EURASIP Journal on Audio, Speech, and Music Processing

REVIEW Open Access

Black-box adversarial attacks through speech distortion for speech emotion recognition

Jinxing Gao, Diqun Yan* and Mingyu Dong

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, and 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 word error rate (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. Our code is available in Github.

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 [1] 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 [2–5]. While the use of hand-crafted acoustic features is still common in SER, lexical features [6, 7] and log Mel spectrgrams are also used as input [8]. Spectrograms are often used with Convolutional Neural Networks (CNNs) 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 [9]. 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) [10–12] and various deep learning techniques, etc. [13–16]. 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 [17–19] has shown that neural net-

*Correspondence: yandiqun@nbu.edu.cn
College of Information Science and Engineering, Ningbo University, Zhejiang, China

works 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 Fast Gradient Sign Method (FGSM)[20] and Project Gradient Descent (PGD)[21], 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) [22] 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. The recognition accuracy largely depends on the acoustic features of the input and the network conditions used. Acoustic features mainly rely on contextual information in the input speech for computation. The combination of Convolution Neural Networks (CNNs) and Long-Short Term Memory (LSTM) has gained a huge advantage in learning contextual information that is crucial for emotion recognition. CNN can overcome the scalability problem of traditional neural networks, while LSTM has long-term memory and solves the problems of vanishing and exploding gradients during training of long sequences.

In [10], 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

In 2021, by Amir Shirian et al., a light-weight depth map method is proposed to solve the task of speech emotion recognition [11]. Following the theory of graph signal processing, modeling speech signals as cyclic graphs or line graphs is a more compact, efficient and scalable form compared to traditional CNN networks. At the same time, compared with the traditional graph structure, the author greatly simplifies the convolution operation on the graph by reducing the operation of weighted edges on the traditional graph, so the parameters that can be learned in the SER task are significantly reduced, and its performance is better than that of LSTM, standard GCN, and other state-of-the-art graph models in SER.

2.3 SER based on CNN-MAA

In SER, emotional features are often represented by multiple energy patterns in the spectrogram. Conventional attention neural network classifiers for SER are usually optimized at a fixed attention granularity. While Xu Mingke et al. [12] applied multiscale area attention in deep convolutional neural networks to focus on emotional features with different granularities, so the classifier could benefit from attention sets with different scales. Meanwhile, channel length perturbation is used for data augmentation 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 distoritation

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

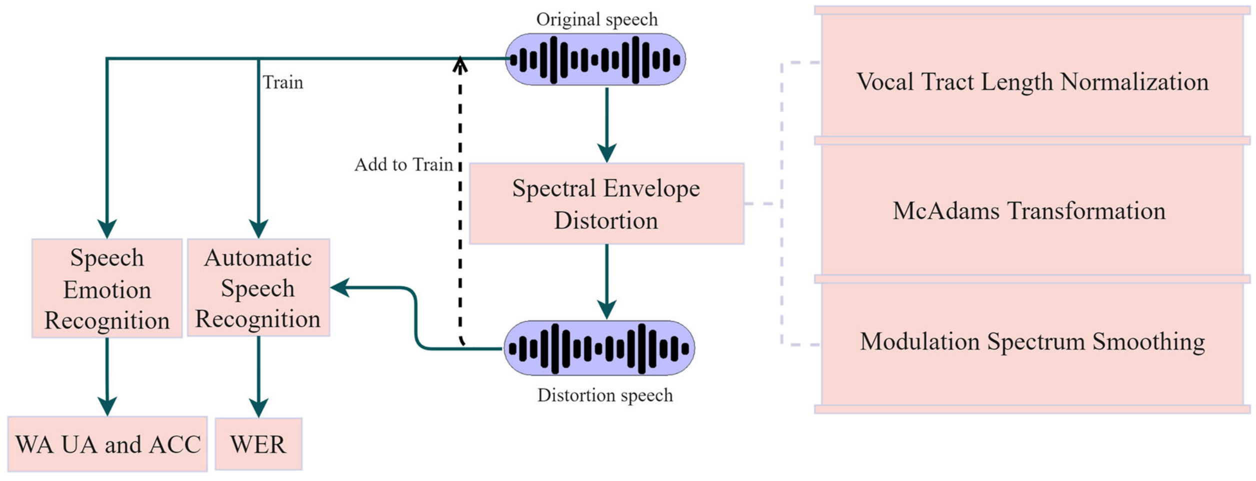
Original speech
Train
Add to Train
Spectral Envelope Distortion
Speech Emotion Recognition
Automatic Speech Recognition
Distortion speech
WA UA and ACC
WER
Vocal Tract Length Normalization
McAdams Transformation
Modulation Spectrum Smoothing

Fig. 1 Adversarial attacks against speech emotion models flow chart

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 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. Figure 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) [23] was originally used for speech-to-text recognition tasks to remove distortions caused by differences in channel lengths by modifying the magnitude spectrum of the original speech through a warping function. Let $\omega_0 \in [0,1]$ and $\omega_1 \in [0,1]$

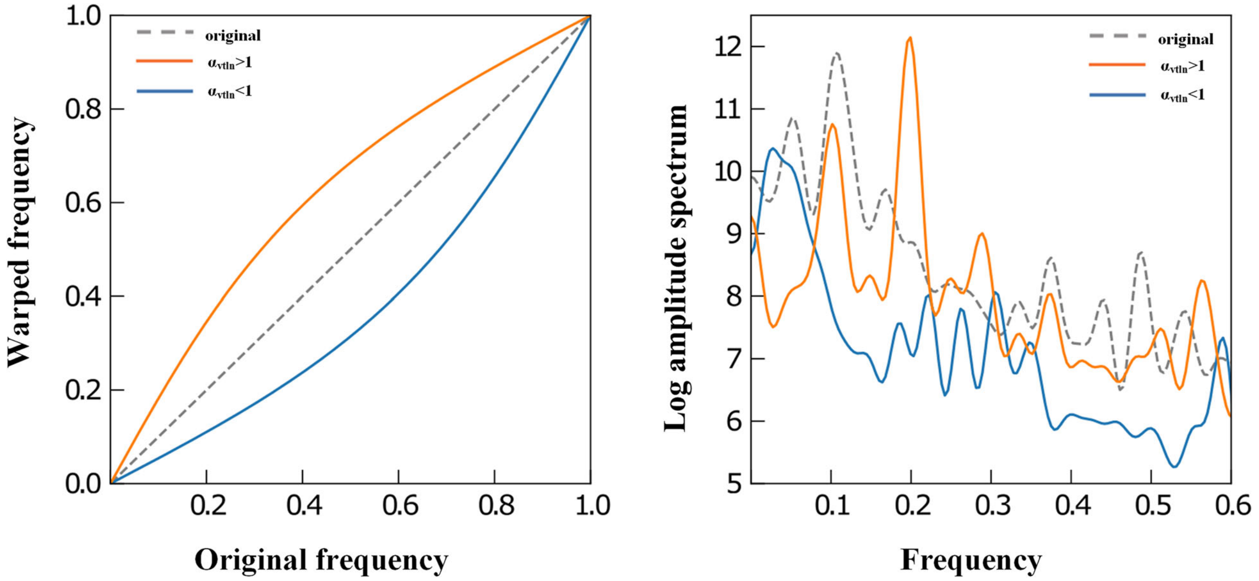

Fig. 2 Examples of VTLN: description of the warping function (left) and the change in the spectral curve after application (right)

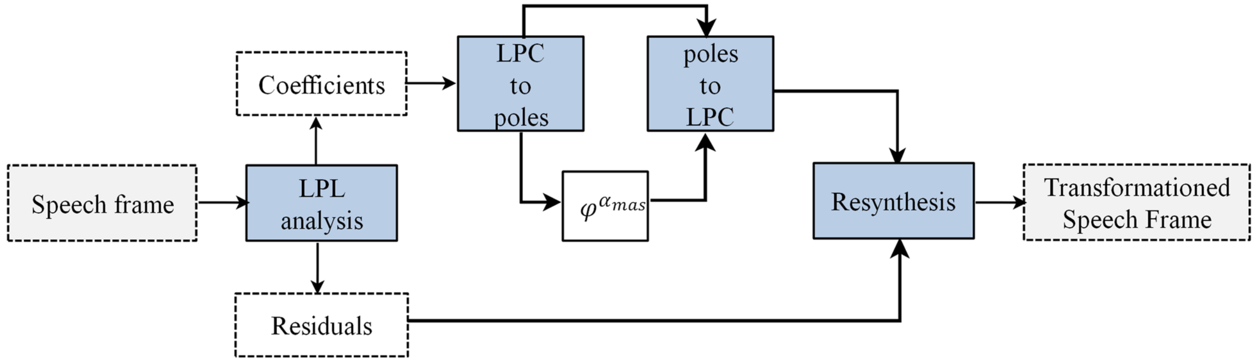

Fig. 3 Pipeline using the McAdams transform method: the pole coordinate coefficients with non-zero imaginary parts are subjected to the power operation of the coefficient α_{mas}, resulting in the distortion of their spectral envelope

the frequency of the original speech and corresponding warped frequency, respectively. $\omega_0 = 1$ is the Nyquist frequency. ω_0 is warped into ω_1 as

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

where $\alpha_{vtln} \in [-1, 1]$ is a hyperparameter of the warping function and also represents the degree of frequency warping. Figure 2 shows the warping results for different hyperparameter choices. When $\alpha_{vtln} < 0$ and $\alpha_{vtln} > 0$, the distorted spectral curves become convex and concave, respectively, which represent the contraction and expansion amplitudes, respectively. When $\alpha_{vtln} = 0$, $\omega_0 = \omega_1$, which means no warping. In this paper, we first obtain the log amplitude spectrum from the original speech using the short-time Fourier transform, and then perform frequency warping with VTLN to obtain the warped log-amplitude spectrum. And finally, the transformed speech is obtained by inverse STFT of the modified amplitude spectrogram and original phase spectrogram.

3.2 McAdams transformation

McAdams transformation [24] achieves the result of speech transformation by modifying the formant frequency of speech. By performing Linear Predictive Coding (LPC) [25] on the original speech, we can get the N poles. Pole $p_n \in \boldsymbol{c}$ is written as $A_n \exp(\theta_n)$ in the polar coordinate, where $A_n \in [0, 1]$ is calculated from the LPC,

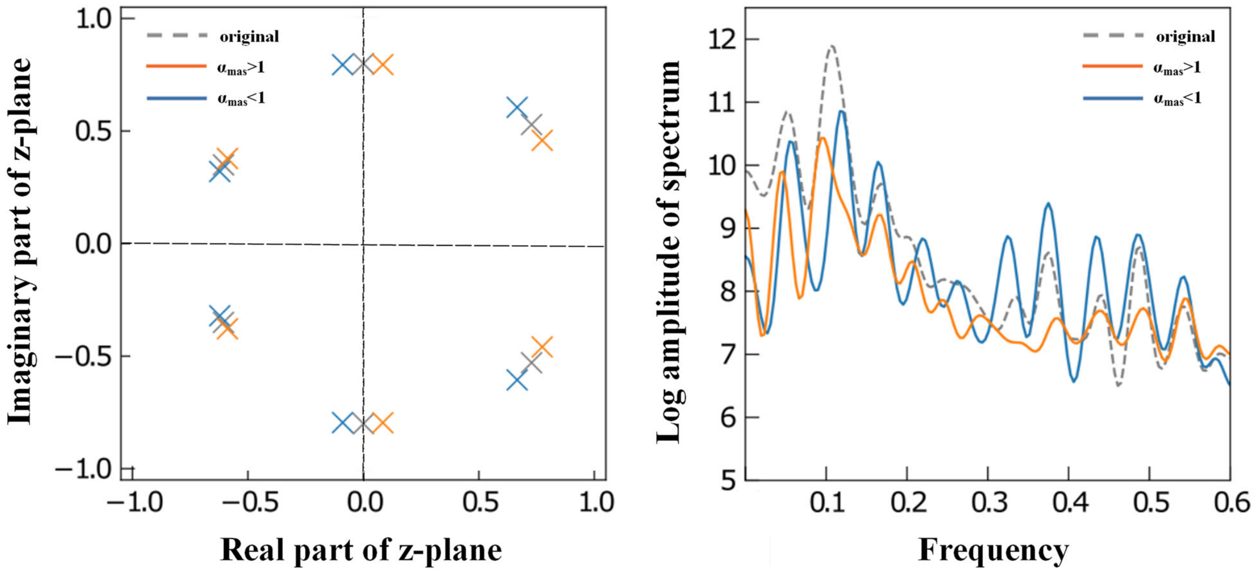

Fig. 4 Examples of McAdams transformation: transformed pole shift (left) and spectrogram change after transformation (right)

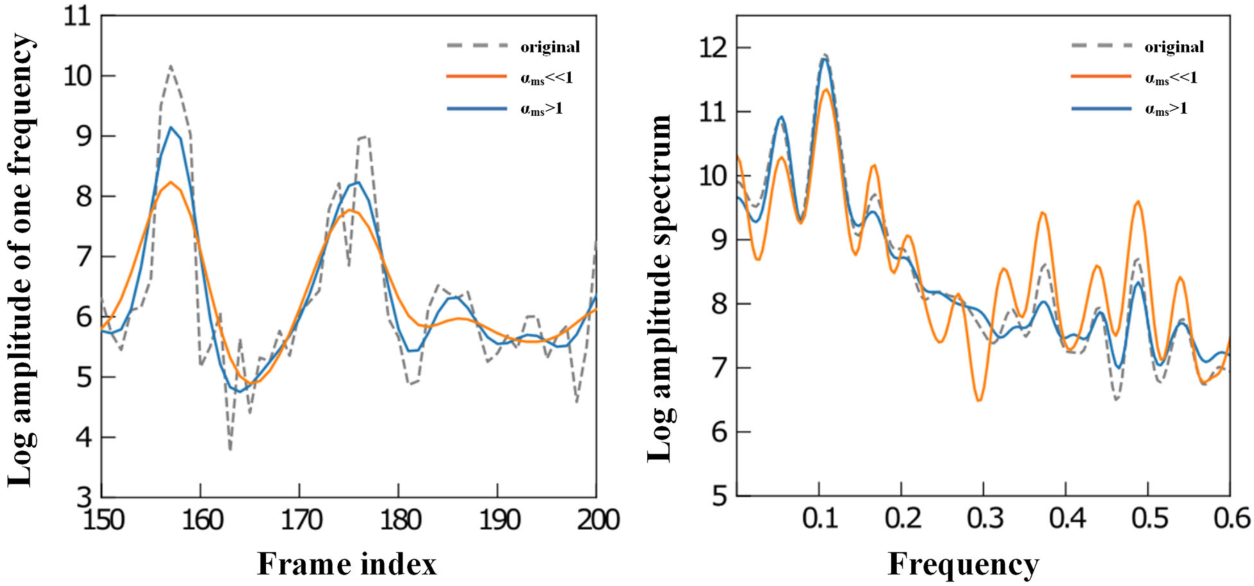

Fig. 5 Examples of modulation spectrum smoothing: temporal smoothing of amplitudes (left) and spectral changes before and after smoothing (right)

which is less than 1, and $\theta_n \in [0, \pi]$ is the offset phase. The pipeline of the McAdams transformation approach is shown in Fig. 3.

The transformed frequency $\theta_n^1 \in [0, \pi]$ is obtained by performing $\theta_n^1 = \theta_n^{\alpha_{mas}}$ on the original frequency $\theta_n \in [0, \pi]$, where $\alpha_{mas} \in R_+$ is the McAdams coefficient. We obtain the corrected pole p_n^1 by combining the original formant intensity with the transformed formant frequency, i.e., $p_n^1 = A_n \exp(j\theta_n^1)$. The general speech transformation is to generate the transformed waveform by adding multiple cosine oscillations to the original oscillation wave:

$$y(t) = \sum_{K=1}^{K} r_k(t) \cos(2\pi (kf_0)^{\alpha_{mas}} t + \varphi_k) \qquad (2)$$

where K is the harmonic index, $r_k(t)$ is amplitude, φ_k is the phase, t is time. Equation 2 represents the synthesis of periodic signals that combine harmonic cosine oscillations, each with a certain amplitude and phase offset. The purpose of the McAdams coefficient is to adjust the frequency of each harmonic, namely θ_n^i, to produce transformed speech by modifying the harmonics in the original speech. Figure 4 shows an example. The picture on the left is the case where the pole position is transformed by McAdams transformation, and the picture on the right is the influence on the spectral envelope.

Table 1 Recognition results of the above three speech emotion recognition models

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

3.3 Modulation spectrum smoothing

Modulation spectrum smoothing achieves the purpose of modifying speech by removing the temporal fluctuation of speech features [26]. The original speech is obtained by short-time Fourier transform to obtain the complex spectrogram $X \in C^{(FT)}$ where F and T are the numbers

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

Bold fonts indicate the best attack performance under the current modes

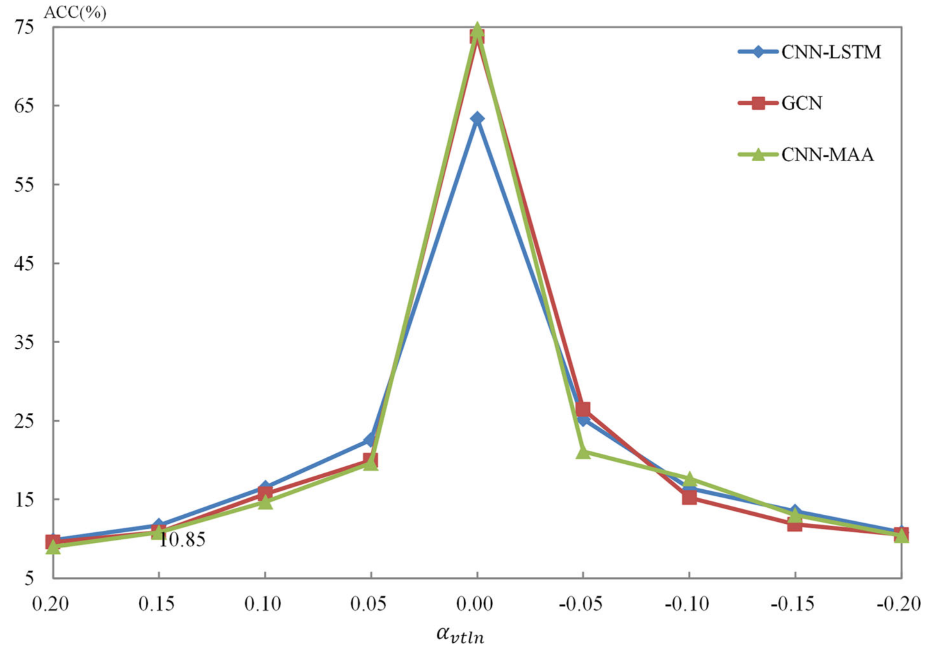

Fig. 6 Description of the three models under the Vocal Tract Length Normalization attack

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*. The cutoff frequency range of the low-pass filter we used is $\alpha_{ms} \in [0, 1]$, after filtering, the inverse short-time Fourier transform is used to combine the smoothed amplitude with the original phase spectrogram to generate the transformed speech. 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

The most widely used data set in the above SER model is Interactive Emotional Dyadic Motion Capture (IEMOCAP)et al. [27]. Therefore, in this paper, in order to explore the effect of black-box attack on the above model, we also use the data in this data set for research. It contains 12 h 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. For the labeled data in the database, we only consider the case of many labels due to the difference of vision aids. In previous studies [28–30], due to the imbalanced data in the dataset (fewer happy data), researchers usually choose more common emotions such as neutral state, sadness, anger, and because of excitement and happiness there is a certain similarity, so the excitement will be replaced by happiness, or the excitement and happiness will be combined to

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	7.32/7.88/7.60

Bold fonts indicate the best attack performance under the current modes

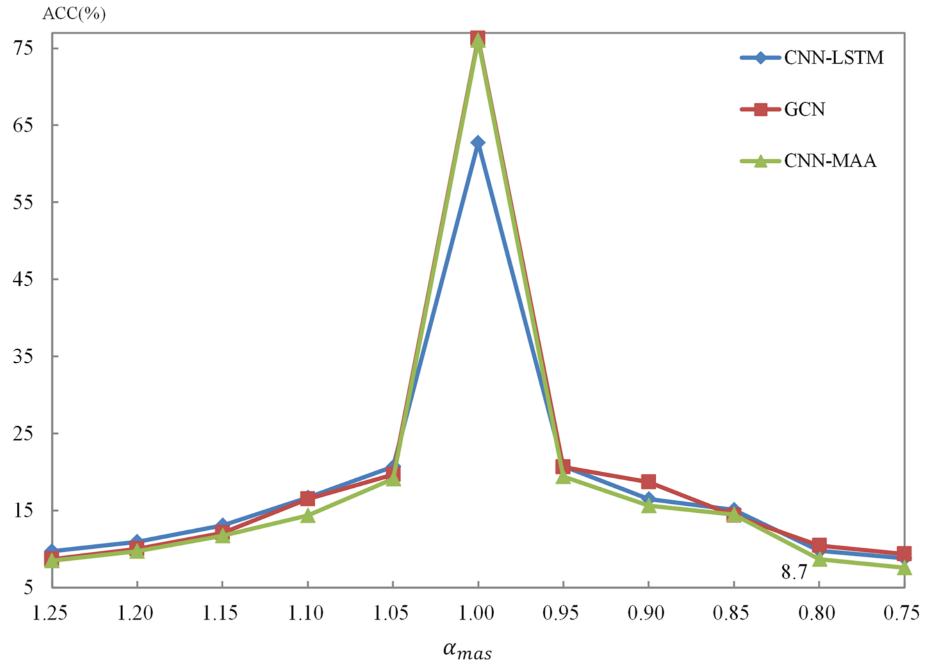

Fig. 7 Description of the three models under the McAdams transform attack

increase the amount of data. In this paper, we also use the four emotions of neutral, excitement, sadness and anger from the IEMOCAP dataset.

4.2 Evaluation metrics

Evaluating the recognition performance in the above SER model uses weighted accuracy (WA) and unweighted accuracy (UA), 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:

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

where P is the number of correct positive instances, N is the number of all negative samples, and True Positive(TP) and True Negative(TN) are the number of positive and negative samples predicted correctly, respectively. And in [31], considering that WA and UA may not reach the maximum value in the same model, their average ACC is used as the final evaluation standard (the smaller the ACC, the better the attack effect on the model is). At the same time, in order to show the actual auditory effect of the transformed speech, we use automatic speech recognition (ASR) as the change standard before and after speech processing. And will calculate the word error rate:

$$WER = \frac{N_{sub} + N_{del} + N_{ins}}{N_{ref}} \tag{4}$$

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 [22]. We will calculate WER on the voice before and after the attack as the standard to judge the voice quality.

4.3 Evaluation setup

In the experiments, we randomly split the dataset into training set (80%) and test set (20%) for cross-validation. First of all, after the above three SER models are trained on the training set, they are tested with the test set, and then the test set is processed with three different black-box attack methods, and then the attack effect is identified and explored. Finally, adversarial training is added to explore the robustness of the model.

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

Bold fonts indicate the best attack performance under the current modes

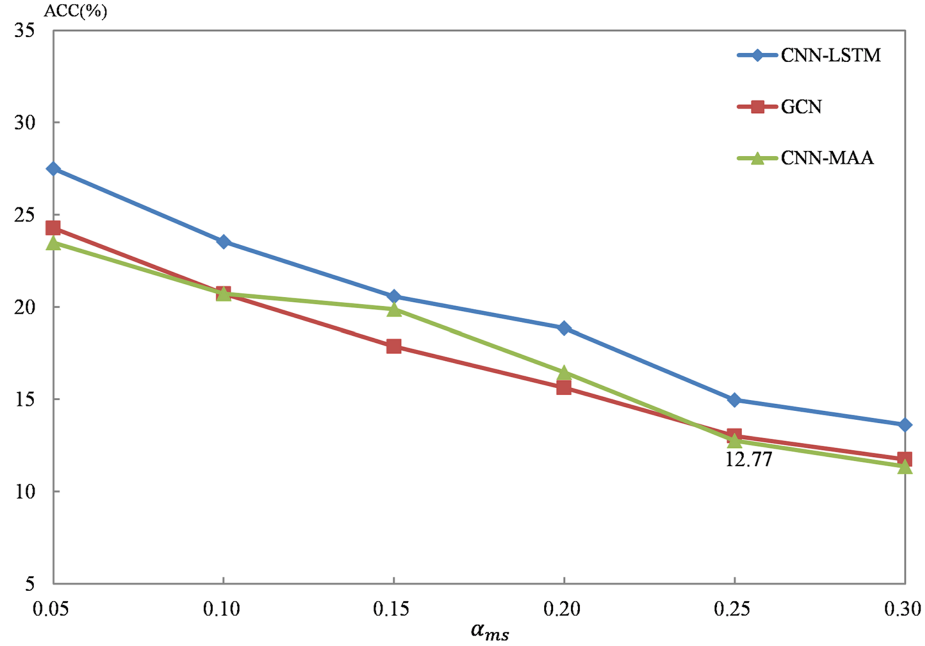

Fig. 8 Description of the three models under the modulation spectrum smoothing attack

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 speech, so the loss of speech quality also needs to be taken into account as a consideration when considering the best attack case i.e. the growth of *WER*. Therefore, in our experiment, we know that it has the best performance when the hyperparameter $\alpha_{vtln} = 0.15$, and the *WER* increases from 11.23 to 21.40% as shown in Table 5, which means that the speech quality effect decreases 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.

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* increased by only 6.24% in the Table 5.

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* increased by 8.83% in the Table 5. 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.

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

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

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

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 = 8.32%. 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.

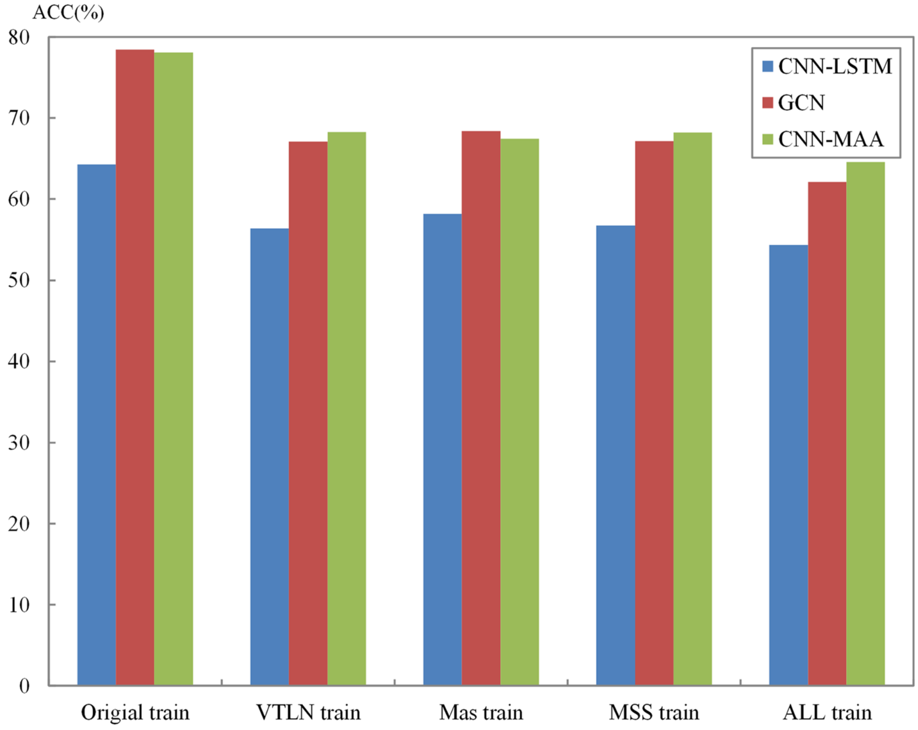

Fig. 9 Three models after adversarial training

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61300055), Zhejiang Natural Science Foundation (Grant No. LY20F020010), Ningbo Natural Science Foundation (Grant No. 202003N4089) and K.C. Wong Magna Fund in Ningbo University.

Authors' contributions

JX Gao conceived the study, conducted the research and design of the attack method, and participated in the discussion and analysis of the results. DQ Yan participated in the design of the study and the analysis of the results, and participated in the design of the structure of the paper. MY Dong participated in the research and design of attack methods, and wrote the thesis. All authors have read and approved the final manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 14 May 2022 Accepted: 25 July 2022
Published online: 17 August 2022

References

1. M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Comm. **116**, 56–76 (2020)
2. D. Tang, J. Zeng, M. Li, in *Interspeech 2018*. An end-to-end deep learning framework for speech emotion recognition of atypical individuals, (Hyderabad, 2018), pp. 162–166
3. Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, C. Li, Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition (2018)
4. C.-W. Huang, S. S. Narayanan, in *Interspeech 2016*. Attention assisted discovery of sub-utterance structure in speech emotion recognition, (San Francisco, 2016), pp. 1387–1391
5. S. Mirsamadi, E. Barsoum, C. Zhang, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Automatic speech emotion recognition using recurrent neural networks with local attention (IEEE, New Orleans, 2017), pp. 2227–2231
6. Z. Aldeneh, S. Khorram, D. Dimitriadis, E. M. Provost, in *Proceedings of the 19th ACM International Conference on Multimodal Interaction 2017*. Pooling acoustic and lexical features for the prediction of valence, (Glasgow, 2017), pp. 68–72
7. Q. Jin, C. Li, S. Chen, H. Wu, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech emotion recognition with acoustic and lexical features (IEEE, Brisbane, 2015), pp. 4749–4753
8. S. Mao, P. Ching, T. Lee, in *Interspeech 2019*. Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition, (Graz, 2019), pp. 1686–1690
9. W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, B. W. Schuller, in *Interspeech 2018*. Towards temporal modelling of categorical speech emotion recognition, (Hyderabad, 2018), pp. 932–936
10. S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, Direct modelling of speech emotion from raw speech. arXiv preprint arXiv:1904.03833 (2019)
11. A. Shirian, T. Guha, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Compact graph architecture for speech emotion recognition (IEEE, Toronto, 2021), pp. 6284–6288
12. M. Xu, F. Zhang, X. Cui, W. Zhang, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech emotion recognition with multiscale area attention and data augmentation (IEEE, Toronto, 2021), pp. 6319–6323
13. F. Albu, D. Hagiescu, L. Vladutu, M.-A. Puica, in *EDULEARN 2015: 7th International Conference on Education and New Learning Technologies*. Neural network approaches for children's emotion recognition in intelligent learning applications, (Spain, 2015)
14. M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recog. **44**(3), 572–587 (2011)
15. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, T. Alhussain, Speech emotion recognition using deep learning techniques: A review. IEEE Access. **7**, 117327–117345 (2019)
16. B. J. Abbaschian, D. Sierra-Sosa, A. Elmaghraby, Deep learning techniques for speech emotion recognition, from databases to models. Sensors. **21**(4), 1249 (2021)
17. J. Chen, Y. Wu, X. Xu, Y. Chen, H. Zheng, Q. Xuan, Fast gradient attack on network embedding. arXiv preprint arXiv:1809.02797 (2018)
18. L. Yang, Q. Song, Y. Wu, Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. Multimed. Tools Appl. **80**(1), 855–875 (2021)
19. Q. Wang, B. Zheng, Q. Li, C. Shen, Z. Ba, Towards query-efficient adversarial attacks against automatic speech recognition systems. IEEE Trans. Inf. Forensics Secur. **16**, 896–908 (2020)
20. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
21. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
22. N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, *et al*, The voiceprivacy 2020 challenge: Results and findings. Comput. Speech Lang. **74**, 101362 (2022)
23. L. Lee, R. Rose, A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Process. **6**(1), 49–60 (1998)
24. S. E. McAdams, *Spectral Fusion, Spectral Parsing and the Formation of Auditory Images*, (1984), pp. 1–354
25. F. Itakura, in *Reports of the 6th International Congress on Acoustics, 1968*. Analysis synthesis telephony based on the maximum likelihood method, (Tokyo, 1968), pp. 280–292
26. S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, S. Nakamura, in *Proc. Blizzard Challenge Workshop, vol. 2*. The naist text-to-speech system for the blizzard challenge 2015 (Language resources and evaluation, Berlin, 2015)
27. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
28. P. Li, Y. Song, I. V. McLoughlin, W. Guo, L.-R. Dai, An attention pooling based representation learning method for speech emotion recognition (2018)
29. Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, B. Schuller, Attention-enhanced connectionist temporal classification for discrete speech emotion recognition (2019)
30. M. Neumann, N. T. Vu, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech (IEEE, Brighton, 2019), pp. 7390–7394
31. M. A. Jalal, R. Milner, T. Hain, in *Interspeech 2020*. Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition, (Shanghai, 2020), pp. 4113–4117

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.