Adversarial Machine Learning And Speech Emotion Recognition: Utilizing Generative Adversarial Networks For Robustness

Siddique Latif[1,2], Rajib Rana[2], and Junaid Qadir[1]

[1]Information Technology University (ITU)-Punjab, Pakistan
[2]University of Southern Queensland, Australia

Abstract

Deep learning has undoubtedly offered tremendous improvements in the performance of state-of-the-art speech emotion recognition (SER) systems. However, recent research on adversarial examples poses enormous challenges on the robustness of SER systems by showing the susceptibility of deep neural networks to adversarial examples as they rely only on small and imperceptible perturbations. In this study, we evaluate how adversarial examples can be used to attack SER systems and propose the first black-box adversarial attack on SER systems. We also explore potential defenses including adversarial training and generative adversarial network (GAN) to enhance robustness. Experimental evaluations suggest various interesting aspects of the effective utilization of adversarial examples useful for achieving robustness for SER systems opening up opportunities for researchers to further innovate in this space.

1 Introduction

Recent progress in machine learning (ML) is reinventing the future of intelligent systems enabling plethora of speech controlled applications [1, 2, 3]. In particular, the emotion-aware systems are on the rise. The breakthrough in deep learning is largely fueling the development of highly accurate and robust emotion recognition systems [4, 5].

Despite the superior performance of deep neural networks (DNNs), recent studies demonstrate that DNNs are highly vulnerable to the malicious attacks that use *adversarial examples*. Adversarial examples are developed by malicious adversaries through the addition of unperceived perturbation with the intention of eliciting wrong responses from ML models. These adversarial examples can debilitate the performance of image recognition, object detection, and speech recognition models [6]. Adversarial attacks can also be used to undermine the performance of speech-based emotion recognition (SER) systems [7], putting security-sensitive paralinguistic applications of SER systems at high risk.

In this paper, we aim to investigate the utility of adversarial examples to achieve robustness in speech emotion classification to adversarial attacks. We consider a “black-box” attack that directly perturbs speech utterances with small and imperceptible noises. The generated adversarial examples are utilized within different schemes highlighting different aspects of robustness of SER systems. We further propose a GAN-based defense for SER systems and show that it can better resits adversarial examples compared to the previously proposed defense solutions such as adversarial training and random noise addition.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Background Literature and Motivation

Existing methods of adversarial attacks including fast gradient sign method (FGSM) [8], Jacobian-based saliency map attack (JSMA) [9], DeepFool [10], and Carlini and Wagner attacks [11] compute the perturbation noise based on the gradient of targeted output with respect to the input. This is computed using backpropagation with the implicit assumption that the attacker has complete knowledge of the network and its parameters (such methods are called *white-box* attacks). While the backpropagation method, which needs to compute the derivative of each layer of the network with respect to the input layers, can be efficiently applied in image recognition due to the differentiability of all layers, the application of such methods is difficult for SER systems since these systems rely on complex acoustic features of the input audio utterances—such as Mel Frequency Cepstral Coefficients (MFCCs), spectrogram, extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [12]. The SER system's first layer is the pre-processing or the feature extraction layer, which does not offer an efficient way to compute derivative, therefore, gradient-based methods [9, 10, 11, 13] are not directly applicable to SER systems.

Adversarial attacks on ML have provoked an active area of research that is focusing on understanding the adversarial attack phenomenon [14] and on techniques that can make ML models robust [15]. For speech-based systems, Carlini [6] proposed a white-box iterative optimization-based attack for DeepSpeech [16], a state-of-the-art speech-to-text model, with 100% success rate. Alzantot et al. [17] proposed an adversarial attack on speech commands classification model by adding a small random noise (background noise) to the audio files. They achieved 87% success without having any information of the underlying model. Song et al. [18] proposed a mechanism that directly attacks the microphone used for sensing voice data and showed that an adversary can exploit the microphone's non-linearity to control the targeted device with inaudible voice commands. Gong et al. [7] presented an architecture to craft adversarial examples for computational paralinguistic applications. They perturbed the raw audio file and were able to cause a significant reduction in performance. Various other studies [19, 20, 21] have also developed adversarial attacks for speech recognition system. However, most of the previous research on targeted attacks for speech-based applications [6, 7, 17, 18, 19, 20, 21] has considered attacks on the model without investigating how adversarial examples may be utilized to make the ML models more robust. Our work is different since we not only propose an adversarial attack for SER system using adversarial examples but also leverage adversarial examples for making ML models more robust.

3 Proposed Audio Adversarial Examples

In this work, we adopt a simple approach to prepare adversarial examples by adding imperceptible noise (δ) to the legitimate samples. We take an audio utterance x with label y, and generate an adversarial example $x^{'} = x + \delta$ such that the SER system fails to correctly classify the given input while ensuring that x and $x^{'}$ are very similar when perceived by humans. Previous speech-related studies have mostly considered "non-real world" random noise as adversarial noise. DolphinAttack exploits inaudible ultrasounds as adversarial noise to control the victim device inconspicuously but the attack sound was out of the human perception. Similarly, Alzantot et al. [17] used random noise for creating an adversarial attack on speech recognition. It was however observed that the state-of-the-art classifiers are relatively robust to random noise [14]. We therefore propose a black-box attack for SER system where an adversary can add "real-world" noise as adversarial perturbation. We empirically show that speech samples imputed with real-world noise can fool the classier while not being perceptible to the human ear.

Generation of δ:

We use three noises: café, meeting, and station from the Demand Noise database [22] and their imputation level is based on the already existing background noises (microphone noise and discussion noise) in the utterances. We estimate the existing noise in utterances using a well-known technique proposed in [23] that estimate noise using spectral and log-amplitude. We make the mean and variance of the above three noises equal to that of the reference noise. We also use ϵ as the variation parameter to further control the extent of perturbation and added the perturbation noise δ to the utterances using $(x_i + \epsilon \times \delta_i)$. Where x_i is i^{th} utterance and δ_i is generated noise for it. In this way, the adversarial noise has a very small value similar to the existing noise and the adversarial

Table 1: Binary class mapping of different emotions

Dataset	Positive class	Negative Class
IEMOCAP	happiness, exited, neutral	anger, sadness
FAU-AIBO	neutral, motherese, and joyful	angry, touchy, reprimanding, and emphatic

example is unrecognizable to the human ear in the human perception test. Because this noise acts as the background noise it does not change the emotional context of a given audio file.

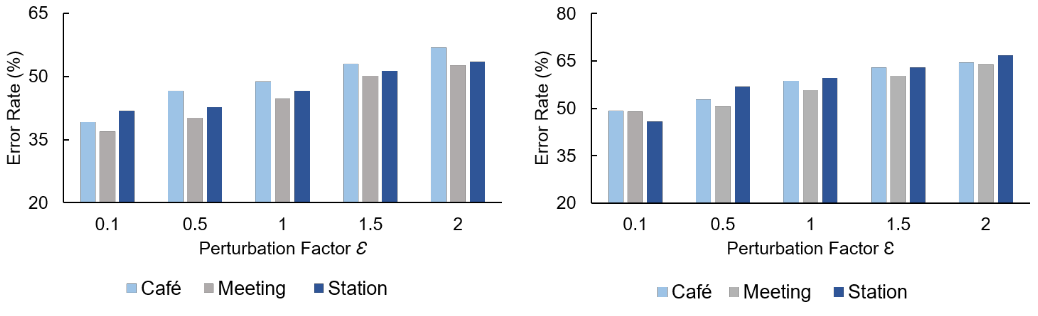
Human Perception and Classifier Test: In order to assess the effect of added adversarial noise on the human listener, we asked five adults (age: 23-30 years) listeners to listen to 200 adversarial examples for different perturbation factor (ϵ) and differentiate it from the original audio file. For the IEMOCAP and FAU-AIBO datasets, 96% and 91% of the samples were indistinguishable from the original utterances. When these examples were given to the classifier, the attack success rate was 72% and 79% for IEMOCAP and FAU-AIBO, respectively.

4 Experimental Setup and Results

We evaluated the generated adversarial examples using two well-known emotional corpora: IEMOCAP and FAU-AIBO. We consider binary classification problem (Positive and Negative) by mapping emotion to binary valance classes as used in [4] and [24]. Table 1 shows the considered emotions and their binary class mapping for both these datasets. We use the eGeMAPS features, a popular features set specifically suited for paralinguistic applications, for representing the audio samples.

Classification Model: We consider LSTM-RNN for emotion classification. LSTM is a popular RNN and widely employed in audio [25] and emotion classification [5] due to their ability to model contextual information. We find the best model structure by evaluating different number of layers. We obtained the best results with two LSTM layers, one dense layer, and a softmax as the last layer. We initially used a learning rate of 0.002 to start training the model and halved this rate after every 5 epochs if performance did not improve on the test set. This process stopped when the learning rate reached below 0.00001.

Emotion Classification Results: For experimentation, we evaluated the model in a speaker independent scheme. IEMOCAP dataset consists of five sessions; we used four session for training and one for testing, consistent with the methodology of previous studies [4, 5]. For FAU-AIBO, we followed the speaker-independent training strategy proposed in the 2009 Interspeech Emotion Challenge [24]. For emotion classification on legitimate examples, we achieved 68.35% and 56.41% unweighted accuracy (UA) on FAU-AIBO and IEMOCAP dataset, respectively. The results on adversarial examples are compared with these results. We generated adversarial examples with different values of ϵ (0.1–2) to evaluate the performance of model with different perturbation factor. This is demonstarted in Figure 1 presents the emotion classification error on adversarial samples with different values of ϵ.



(a) (b)

Figure 1: The error rate (%) with different perturbation factors for speech emotion classification for FAU-AIBO (left) and IEMOCAP (right) datasets.

Based on Figure 1 the proposed attack is effective in fooling the classifier for emotion classification tasks. With the perturbation factor 2.0, the classification error rate is increased from 31.65 and 43.59 to 56.87 and 66.87 for FAU-AIBO and IEMOCAP dataset respectively.

5 Defense Mechanisms

5.1 Training with Adversarial Examples

Adversarial training of model is considered as a possible defense to adversarial attacks when the exact nature of the attack is known. Model training on the mixture of clean and adversarial examples can somewhat help regularization [26]. Training on adversarial samples is different from data augmentation methods that are performed based on the expected translations in test data. To the best of our knowledge, adversarial training is not explored for SER systems and other speech/audio classification systems. We explore this phenomenon by mixing adversarial examples with training data to highlight the robustness of model against attack. We trained the model with training data comprising of a varying percentage of adversarial examples (10% to 100% of training data). Figure 2 shows the classification error rate (%) significantly decreases with the increase of percentage of adversarial examples in the training data.

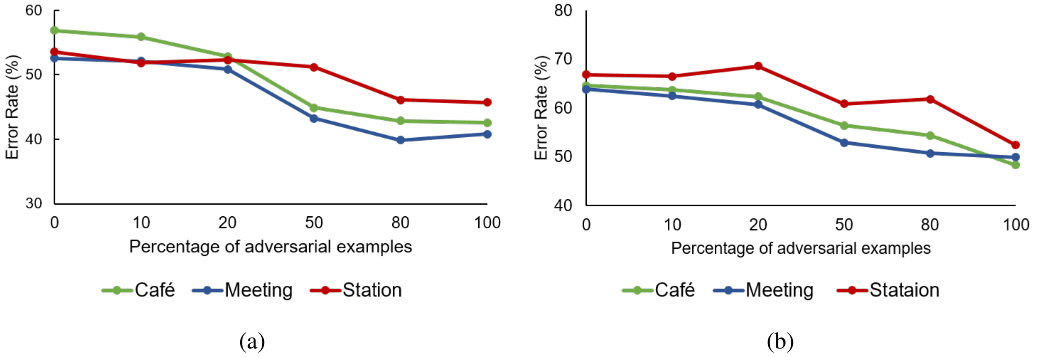


Figure 2: The error rate (%) with varying the percentage of adversarial samples as training data for FAU-AIBO (left) and IEMOCAP (right) datasets.

5.2 Training with Random Noise

It is reported in [27] that the addition of a random noise layer to the neural network can prevent strong gradient-based attacks in the image domain. We evaluated this phenomenon in speech emotion classification system by adding a small random noise to overall training data and evaluated the performance against the proposed attacks. Table 2 shows that emotion classification error reduces only slightly with the addition of random noise in training data, which indicates that this strategy is not particularly effective in the SER settings.

Table 2: Emotion classification error (%) while adding random noise in training data

Dataset	Adversarial Perturbations	Error (max) with adversarial attack	Error by training with random noise
FAU-AIBO	Café	56.87	54.02
	Meeting	52.58	49.24
	Station	53.57	48.51
IEMOCAP	Café	64.58	56.73
	Meeting	63.88	52.57
	Station	66.87	60.87

5.3 Using Generative Adversarial Network

Generative adversarial networks (GANs) [28] are deep models that learn to generate samples, ideally indistinguishable from the real data x, that are supposed to belong to an unknown data distribution, $p_{data}(x)$. GANs consist of two networks, a generator (G) and a discriminator (D). The generator network (G) maps latent vectors from some known prior p_z to samples and discriminator tasked to differentiate between the real sample x or fake $G(z)$. Mathematically, this is represented by the following optimization program:

$$\min_{G} \max_{D} \quad \mathrm{E}_x[\log(D(x))] + \mathrm{E}_y[\log(1 - D(G(z)))] \tag{1}$$

where G and D play this game to fool each other using this min-max optimization program. In our case, G network is tasked to remove the adversarial noise from the adversarial examples z. The G network is structured like an autoencoder using LSTM layers. In the G network, the encoder part compresses the contextual (emotional) information of the input speech features and the decoder uses this representation for reconstruction. The D network follows the same encoder-decoder architecture. For training G and D for different possible scenarios, we used the training data from both the datasets to train the GAN. For each G step, the discriminator was updated twice. For faster convergence, we pretrained the G network in each case. We trained GAN using RMSProp optimizer with learning rate 1×10^{-4} and batch size of 32, until convergence. For training we used utterances corrupted by the three adversarial noises: café, meeting, station as noisy data and it was tasked to clean the utterances. Data cleaned by GAN was given to the classifier for emotion classification.

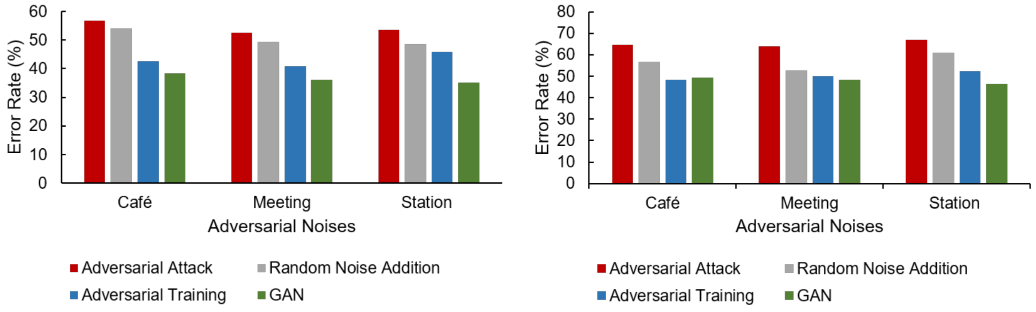
Table 3 shows emotion classification results on audio utterances cleaned by GAN. It can be noted that the classification error significantly reduces by removing adversarial noise from data prior to classification.

Table 3: Emotion classification error (%) by utilizing GAN as defense against adversarial noise removal

Dataset	Adversarial Perturbations	Error (max) with adversarial attack	Error by employing GAN prior to classification
FAU-AIBO	Café	68.82	38.31
	Meeting	62.58	36.02
	Station	66.87	35.14
IEMOCAP	Café	65.87	49.20
	Meeting	67.70	48.18
	Station	69.87	46.24

6 Discussion

From the experimental evaluations we find that GAN-based defense against adversarial audio examples better withstands adversarial examples compared to other approaches. Figure 3 shows a comparison of the different defense mechanisms using two well-known datasets: IEMOCAP and FAU-AIBO. The addition of random noise in training utterances slightly reduces speech emotion classification error, however, using adversarial training, classification error is significantly reduced. This supports that training with random noise is not adequate to avoid adversarial attacks. The best results are however achieved using GAN. This motivates further research for its utilization in other speech-based intelligent systems for the minimization of adversarial perturbations. It is worth



(a)

(b)

Figure 3: The error rate (%) with three different approaches against adversarial examples for FAU-AIBO (left) and IEMOCAP (right) datasets.

pointing out that GANs require information about the exact type and nature of adversarial examples for its training, but this is also an essential requirement for the adversarial training mechanism.

7 Conclusions

In this paper, we propose a black-box method to generate adversarial perturbations in audio examples of speech emotion recognition system (SER). We also propose a defence strategy using Generative Adversarial Network (GAN) for enhancing the robustness of SER system by first cleaning the perturbed utterances through GANs and then running a classifier on it. We compared our GAN-based defense against adversarial training and the addition of random noise in training examples and showed that our GAN-based defense provides consistently better results in speech emotion recognition. We anticipate that the attack and defense that we propose can also be utilized more generally for other speech-based intelligent systems.

References

[1] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[2] Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104–116, 2015.

[3] Rajib Rana. Poster: Context-driven mood mining. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*, pages 143–143. ACM, 2016.

[4] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. Transfer learning for improving speech emotion classification accuracy. *Proc. Interspeech 2018*, pages 257–261, 2018.

[5] Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. In *Proc. Interspeech 2018*, pages 3107–3111, 2018.

[6] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint arXiv:1801.01944*, 2018.

[7] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280*, 2017.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples (2014). *arXiv preprint arXiv:1412.6572*.

[9] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[11] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[12] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[13] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.

[14] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016.

[15] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. *arXiv preprint arXiv:1704.08847*, 2017.

[16] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[17] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

[18] Liwei Song and Prateek Mittal. Inaudible voice commands. *arXiv preprint arXiv:1708.07238*, 2017.

[19] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 2–14. ACM, 2017.

[20] Dan Iter, Jade Huang, and Mike Jermann. Generating adversarial examples for speech recognition. `http://web.stanford.edu/class/cs224s/reports/Dan_Iter.pdf`, 2017.

[21] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

[22] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, 2013.

[23] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[24] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[25] Siddique Latif, Muhammad Usman, Rajib Rana, and Junaid Qadir. Phonocardiographic sensing using deep learning for abnormal heartbeat detection. *IEEE Sensors Journal*, 2018.

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[27] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. *arXiv preprint arXiv:1712.00673*, 2017.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.