Liu *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:22
https://doi.org/10.1186/s13636-023-00289-4

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Speech emotion recognition based on emotion perception

Gang Liu[1*], Shifang Cai[1] and Ce Wang[1]

Abstract

Speech emotion recognition (SER) is a hot topic in speech signal processing. With the advanced development of the cheap computing power and proliferation of research in data-driven methods, deep learning approaches are prominent solutions to SER nowadays. SER is a challenging task due to the scarcity of datasets and the lack of emotion perception. Most existing networks of SER are based on computer vision and natural language processing, so the applicability for extracting emotion is not strong. Drawing on the research results of brain science on emotion computing and inspired by the emotional perceptive process of the human brain, we propose an approach based on emotional perception, which designs a human-like implicit emotional attribute classification and introduces implicit emotional information through multi-task learning. Preliminary experiments show that the unweighted accuracy (UA) of the proposed method has increased by 2.44%, and weighted accuracy (WA) 3.18% (both absolute values) on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which verifies the effectiveness of our method.

Keywords Speech emotion recognition, Emotion perception, Implicit emotional attribute, Multi-task learning

1 Introduction

Speech emotion recognition (SER) usually refers to the process by which a machine automatically recognizes human emotions and emotion-related states from speech. Emotion plays an important role in human intelligence, rational decision-making, social interaction, perception, memory, learning, and creation. As a higher creature, an important factor that distinguishes human beings from animals is the transmission of emotions. Speech emotion recognition has a wide range of practical application scenarios, such as depression diagnosis [1], call center [2], online classroom [3], etc.

With deep learning flourishing, the neural network approach has swept many fields. Driven by deep learning, great progress has been made in the field of speech emotion recognition [4–9] and the performance has been greatly improved. Deep learning has become the mainstream method of speech emotion recognition. As early as 2002, Chang-hyun Park et al. [10] have used recurrent neural networks (RNN) in speech emotion recognition. Gradually, purely deep learning and end-to-end approaches emerged. In 2014, Jianwei Niu et al. [11] pioneered the use of deep neural networks (DNN) in speech emotion recognition. Qirong Mao et al. [12] used convolutional neural networks (CNN) to learn the invariance of speech emotion features. In 2015, Lee and Tashev [13] used RNN with long short-term memory (LSTM) units to tackle the tricky problem. In 2019, MA Jalal et al. [14] used the capsule neural network for temporal modeling of speech emotion. R Shankar et al. [15] applied highway neural network to speech emotion research, and in 2020, Shamane Siriwardhana et al. [16] exploited transformer to perform multi-modal emotion recognition including speech.

Deep learning improves the recognition performance of speech emotion recognition to some extent, but the network structure is mostly borrowed from the field of computer vision and natural language processing. The main network structure is also specially designed to

*Correspondence:
Gang Liu
liugang@bupt.edu.cn
[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

Springer Open

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

solve problems in other fields. How to reasonably use the network in other fields to improve the ability to model emotional information is a major problem in speech emotion recognition. Moreover, the scarcity of datasets and emotion perception makes the recognition task more challenging. Therefore, the performance of speech emotion recognition is still not ideal. In recent years, brain science is strengthening the exploration of the structure and function of various brain areas that produce emotions, thought and consciousness in the human brain. For example, emotion perception mainly depends on the limbic system of the human brain [17, 18], and different parts of the limbic system have different perception of different emotions [19, 20]. In this paper, an approach inspired by emotion perception is proposed based on the human brain's perceptive process of emotion, and a human brain-like implicit emotion attribute classification is designed. The implicit emotion attribute information is introduced through multi-task learning to increase the extraction of emotion information. Preliminary experiments show that the unweighted accuracy (UA) on the the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is improved by 2.44%, and the weighted accuracy (WA) by 3.18% (both absolute values), which verifies that the proposed human brain-like implicit emotion attribute classification is beneficial to extract emotion information.

The paper is organized as follows: Section 2 introduces the characteristics of the human brain's emotion perception. Section 3 introduces the network designed according to the characteristics of emotion perception. Section 4 elaborates the experimental results and conclusions. Section 5 summarizes the paper and further looks forward to the future development of speech emotion recognition.

2 Emotion perception

Most researchers are likely to improve the recognition performance by changing neural networks structure without referring to the perceptual mechanism of the human brain for emotion perception (perceptual network, perceptual process, perceptual characteristics, etc.). Due to the complexity of the human brain structure and the insufficiency of existing research techniques, it is difficult for brain science to see the full picture of the emotional cognitive mechanism of the human brain. How the human brain processes information in speech to recognize emotion is still a mystery. However, various imaging technologies and electrophysiological signals are used to establish topological structures of high-order connections of brain networks at various levels, and there have been many experiments carried out in related fields such as brain network modeling and emotional computing [21–25]. Existing research results also reveal some potential mechanisms of human emotional cognition. For example, different parts of the brain perceive different emotions differently [21, 25].

Research has shown that emotional perception is linked to a set of structures in the brain called the limbic system, which includes the hypothalamus, the cingulate cortex, the hippocampus, and others. Different parts play different roles in the perception of different emotions. For example, removing the amygdala leads to a reduction in fear, and the posterior hypothalamus may be particularly important for anger and aggression. The frontal cortex of the brain is more sensitive to intense emotions such as happiness and anger. A part of the brain called the hypothalamus is more active during the process of feeling sadness. Meanwhile, a part of the brain called the hippocampus plays a significant role in the perception of sadness.

Table 1 lists emotions and the parts of the brain associated with them. According to Table 1, the following characteristics related to the human brain perception can be concluded.

1) The human brain's perception of an emotion is related to multiple parts of the limbic system, which indicates that the human brain's emotional perception network may have a certain structure, and the

Table 1 Emotion and its related brain parts

Emotion	Brain parts								
	Amygdala	Cortex	Hippocampus	Cortex	Thalamus	Island	Gyrus	Occipital lobe	Neurite
Happy	Left	Right frontal				Left			Anterior
Fear	Bilateral	Left frontal			Hypothalamus				
Sad	Whole		Whole		Left	Left		Right	
Disgust	Left	Left subfrontal		Island					
Anger	Whole	Prefrontal	Right	Island					
Joy			Bilateral				Bilateral lower frontal		

structure causes the relevant parts of the human brain to be more sensitive to certain emotions. Could this structure be introduced into speech emotion recognition? Many parts of the human brain are sensitive to the same emotion, but there are differences in sensitivity. So, are there similar but not identical internal structures in the various parts of the human brain?

2) Different emotions are involved in different parts. For example, both anger and sadness are linked to the amygdala, but sadness is linked to the left thalamus, whereas anger is not. This means that the human brain has differences in the perception of different emotions and also shows that the differences in the perception of different parts of the human brain are related to the internal structure of the parts. So what exactly are these parts perceptual about?
3) One part of the limbic system is related to the perception of multiple emotions. For example, the amygdala is associated with the perception of happiness, sadness, anger, and other emotions. What common information does the amygdala perceive in these emotions?

According to above analysis, we propose a conjecture that some parts of the human brain's limbic system can perceive certain attribute information in emotions, and the attribute information is the common information of many emotions that this part can perceive. The specific attribute is unknown, so this paper calls it implicit attribute information. The perceptual network of the human brain for emotion has a certain structure. Therefore, implicit attribute information is extracted through some parts of the limbic system and then sent to the brain center with the underlying information for emotion recognition.

Based on these assumptions, this paper adopts artificial neural networks to simulate the parts of the human limbic system that draw on its mechanism of extracting and perceiving emotional information and proposes a method based on emotional perception. According to the limbic system's perception of different emotions, implicit attribute classification is defined, and the information of implicit attribute is extracted through multi-task learning which is then added into the emotion recognition system as auxiliary information for recognition.

3 Emotion recognition based on emotion perception

3.1 Implicit emotion attribute classification design

A part of the human limbic system can sense some implicit attribute information of emotions. If a part can sense some emotions, it means that these emotions contain the same implicit attribute information. At the same time, the fact that the same emotion can be sensed by different parts suggests that these parts have some similar structural features. Based on this, this paper designs an implicit emotion attribute classifier to simulate the emotion perception of the human brain. So, an implicit attribute binary classifier is designed according to whether the perception of different emotions by certain parts is related, as shown in Table 2. For example, if the frontal cortex of the human brain has a strong perception of happiness and anger, it is believed that happiness and anger have the same implicit attribute (denoted as attribute A), while other emotions (sadness and neutral, etc.) do not have attribute A. So in the binary classifier for attribute A, the classification label of happiness and anger is set to 1, while the classification label of other emotions is set to 0. In this paper, four parts with high degree of distinction are introduced, and four implicit attributes of A–D and corresponding classifiers are defined, as shown in the Table 2.

3.2 Multi-task learning based on implicit attribute classification

In order to train the four implicit attribute classification and speech emotion classifiers at the same time, this paper adopts the way of multi-task learning, and the loss of the implicit emotion attribute binary classification task will be added to the total loss of the model with a certain weight. At the same time, referring to the structured network characteristics of the human brain for emotion recognition, the network in this paper also introduces the implicit emotion attribute information, which increases the difference between different emotions, and is conducive to the network to recognize different emotions.

The specific structure of the network is shown in the Fig. 1. The network consists of four CNN layers, four categories of implicit emotional attributes, gated recurrent unit (GRU), and an attention layer. Firstly, the logMel spectrum extracted by Librosa [26] is used as the input of the network, and then the extracted features are input into four continuous CNN layers,

Table 2 Implicit attribute classification

Attribute	parts	Label 1	Label 0
A	Frontal cortex	Happy, angry	Sad, neutral
B	Thalamus	Sad	Happy, angry, neutral
C	Hippocampus	Sad, angry	Happy, neutral
D	Anterior neurite	Happy	Anger, sad, neutral

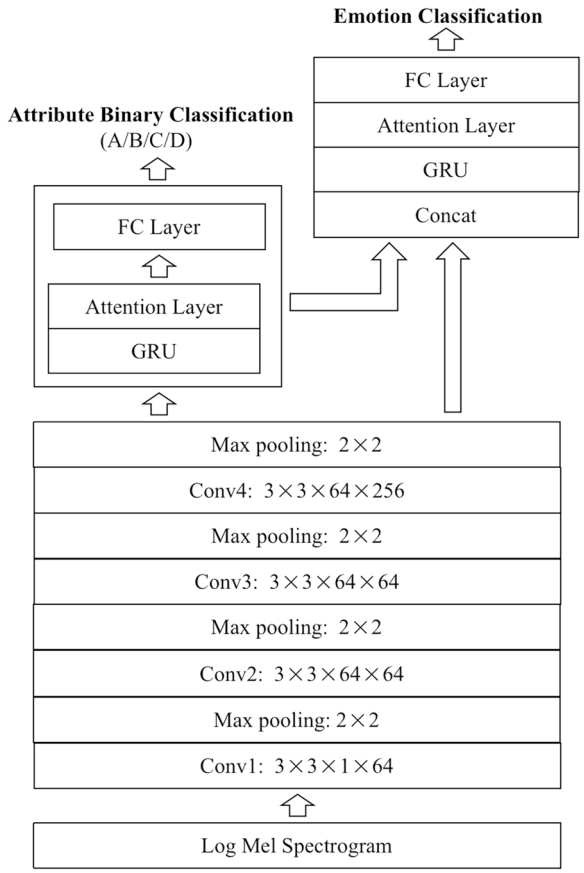

Fig. 1 Specific structure of the network

and then through the implicit emotion attribute classification network composed of the GRU and attention layer. The output of the attention layer of the binary classification network is combined with the output of the last CNN layer to get the input of the final emotion classification network. After classification, the emotion predicted by the model is obtained. Cross entropy loss function is used to optimize the network training. The attention layer in the network is conducive to extracting the difference of implicit emotional attribute information, simulating the difference in the sensitivity of different parts of the brain to different emotion perception, and reflecting the difference in the internal structure of different parts of the limbic system.

4 Experiment results and evaluations

4.1 Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [27] is the most widely used dataset in SER. It consists of 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, depending on whether the actors perform according to a fixed script. The dataset is labeled for 9 types of emotion-anger, excitement, happiness, sadness, frustration, fear, neutral, surprise, and other.

Because of the imbalance in the dataset, researchers usually choose the most common emotions, such as neutral state, happiness, sadness, and anger. Because excitement and happiness are similar to each other to a certain extent, and there are too few utterances about happiness, so researchers sometimes replace happiness with excitement or combine excitement and happiness to avoid the problem of too few utterances about happiness [28–30]. In addition, existing studies have shown that the accuracy of using improvised data is higher than that of using scripted data [28, 31] which may be due to the fact that actors in improvised data concentrate on emotional expression.

In this paper, we employ improvised data from the IEMOCAP dataset, which includes four types of emotion: excitement, sadness, neutral, and anger.

Table 3 Binary classification results of implicit attribute

Attribute categories	UA
A	84.59%
B	95.06%
C	77.54%
D	66.39%

4.2 Experiments on implicit attribute binary classification

In order to verify the reliability of the implicit attribute hypothesis, this paper first carries out a binary classification experiment of implicit emotion attributes, and the results are shown in Table 3. It can be seen that the experimental results of four kinds of implicit attribute binary classification experiments are relatively ideal, indicating that the hypothesis of implicit attribute has certain credibility. At the same time, the result of attribute B is the best, achieving 95.06%, while the result of attribute D is only 66.39%. The difference is relatively large, indicating that there are differences in the stability of different implicit attributes, or that different parts of the limbic system perceive emotional information differently. The internal structure of different parts leads to this difference. In Fig. 1, the weight of the attention layer reflects the difference of internal structure in different parts of the limbic system.

4.3 Emotion classification based on multi-task learning

In order to verify the effect of different implicit attributes on speech emotion recognition, different multi-task experiments are designed in this paper, namely, multi-task experiments based on 1–4 implicit attributes are respectively carried out. The experimental results are shown in Table 4. The following points can be inferred from Table 4.

1) The experimental performance of adopting four attributes is the best, achieving that the UA index is 2.42% higher than that of the baseline system (single task), and the WA index is 3.18% higher than that of the baseline system (absolute value), indicating the effectiveness of our method based on emotion perception. It is conducive to extract emotional information which draws lessons from the exploration results of brain science on the structure and function of the brain area that produces emotions in the human brain, combined with deep learning to simulate the neural network of the human brain.
2) In the multi-task experiment with single attribute, A, B, and C all are improved, but the introduction of attribute D reduces the performance of multi-task, which is consistent with the experimental results of binary classification. It is possible that the emotional information in implicit attribute D is unstable. However, in the mixed use experiment of attribute D and other attributes, the system performance is generally better than the baseline. It indicates that although the stability of attribute D is not good, it can play a positive role in emotion classification only with the assistance of other attribute information. This further verifies the credibility of the implicit attribute hypothesis in this paper. Meanwhile, the experimental results indirectly prove that the part where the human brain recognizes emotions may share some kind of information. When recognizing the same emotion, the level of sensitivity varies differently as for the parts of the limbic system. Combining a variety of information to jointly judge the emotional changes of the surrounding characters, this also further verifies the credibility of the implicit attribute hypothesis in this article.
3) The mixed use of two and three attributes has high and low experimental performance, indicating that the relationship between different implicit attributes may be complementary or cancel each other for the effect of emotion recognition. And the effect of using all four attributes is the best, demonstrating that the positive effect of these implicit attributes needs more attributes to participate. Different parts of the human limbic system have certain implicit emotional attributes. Experiments have shown that multiple parts are involved when recognizing a certain emotion, but the states of inhibition or activation when different parts recognize emotions are different.

Table 4 Emotion recognition results based on multi-task learning

Multi-task learning	UA	WA
A	68.57%	67.79%
B	68.91%	67.42%
C	68.31%	67.23%
D	67.44%	66.67%
A+B	68.61%	65.36%
A+C	67.46%	67.23%
A+D	67.91%	64.23%
B+C	67.25%	67.42%
B+D	68.75%	68.16%
C+D	68.23%	68.16%
A+B+C	67.84%	65.92%
A+B+D	68.30%	66.29%
A+C+D	68.14%	66.29%
B+C+D	69.38%	67.04%
A+B+C+D	*70.42%*	*67.79%*
Baseline (single task)	67.98%	64.61%

5 Conclusion

Brain science is constantly studying the brain structure and underlying mechanism of emotions. Combined with the continuous simulation of the human brain by artificial intelligence, this paper draws on the mechanism of the human brain emotional perception and designs the implicit emotional attributes classification to imitate the brain structure related to emotions. Implicit emotion information is introduced through multi-task learning as auxiliary information to recognize emotion, improving the effect of speech emotion recognition and proving the effectiveness of the network proposed in this paper. In the future, we can learn from the human brain's mechanism of cognitive emotions and add more attribute information. Meanwhile, we can also adopt different approaches instead of multi-task learning to mine emotional information.

Abbreviations

SER	Speech emotion recognition
UA	Unweighted accuracy
WA	Weighted accuracy
IEMOCAP	The Interactive Emotional Dyadic Motion Capture
RNN	Recurrent neural networks

DNN	Deep neural networks
CNN	Convolutional neural networks
LSTM	Long short-term memory
GRU	Gated recurrent unit

Acknowledgements
The authors thank the editors and the anonymous reviewers for their constructive comments and useful suggestions.

Authors' contributions
Liu innovatively proposed and designed the complete experiments for the paper and was a major contributor in writing the manuscript. Cai and Wang worked together to filter and sort emotional data and wrote programs to complete the entire experiments. After the experiment, all of us analyzed the whole experimental results to further corroborate our conjecture. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The Interactive Emotional Dyadic Motion Capture (IEMOCAP) is the most widely used dataset in SER. The dataset is accessible at https://sail.usc.edu/iemocap/. It consists of 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, depending on whether the actors perform according to a fixed script. The dataset is labeled for 9 types of emotion-anger, excitement, happiness, sadness, frustration, fear, neutral, surprise, and other. Our experiments choose four main emotions-anger, excitement, happiness, and sadness. Furthermore, the experimental code implementation is available at https://github.com/FlowerCai/speech-emotion-recognition. We can research further in the field of SER based on the experiment in the future.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that we have no competing interests.

Received: 5 April 2022 Accepted: 29 April 2023
Published online: 12 May 2023

References

1. L.S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans. Biomed. Eng. **58**(3), 574–586 (2010)
2. X. Huahu, G. Jue, Y. Jian, in Proceedings of the *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1. Application of speech emotion recognition in intelligent household robot, (IEEE, Sanya, 2010), pp. 537–541
3. W.J. Yoon, Y.H. Cho, K.S. Park, in *International Conference on Ubiquitous Intelligence and Computing*. A study of speech emotion recognition and its application to mobile services (Springer, Hong Kong China, 2007), pp. 758–766
4. K. Han, D. Yu, I. Tashev, in Proceedings of Interspeech 2014. Speech emotion recognition using deep neural network and extreme learning machine (ISCA, Singapore, 2014)
5. M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. **25**(10), 1440–1444 (2018)
6. X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, et al., in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech emotion recognition using capsule networks (IEEE, Brighton UK, 2019), pp. 6695–6699
7. Y. Xu, H. Xu, J. Zou, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hgfm: a hierarchical grained and feature model for acoustic emotion recognition (IEEE, Barcelona, 2020), pp. 6499–6503
8. D. Priyasad, T. Fernando, S. Denman, S. Sridharan, C. Fookes, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Attention driven fusion for multi-modal emotion recognition (IEEE, Barcelona, 2020), pp. 3227–3231
9. A. Nediyanchath, P. Paramasivam, P. Yenigalla, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition (IEEE, Barcelona, 2020), pp. 7179–7183
10. C.H. Park, D.W. Lee, K.B. Sim, Emotion recognition of speech based on rnn. Nurse Lead. **4**, 2210–2213 (2002). https://doi.org/10.1109/ICMLC.2002.1175432
11. J. Niu, Y. Qian, K. Yu, in *The 9th International Symposium on Chinese Spoken Language Processing*. Acoustic emotion recognition using deep neural network (IEEE, Singapore, 2014), pp. 128–132
12. Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimedia **16**(8), 2203–2213 (2014)
13. J. Lee, I. Tashev, in *Proceedings of Interspeech 2015*. High-level feature representation using recurrent neural network for speech emotion recognition (ISCA, Dresden Germany, 2015)
14. M.A. Jalal, E. Loweimi, R.K. Moore, T. Hain, in *Proceedings of Interspeech 2019*. Learning temporal clusters using capsule routing for speech emotion recognition (ISCA, Graz, 2019), pp. 1701–1705
15. R. Shankar, H.W. Hsieh, N. Charon, A. Venkataraman, in Proceedings of Interspeech 2019. Automated emotion morphing in speech based on diffeomorphic curve registration and highway networks(ISCA, Graz, 2019), pp. 4499–4503
16. S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, S. Nanayakkara, Multimodal emotion recognition with transformer-based self supervised feature fusion. IEEE Access **8**, 176274–176285 (2020)
17. S. Costantini, G. De Gasperis, P. Migliarini, in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. Multi-agent system engineering for emphatic human-robot interaction (IEEE, Sardinia Italy, 2019), pp. 36–42
18. H. Okon-Singer, T. Hendler, L. Pessoa, A.J. Shackman, The neurobiology of emotion-cognition interactions: fundamental questions and strategies for future research. Front. Hum. Neurosci. **9**, 58 (2015)
19. Q. Ma, D. Guo, Research on brain mechanisms of emotion. Adv. Psychol. Sci. **11**(03), 328 (2003)
20. S. Lee, S. Yildirim, A. Kazemzadeh, S. Narayanan, in *Ninth European Conference on Speech Communication and Technology*. An articulatory study of emotional speech production (ISCA, Lisbon Portugal, 2005)
21. J. LeDoux, Rethinking the emotional brain. Neuron **73**(4), 653–676 (2012)
22. V.R. Rao, K.K. Sellers, D.L. Wallace, M.B. Lee, M. Bijanzadeh, O.G. Sani, Y. Yang, M.M. Shanechi, H.E. Dawes, E.F. Chang, Direct electrical stimulation of lateral orbitofrontal cortex acutely improves mood in individuals with symptoms of depression. Curr. Biol. **28**(24), 3893–3902 (2018)
23. P. Fusar-Poli, A. Placentino, F. Carletti, P. Landi, P. Allen, S. Surguladze, F. Benedetti, M. Abbamonte, R. Gasparotti, F. Barale et al., Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J. Psychiatry Neurosci. **34**(6), 418–432 (2009)
24. F. Ahs, C.F. Davis, A.X. Gorka, A.R. Hariri, Feature-based representations of emotional facial expressions in the human amygdala. Soc. Cogn. Affect. Neurosci. **9**(9), 1372–1378 (2014)
25. M.D. Pell, Recognition of prosody following unilateral brain lesion: influence of functional and structural attributes of prosodic contours. Neuropsychologia **36**(8), 701–715 (1998)
26. B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, in *Proceedings of the 14th python in science conference*, vol. 8. librosa: audio and music signal analysis in python (SciPy, Texas US, 2015), pp. 18–25
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, in Proceedings of Interspeech 2018. An attention pooling based representation learning method for speech emotion recognition (ISCA, Hyderabad India, 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 UK, 2019), pp. 7390–7394
31. L. Tarantino, P.N. Garner, A. Lazaridis, et al., in *Proceedings of Interspeech 2019*. Self-attention for speech emotion recognition (ISCA, Graz, 2019), pp. 2578–2582

Publisher's Note

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

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com