On the Robustness of Speech Emotion Models to Black-box Adversarial Attack

Jinxing Gao
Ningbo University
Diqun Yan (yandiqun@nbu.edu.cn)
Ningbo University
Mingyu Dong
Ningbo University

Research Article

Keywords: Convolutional Neural Network, robustness, speech emotion recognition, adversarial attack, adversarial training

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1549399/v1

On the Robustness of Speech Emotion Models to Black-box Adversarial Attack

Jinxing Gao, Diqun Yan* and Mingyu Dong

College of Information Science and Engineering, Ningbo University, Ningbo, 315211, Zhejiang, China.

Contributing authors: 2011082290@nbu.edu.cn; yandiqun@nbu.edu.cn; 1253173217@qq.com;

Abstract

Speech emotion recognition is a key branch of affective computing. Nowadays, it is common to detect emotional diseases through speech emotion recognition. Various detection methods of emotion recognition, such as LTSM, GCN, CNN, show excellent performance. However, due to the robustness of the model, the recognition results of the above models will have a large deviation. So in this article, we use black boxes to combat sample attacks to explore the robustness of the model. After using three different black-box attacks, the accuracy of the CNN-MAA model decreased by 69.38% at the best attack scenario, while the *WER* of voice decreased by only 6.24%, indicating that the robustness of the model does not perform well under our black-box attack method. After adversarial training, the model accuracy only decreased by 13.48%, which shows the effectiveness of adversarial training against sample attacks.

Keywords: Convolutional Neural Network, robustness, speech emotion recognition, adversarial attack, adversarial training

1 Introduction

Machine recognition of emotional content in speech is crucial in many human-centric systems, such as behavioral health monitoring and empathetic conversational systems. Speech emotion recognition is the simulation of human emotion perception and understanding process by computer. Its task is to extract the acoustic features expressing emotion from the collected speech signals, and find the mapping relationship between these acoustic features and human emotion. Therefore Speech

Emotion Recognition (SER) in general is a challenging task due to the huge variability in emotion expression and perception across speakers, languages and culture.

Many SER approaches follow a two-stage framework, In this framework, a set of Low-Level Descriptors (LLDs) are first extracted from raw speech. Then the LLDs are fed to a deep learning model to generate discrete (or continuous) emotion labels [1, 2, 3, 4]. While the use of handcrafted acoustic features is still common in SER, lexical features [5, 6] and log Mel spectrgrams are also used as input [7]. Spectrograms are often used with Convolutional Neural Networks (CNNs) [7] that does not explicitly model the speech dynamics. Explicit modeling of the temporal dynamics is important in SER as it reflects the changes in emotion dynamics [8]. The deep learning model of time series shows excellent performance in this regard, such as Long-Short Term Memory networks (CNN-LSTM), Graph Convolution Network (GCN), Convolutional Neural Networks with Multiscale Area Attention (CNN-MAA) and et al. [9, 10, 11]. The above models are very outstanding in capturing the temporal dynamics of emotion, and their performance effect in SER is the best so far.

Despite their outstanding performance accuracies in SER, recent research has shown that neural networks are easily fooled by malicious attackers who can force the model to produce wrong result or to even generate a targeted output value. And the robustness of SER models against intentional attacks has been largely neglected. However, understanding the robustness against intentional attacks is important for the following reasons: (i) The speech privacy protection method was migrated to the field of speech emotion recognition for black-box adversarial attack; (ii) If the speaker itself has emotional problems and does not want his own voice to be analyzed and used to explore privacy, such an operation can protect his own privacy, while the interference to the original signal content is minimal and imperceptible. The former is regarded as a defense against speech emotion recognition attacks, while the latter is regarded as a protection of speaker emotional privacy to prevent privacy leakage.

To solve these problems, there are gradient-based adversarial attack methods to enhance model robustness, such as FGSM [12], PGD [13], etc., but such methods require the attacker to understand the structure and parameters of the original recognition model. We also need to train alternative models. However, we found that the method of spectral envelope distortion, which is common in the field of speech privacy protection, can play a good adversarial attack effect in speech emotion recognition system. These methods do not need to understand the original recognition model and additional corpus, and can be used for black box attack without training substitutive models. In this paper, we use McAdams transformation, Vocal Trace Length Normalization (VTLN) and Modulation Spectrum Smoothing (MSS) [14] to explore the impact on the current

advanced SER system. To our knowledge, this is the first work that investigates adversarial examples for the field of speech emotional recognition.

The contributions of this work are summarized as follows: (i) We have migrated voice privacy protection methods for use in the field of voice emotional recognition to black-box adversarial attack; (ii) We use the above methods to adversarial attacks against SER and summarize the results to get the best performing hyperparameters α; (iii) We are the first to propose black-box adversarial attack methods to analyze the robustness of the SER models.

Firstly, Section 2 introduces three different SER models (CNN-LSTM, GCN, CNN-MAA) studied in this paper. Then Section 3 will show three speech transformations (McAdams, VTLN, MSS). Section 4 will present the experimental setup and results, Finally, Section 5 concludes the article.

2 Related Work

This section reviews three advanced speech emotion recognition models, which are used as test models in the subsequent parts.

2.1 SER based on CNN-LSTM

Speech emotion recognition is a challenging task and heavily depends on hand-engineered acoustic features, which are typically crafted to echo human perception of speech signals. However, a filter bank that is designed from perceptual evidence is not always guaranteed to be the best in a statistical modeling framework where the end goal is, for example, emotion classification. However a combination of Convolution Neural Networks (CNNs) and Long Short Term Memory (LSTM) have gained great traction for the intrinsic property of LSTM in learning contextual information crucial for emotion recognition. Among them, CNNs can overcome the scalability problem of conventional neural networks. In [9], Siddique Latif et al. proposed the use of parallel convolutional layers to harness multiple temporal resolutions in the feature extraction block, which is jointly trained with an LSTM-based classification network for emotion recognition tasks and achieved better performance results.

2.2 SER based on GCN

Amir Shirian et al. proposed a deep graph approach to address the task of speech emotion recognition [10]. Following the theory of graph signal processing, compared with the traditional CNN network, modeling speech signal as cyclic graph or line graph is a more compact, efficient and scalable form. Meanwhile, compared with the traditional graph structure, the graph with simple structure proposed by the author greatly simplifies the convolution operation on the graph and is not weighted on the edge of the graph, so the parameters that can be learned in SER task are significantly reduced, and its performance is better than LSTM, standard GCN and other latest graph models in SER.

2.3 SER based on CNN-MAA

In SER, emotional characteristics often appear in various forms of energy patterns in spectrograms.

Typical attention neural network classifiers of SER are usually optimized on a fixed attention granularity. Mingke Xu et al [11]. apply multiscale area attention in a deep convolutional neural network to attend emotional characteristics with varied granularities and therefore the classifier can benefit from an ensemble of attentions with different scales. Meanwhile, vocal tract length perturbation is used for data enhancement to improve the generalization ability of the classifier. Compared with other emotion recognition models, more advanced recognition results are obtained.

3 Adversarial Attacks Based on Speech Envelope Distorition

In speech privacy protection, there have been many methods of spectral envelope distortion to protect the personal information contained in speech. Our research found that the speech after spectral envelope distortion has a very excellent adversarial attack effect on the speech emotion recognition system trained by the original speech. Here, we here describe signal processing-based methods that are a part of our voice modification module, each method has an individual scalar hyper parameter α_*. By adjusting the α_*, we explore the adversarial attacks against the SER model and the robustness of the model. Fig 1 shows the flow of black box attack on SER model. After the attack, in order to test the robustness of the model, we use the adversarial training method to add the samples generated by the attack to the training samples, use the original label as the correct label for training, and use the trained model to identify the normal samples and countermeasure samples.

3.1 Vocal Tract Length Normalization

Vocal tract length normalization (VTLN) is mainly used in speech-to-text recognition to remove distortion caused by the difference of length in vocal tracts [15]. This method modifies the amplitude spectrum of the original speech, according to the warping function. Let $\omega_0 \in [0, 1]$ and $\omega_1 \in [0, 1]$ be an original normalized frequency and warped one, respectively. $\omega_0 = 1$ corresponds to the Nyquist frequency. ω_0 is warped into ω_1 as

$$\omega_1 = \pi\omega_0 + 2\tan^{-1}\frac{\alpha_{vtln}\sin(\pi\omega)}{1-\alpha_{vtln}\cos(\pi\omega)} \quad (1)$$

where $\alpha_{vtln} \in [-1, 1]$ denotes the warping factor. Fig 2 shows this warping. When $\alpha_{vtln} < 0$ and $\alpha_{vtln} > 0$, the warping curves (left in the figure) become convex and concave, respectively. The parameters contract and expand an amplitude spectrum, respectively. When $\alpha_{vtln} = 0$, $\omega_0 = \omega_1$, which means no warping. In this paper, we warp frequencies of the log amplitude spectra obtained by applying short-time Fourier transform (STFT) to the original speech. The transformed speech is obtained by inverse STFT of the modified amplitude spectrogram and original phase spectrogram.

3.2 McAdams Transformation

McAdams transformation [16] modifies the resonance frequencies (i.e.,

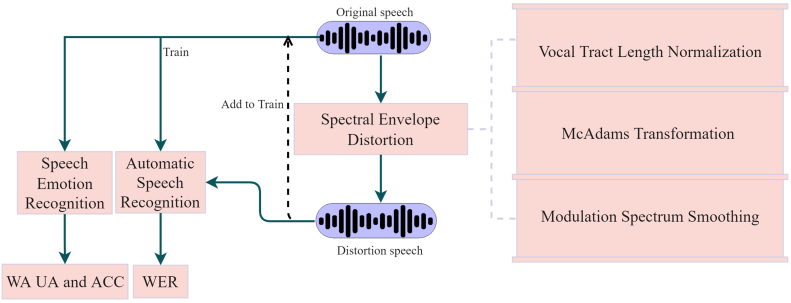

Fig. 1 Adversarial attacks against speech emotion models flow chart

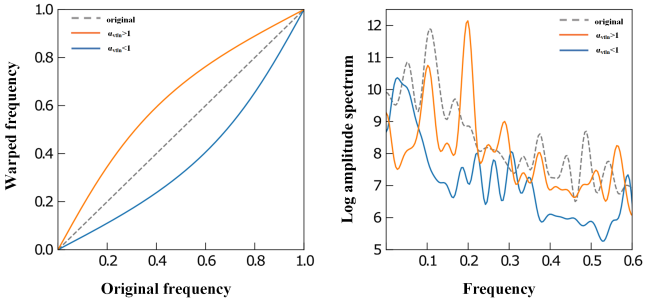

Fig. 2 Examples of VTLN: Description of the warping function (left) and the change in the spectral curve after application (right).

formant frequencies) of the original speech [16]. We obtain N poles [p_1,...,p_n,...,p_N] by taking linear predictive coding (LPC) [17] to original speech. Pole p_n $\in$is written as $A_n \exp(\theta_n)$ in the polar coordinate, where $A_n \in [0,1]$ and $\theta_n \in [0,\pi]$ are the absolute value and phase in the upper half of z-plane, respectively (Poles in the lower half of z-plane are the complex conjugate.). The pipeline of the McAdams transformation approach is shown in Fig 3.

They correspond to the formant strength and frequency, respectively. The imaginary unit is represented as j. McAdams transformation obtains the modified formant frequency $\theta_n^1 \in [0,\pi]$ by $\theta_n^1 = \theta_{n\,mas}^{\alpha}$, where $\alpha_{mas} \in R_+$ is the McAdams coefficient. The $n-th$ modified pole p_n^1 consists of the original formant strength and modified formant frequency, i.e., $p_n^1 = A_n \exp(j\theta_n^1)$. The technique is used to generate timbre through the addition of multiple cosinusoidal oscillations:

$$y(t) = \sum_{K=1}^{K} r_k(t) \cos(2\pi (k f_0)^{\alpha_{mas}} t + \varphi_k) \quad (2)$$

where K is the harmonic index, $r_k(t)$ is amplitude, φ_k is the phase, t is time. Equation 2 resembles the synthesis of a periodic signal with an inverse Fourier series which combines harmonic cosinusoidal oscillations, each with some magnitude and some phase shift. The McAdams coefficient is used to adjust the frequency

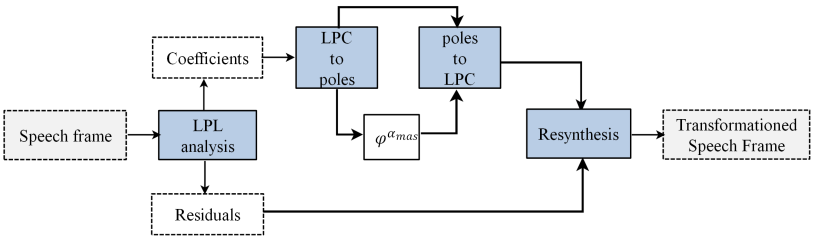

Fig. 3 Pipeline using the McAdams transform method: The pole coordinate coefficients with non-zero imaginary parts are subjected to the power operation of the coefficient α_{mas}, resulting in the distortion of their spectral envelope.

of each harmonic$(\theta_n^i,)$. Adjustments to the distribution of harmonics act to modify the resulting timbre. Fig 4 shows an example. The figure on the left shows the pole position changed by McAdams transformation, while the figure on the right shows the influence on the spectrum envelope

3.3 Modulation Spectrum Smoothing

Modulation spectrum smoothing removes the temporal fluctuation of speech features [18]. Let $X \in C^{(FT)}$ be the complex spectrogram obtained by STFT of the original speech, where F and T are the numbers of frequency bins and frames, respectively. A temporal sequence of the log amplitude spectrogram at frequency f, $[\log|X_{(f,1)}|, ..., \log|X_{(f,1)}|]$, is filtered by a zero-phase low pass filter, where $X_{(f,t)}$ is the $f, t-th$ component of X. A cutoff modulation frequency of the low pass filter is denoted as $\alpha_{ms} \in [0, 1]$, where 1 corresponds to the Nyquist modulation frequency. The transformed speech is synthesized by inverse STFT of the smoothed amplitude and original phase spectrograms. In Fig 5, the left side shows the smoothing effect of a certain frequency in the spectrum envelope, and the right side shows the complete smoothing effect.

4 Experiment

4.1 Dataset

Interactive Emotional Dyadic Motion Capture (IEMOCAP) et al. [19] is the most widely used dataset in the SER field. It contains 12 hours of emotional speech performed by 10 actors from the Drama Department of University of Southern California. The performance is divided into two parts, improvised and scripted, according to whether the actors perform according to a fixed script. The utterances are labeled with 9 types of emotion-anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state. For the databases, a single utterance may have multiple labels owing to different annotators. We consider only the label that has majority agreement.

Due to the imbalances in the dataset, researchers usually choose the most common emotions, such as neutral state, happiness, sadness, and anger. Because excitement and happiness have a certain degree of similarity and there are too few happy utterances.

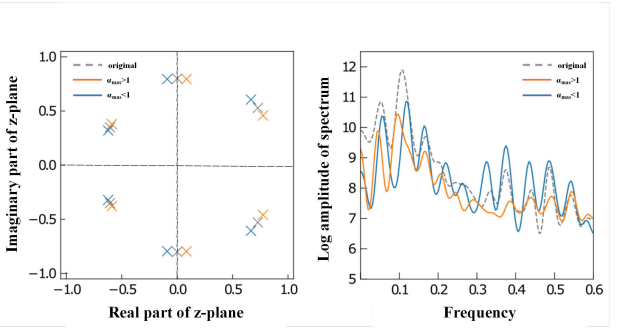

Fig. 4 Examples of McAdams transformation: Transformed pole shift (left) and spectrogram change after transformation (right).

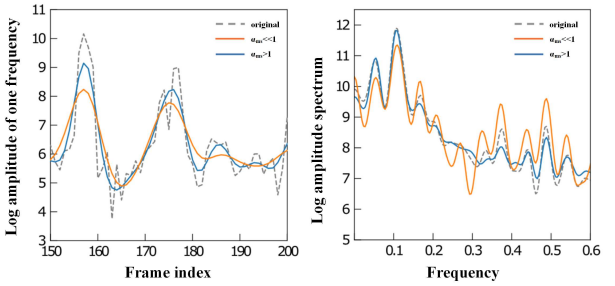

Fig. 5 Examples of Modulation Spectrum Smoothing: Temporal smoothing of amplitudes (left) and spectral changes before and after smoothing(right).

Researchers sometimes replace happiness with excitement or combine excitement and happiness to increase the amount of data [20, 21, 22]. In addition, previous studies have shown that the accuracy of using improvised data is higher than that of scripted data, [20, 23] which may be due to the fact that actors pay more attention to expressing their emotions rather than to the script during improvisation. In this paper, following other published works, we use data in the IEMOCAP dataset with four types of emotion neutral, excitement, sadness, and anger.

4.2 Evaluation Metrics

The paper uses weighted accuracy (WA) and unweighted accuracy (UA) for evaluation, where WA weighs each class according to the number of samples in that class and UA calculates accuracy in terms of the total correct predictions divided by total samples, which gives equal weight to each class:

$$UA = \frac{TP + TN}{P + N}, WA = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \quad (3)$$

where P is the number of correct positive instances (equivalent to $TP + FN$) and N is the number of correct negative instances(equivalent to $TN + FP$). Considering that WA and UA may not reach the maximum in the same model, we calculate the average of WA and UA as

the final evaluation criterion (indicated by ACC below), i.e., we save the model with the largest average of WA and UA as the optimal model [24]. Meanwhile, Automatic Speech Recognition(ASR) is used as the change standard before and after speech processing. And the word error rate (WER) will be calculated:

$$WER = \frac{N_{sub} + N_{del} + N_{ins}}{N_{ref}} \quad (4)$$

where N_{sub}, N_{del}, and N_{ins} are the number of substitution, deletion, and insertion errors, respectively, and N_{ref} the number of words in the reference [14]. We will calculate WER on the voice before and after the attack as the standard to judge the voice quality.

4.3 Evaluation Setup

We randomly divide the dataset into a training set (80% of data) and a test set (20% of data) for a 5-fold cross-validation. Each utterance is divided into 2-second segments, with 1 second (in training) or 1.6 seconds (in testing) overlap between segments. Although divided, the test is still based on the utterance, and the prediction results from the same utterance are averaged as the prediction result of the utterance.

Table 1 Results of three different online emotion recognition

Model	UA(%)	WA(%)	ACC(%)
CNN-LSTM	63.23	65.36	64.30
GCN	77.54	79.34	78.44
CNN-MAA	77.05	79.11	78.08

4.4 Evaluation Results

Table 1 shows the results of three different online emotion recognition for the dataset (IEMOCAP). Firstly, VTLN is used to attack three different models. Table 2 describes the performance of the three models under adversarial attack. With the continuous adjustment of super parameters α_{vtln}, the success rate of attack is also increasing. However, due to excessive and obvious transformation, the original voice content will change too much, which is a loss for the value of voice. Therefore, in our experiment, we know that it has the best performance when the hyperparameter $\alpha_{vtln} = 0.15$, and the WER is reduced from 11.23% to 21.40%, down by 10.17%. The recognition accuracy of the three models is reduced to about 10% in Fig 6, indicating that they have good resistance to the emotion recognition system.

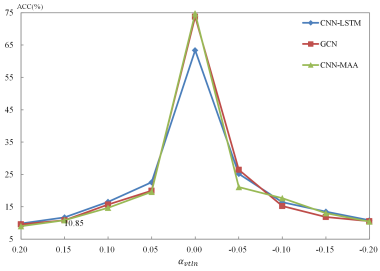

Fig. 6 Description of the three models under the Vocal Tract Length Normalization attack.

Table 3 shows the recognition results of the three models under McAdams transformation attack. Due to the particularity of McAdams coefficient, there are two relatively symmetric transformation modes in

forward and reverse, so the recognition results in the table also show a symmetry. According to the experimental results in the Fig 7, the best attack performance will be obtained when the $\alpha_{mas} = 1.20$ (reverse is 0.80), reducing the recognition accuracy of the three models to 8% - 10%. Meanwhile, WER decreased by only 6.24%.

Table 4 shows the results of the three models on the Modulation Spectrum Smoothing attack method. According to the analysis of the experimental results, as shown in Fig 8, when the $\alpha_{ms} = 0.25$, the best attack effect can be obtained, and the accuracy of emotion recognition can be reduced to 12% - 14%, and WER reduced by 8.83%. After the three attack methods, the recognition accuracy of the model dropped significantly. At the initial hyperparameter α_* (0.05, 0.95.0.05, respectively), the model accuracy dropped to 20%-25%, indicating that the three black-box confrontation attacks effectiveness, the robustness of the model is not excellent.

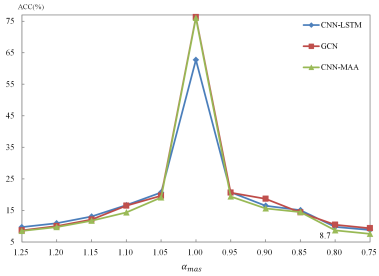

Fig. 7 Description of the three models under the McAdams transform attack.

Table 2 Performance of the three models under the Vocal Tract Length Normalization attack($UA/WA/ACC$)

α_{vtln}	CNN-LSTM(%)	GCN(%)	CNN-MAA(%)
0.20	9.25/10.36/9.80	9.54/9.61/9.58	8.85/9.12/8.99
0.15	11.56/11.84/**11.70**	10.22/11.50/**10.86**	10.75/10.94/**10.85**
0.10	15.62/17.45/16.54	14.12/17.21/15.67	14.55/14.85/14.70
0.05	20.55/24.63/22.59	19.31/20.56/19.94	19.42/19.78/19.60
0.00	62.54/64.27/63.41	75.23/72.32/73.78	76.24/73.32/74.78
-0.05	24.77/25.61/25.19	25.49/27.38/26.44	20.55/21.64/21.10
-0.10	16.51/16.35/16.43	14.85/15.64/15.25	17.43/17.87/17.65
-0.15	12.43/14.62/13.53	11.26/12.43/**11.85**	12.65/13.44/13.05
-0.20	10.45/11.24/10.85	10.46/10.65/10.56	10.32/10.55/10.44

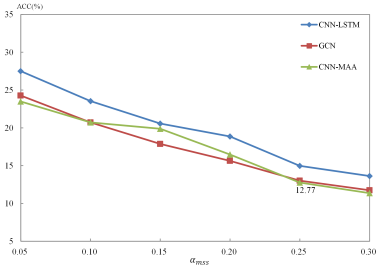

Fig. 8 Description of the three models under the Modulation Spectrum Smoothing attack.

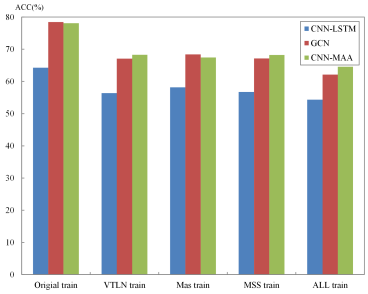

Fig. 9 Three models after adversarial training.

Table 3 Performance of the three models under the McAdams transform attack($UA/WA/ACC$)

α_{mas}	CNN-LSTM(%)	GCN(%)	CNN-MAA(%)
1.25	9.54/09.95/9.75	8.66/08.72/8.69	8.02/08.94/8.48
1.20	10.54/11.32/10.93	9.54/10.55/**10.05**	9.33/10.06/9.70
1.15	12.75/13.44/13.10	11.83/12.31/12.07	11.45/12.02/11.74
1.10	15.56/17.75/16.66	15.66/17.37/16.52	14.03/14.69/14.36
1.05	18.64/20.68/20.68	19.55/19.73/19.64	18.40/19.83/19.12
1.00	61.77/63.64/62.71	77.44/76.27/76.86	75.34/76.73/76.04
0.95	19.94/21.55/20.75	20.33/20.97/20.65	19.21/19.63/19.42
0.90	16.03/16.94/16.49	18.42/18.93/18.68	15.42/15.88/15.65
0.85	14.88/15.32/15.10	13.03/15.64/14.34	14.03/14.86/14.45
0.80	9.57/10.04/**9.81**	9.93/11.01/10.47	08.32/09.07/**8.70**
0.75	8.57/09.04/8.81	8.93/09.77/9.35	7.32/07.88/7.60

Table 4 Performance of the three models under the Modulation Spectrum Smoothing attack($UA/WA/ACC$)

α_{ms}	CNN-LSTM(%)	GCN(%)	CNN-MAA(%)
0.30	13.74/13.49/13.62	11.53/11.95/11.74	11.14/11.59/11.37
0.25	14.94/14.99/**14.97**	12.46/13.55/**13.01**	12.22/13.32/**12.77**
0.20	18.44/19.29/18.87	14.93/16.32/15.63	15.74/17.22/16.48
0.15	20.34/20.77/20.56	17.44/18.30/17.87	19.32/20.43/19.88
0.10	23.21/23.89/23.55	19.92/21.49/20.71	20.11/21.34/20.73
0.05	26.56/28.44/27.50	23.93/24.60/24.27	22.94/24.05/23.50

Table 5 Changes of speech quality before and after the change

Method	α_*	$WER(\%)$
Original	-	11.23
VTLN	0.15	21.40
	-0.15	23.84
McAdams	1.20	18.69
	0.80	17.47
MSS	0.25	20.06

After we add three kinds of adversarial samples into the training, as shown in Table 6, three different adversarial samples are added. As shown in Fig 9, VTLN train, Mas train and MSS train respectively add one adversarial sample to the training with the correct label, and then test the accuracy of the model. The best performance is the adversarial samples produced by adding McAdams. The recognition result of GCN model can reach 68.40% after adversarial training. After adding three kinds of samples together into the adversarial training (All train in Fig 9), the best performance model is CNN-MAA, and the recognition accuracy is 64.60%. According to our analysis, the above two models still have strong robustness after adversarial training because they have better learning effect on sample dispersion by incorporating graph structure and area attention mechanism.

5 Conclusion

By transferring the method of voice privacy protection to the field of SER, a black-box attack is carried out under the condition of an unknown emotional recognition system, and it is found that warp transformation processing has a strong resistance to emotional recognition. After simple warp transformation, the voice is well protected in the trained SER and the usability of voice content is guaranteed. In different speech transformation processing, the final attack effect is not the same. Experiments show that, among which the McAdams attack method has the best attack effect $WA = 08.32\%$. Different emotional recognition models have high mobility and low time cost.

This kind of black box attack is a kind of no-target attack. There is no actual direction and prediction for the result of the attack. Meanwhile, after the adversarial samples are added to the training, although the accuracy of the model decreases to a certain extent, the recognition results still have a certain accuracy, and the model has a certain robustness to such adversarial samples.

Our work in the future should be to study how to make a clear and targeted attack through the voice warp transformation.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 6217011361, U1736215, 61901237), Zhejiang Natural Science Foundation (Grant No. LY20F020010), Ningbo Natural Science Foundation (Grant

Table 6 Performance of the three models under the Modulation Spectrum Smoothing attack($UA/WA/ACC$)

Model	VTLN	McAdams	MSS	ALL
CNN-LSTM	55.57/57.21/56.39	57.84/58.49/58.17	56.04/57.39/56.72	52.74/55.93/54.34
GCN	66.73/67.38/67.06	67.32/69.47/68.40	66.94/67.30/67.12	61.93/62.29/62.11
CNN-MAA	67.32/69.19/68.26	66.04/68.84/67.44	66.73/69.72/68.23	63.83/65.37/64.60

No. 202003N4089) and K.C. Wong Magna Fund in Ningbo University.

References

[1] Tang, D., Zeng, J., Li, M.: An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In: Interspeech, pp. 162–166 (2018)

[2] Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., Li, C.: Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition (2018)

[3] Huang, C.-W., Narayanan, S.S.: Attention assisted discovery of sub-utterance structure in speech emotion recognition. In: Interspeech, pp. 1387–1391 (2016)

[4] Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231 (2017). IEEE

[5] Aldeneh, Z., Khorram, S., Dimitriadis, D., Provost, E.M.: Pooling acoustic and lexical features for the prediction of valence. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 68–72 (2017)

[6] Jin, Q., Li, C., Chen, S., Wu, H.: Speech emotion recognition with acoustic and lexical features. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4749–4753 (2015). IEEE

[7] Mao, S., Ching, P., Lee, T.: Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition. In: Interspeech, pp. 1686–1690 (2019)

[8] Han, W., Ruan, H., Chen, X., Wang, Z., Li, H., Schuller, B.W.: Towards temporal modelling of categorical speech emotion recognition. In: Interspeech, pp. 932–936 (2018)

[9] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J.: Direct modelling of speech emotion from raw speech. arXiv preprint arXiv:1904.03833 (2019)

[10] Shirian, A., Guha, T.: Compact graph architecture for speech emotion recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6284–6288 (2021). IEEE

[11] Xu, M., Zhang, F., Cui, X.,

Zhang, W.: Speech emotion recognition with multiscale area attention and data augmentation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6319–6323 (2021). IEEE

[12] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

[13] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)

[14] Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B.M.L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., *et al.*: The voiceprivacy 2020 challenge: Results and findings. Computer Speech & Language **74**, 101362 (2022)

[15] Lee, L., Rose, R.: A frequency warping approach to speaker normalization. IEEE Transactions on speech and audio processing **6**(1), 49–60 (1998)

[16] McAdams, S.E.: Spectral Fusion, Spectral Parsing and the Formation of Auditory Images. Stanford university, ??? (1984)

[17] Itakura, F.: Analysis synthesis telephony based on the maximum likelihood method. In: The 6th International Congress on Acoustics, 1968, pp. 280–292 (1968)

[18] Takamichi, S., Kobayashi, K., Tanaka, K., Toda, T., Nakamura, S.: The naist text-to-speech system for the blizzard challenge 2015. In: Proc. Blizzard Challenge Workshop, vol. 2 (2015). Berlin, Germany

[19] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**(4), 335–359 (2008)

[20] Li, P., Song, Y., McLoughlin, I.V., Guo, W., Dai, L.-R.: An attention pooling based representation learning method for speech emotion recognition (2018)

[21] Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., Schuller, B.: Attention-enhanced connectionist temporal classification for discrete speech emotion recognition (2019)

[22] Neumann, M., Vu, N.T.: Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7390–7394 (2019). IEEE

[23] Tarantino, L., Garner, P.N., Lazaridis, A., *et al.*: Self-attention for speech emotion recognition. In: Interspeech, pp. 2578–2582 (2019)

[24] Jalal, M.A., Milner, R., Hain, T.: Empirical interpretation of speech emotion perception with attention based model for speech

emotion recognition. In: INTERSPEECH, pp. 4113–4117 (2020)