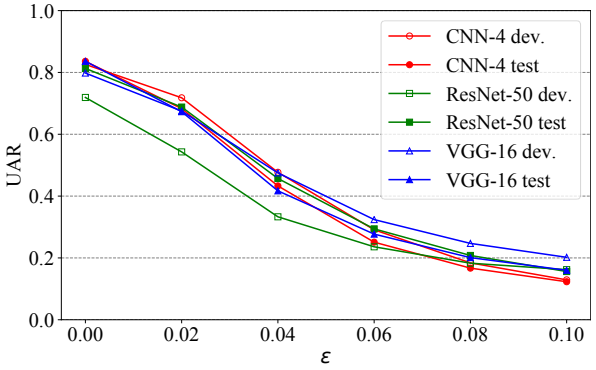

Fig. 3. The performance of the CNN models obtained by single training, while validated on the adversarial (dev)elopment and test data utilising the DEMoS dataset. The adversarial development/test data is equal to the real data while $\epsilon = 0.00$.

Table 3. Performance comparison between our proposed approach and other data augmentation methods using DEMoS, reporting UAR.

UAR	*Dev.*	*Test*
WaveNet (two classes) [23]	.857	.741
Raw audio augmentation by random noise	.795	.833
Spectrogram augmentation by random noise	.808	.833
Our proposed approach	**.875**	**.867**

can affect the data distribution. Furthermore, when comparing the two adversarial training approaches, the vanilla adversarial training performs better on the real data than the similarity-based one in most cases; the similarity-based adversarial training can defend against attacks more effectively than the other. The similarity-based loss can reduce the difference between the features of real and adversarial data, but it affects the feature vectors extracted from the real data.

Among the three CNN models, VGG-16 performs best, whilst ResNet-50 has the worst performance on both real and adversarial data. This might be caused by the architecture of ResNet, containing more convolutional layers in the Inception architecture than the other two CNN models. The performance of CNN models are highly related to the number of layers [30]. Too many convolutional layers might slow down the convergence. While comparing CNN-4 and VGG-16, VGG-16 is more stable and more robust. We think this is because more convolutional layers can extract higher-level features. Further, the difference between real and adversarial data is becoming larger when the model is going deeper, although they are similar as input. Therefore, more convolutional layers can increase such a difference, and help the model learn to reduce the difference. Finally, our best result on the real data (development: .875, test: .867) achieves a significant improvement, compared to that obtained by single training (development: .826, test: .836) (in a one-tailed z-test, $p < .001$).

Further, we compare our results with the state-of-the-art methods for data augmentation in Table 3. WaveNet can achieve a good performance, however this was applied only for two classes (*happiness*, and *sadness*) in [23]. Through additional experiments, we see that our training result also performs significantly better than data (raw audio and log-Mel spectrogram image) augmentation methods using random noise (in a one-tailed z-test, $p < .001$).

5. CONCLUSIONS AND FUTURE WORK

For this study, we proposed a system for training a deep speech emotion recognition Convolutional Neural Network (CNN) model to be robust against adversarial attacks. We applied the vanilla and similarity-based adversarial-based training for defense (i. e., adversarial training), to three deep CNN models, namely CNN-4, ResNet-50, and VGG-16. From these experiments, we found that the model by adversarial training worked better on real data than that by single training due to adversarial-based augmentation. Further, the similarity-based adversarial training produced an improved performance on fake data than the vanilla adversarial training approach.

In future efforts, we will investigate generating black-box fake data (rather than the white-box approach implemented here) for attacking deep learning models. A black-box approach is applied while the attacked model parameters are unknown, and is inherently closer to a real world situation. Moreover, transferring the fake data across deep models will help to validate model robustness. Additionally, to improve the performance when using the adversarially generated fake data, we look to train a detector for recognising this.

6. REFERENCES

[1] Li Zhang, Alamgir Hossain, and Ming Jiang, "Intelligent facial action and emotion recognition for humanoid robots," in *Proc. IJCNN*, Beijing, China, 2014, pp. 739–746.

[2] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, Apr. 2014.

[3] Karla Welch, "Physiological signals of autistic children can be useful," *IEEE Instrumentation & Measurement Magazine*, vol. 15, no. 1, pp. 28–32, Feb. 2012.

[4] Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller, "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 5861–5865.

[5] Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Transactions on Affect Computing*, July 2019, 12 pages.

[6] Björn Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, May 2018.

[7] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, Oct. 2017.

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, "Robust physical-world attacks on deep learning models," in *Proc. CVPR*, Salt Lake City, UT, 2018, pp. 1625–1634.

[9] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, San Diego, CA, 2015, 11 pages.

[10] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, Jan. 2019.

[11] Leong Chan, Ian Morgan, Hayden Simon, Fares Alshabanat, Devin Ober, James Gentry, David Min, and Renzhi Cao, "Survey of AI in cybersecurity for information technology management," in *Proc. TEMSCON*, Atlanta, GA, 2019, 8 pages.

[12] Yuan Gong and Christian Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," in *Proc. DYNAMICS*, San Juan, PR, 2017, 8 pages.

[13] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. ICCV*, Venice, Italy, 2017, pp. 1369–1378.

[14] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," in *Proc. ACL*, Melbourne, Australia, 2018, pp. 2587–2597.

[15] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. ICLR*, Vancouver, Canada, 2018, 20 pages.

[16] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 2227–2231.

[17] Maximilian Schmitt and Björn Schuller, "Deep recurrent neural networks for emotion recognition in speech," in *Proc. DAGA*, Munich, Germany, 2018, 4 pages.

[18] Ziping Zhao, Zhongtian Bao, Yiqin Zhao, Zixing Zhang, Nicholas Cummins, Zhao Ren, and Björn Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, July 2019.

[19] Andreas Triantafyllopoulos, Gil Keren, Johannes Wagner, Ingmar Steiner, and Björn Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1691–1695.

[20] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. SPW*, San Francisco, CA, 2018, pp. 1–7.

[21] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. ICLR*, Toulon, France, 2017, 14 pages.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICML*, Sydney, Australia, 2017, 10 pages.

[23] Alice Baird, Amiriparian Shahin, and Björn Schuller, "Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation," in *Proc. MMSP*, Kuala Lumpur, Malaysia, 2019, 5 pages.

[24] Zhao Ren, Qiuqiang Kong, Kun Qian, Mark Plumbley, and Björn Schuller, "Attention-based convolutional neural networks for acoustic scene classification," in *Proc. DCASE*, Surrey, UK, 2018, pp. 39–43.

[25] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, San Francisco, CA, 2017, pp. 4278–4284.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, 2016, pp. 770–778.

[27] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, 2015, 10 pages.

[28] Zhao Ren, Nicholas Cummins, Vedhas Pandit, Jing Han, Kun Qian, and Björn Schuller, "Learning image-based representations for heart sound classification," in *Proc. DH*, Lyon, France, 2018, pp. 143–147.

[29] Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Maximilian Schmitt, and Björn Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, pp. 1–43, Feb. 2019.

[30] Tara Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 8614–8618.