Improvement On Speech Emotion Recognition Based on Deep Convolutional Neural Networks

Yafeng Niu[1] Dongsheng Zou[1*] Yadong Niu[2] Zhongshi He[1] Hua Tan[1]
1 College of Computer Science, Chongqing University, Chongqing 400044, China.
2. School of Electronics Engineering and Computer science. Peking University, Beijing 100871, China.
Email: dszou@cqu.edu.cn

ABSTRACT
Speech emotion recognition (SER) is to study the formation and change of speaker's emotional state from the speech signal perspective, so as to make the interaction between human and computer more intelligent. SER is a challenging task that has encountered the problem of less training data and low prediction accuracy. Here we propose a data processing algorithm based on the imaging principle of the retina and convex lens (DPARIP), to acquire the different sizes of spectrogram and get different training data by changing the distance between the spectrogram and the convex lens. Meanwhile, with the help of deep learning to get the high-level features, we apply the AlexNet on the IEMOCAP database and achieve the average accuracy over 48.8% on six emotions. The experimental results indicate that our proposed data preprocessing algorithm is effective and more accurate compared to existing emotion recognition algorithms.

CCS Concepts
•Human-centered computing→Human computer interaction (HCI)→Field studies

•Computing methodologies →Machine learning →Learning paradigms →Supervised learning by classification

Keywords
Speech emotion recognition; deep learning; speech spectrogram.

1. INTRODUCTION
SER is using computer to analyze the speaker's voice signal and its change process, to find their inner emotions and ideological activities, and finally to achieve a more intelligent and natural human-computer interaction (HCI), which is of great significance to develop new HCI system and to realize artificial intelligence [1] - [3].

Until now the method of SER can be divided into two categories: the traditional machine learning method and the deep learning method.

The key to the traditional machine learning method of SER is feature selection, which is directly related to the accuracy of recognition. By far the most common feature extraction methods include the pitch frequency feature, the energy-related feature, the formant feature, the spectral feature, etc. After feature extracted, the machine learning method is used to train and predict Artificial Neural Network (ANN) [4] – [7], Bayesian network model [8], Hidden Markov Model (HMM) [9] – [12], Support Vector Machine (SVM) [13], [14], Gauss Mixed Model (GMM) [15], and multi-classifier fusion [16], [17]. The primary advantage of this method is that it could train model without very large data. While the disadvantage is that it is difficult to judge the quality of the feature and may lose some key features, which will decrease the accuracy of recognition. In the meantime, it is difficult to ensure the good results can be achieved in a variety of databases.

Compared with the traditional machine learning method, the deep learning can extract the high-level features [18], [19], and it has been shown to exceed human performance in visual tasks [20], [21]. Currently, the deep learning has been applied to the SER by many researchers. Yelin Kim et al [22] proposed and evaluated a suite of Deep Belief Network (DBN) models, which can capture none linear features, and that models show improvement in emotion classification performance over baselines that do not employ deep learning. W Zheng et al [23] proposed a DBN-HMM model, which improves the accuracy of emotion classification in comparison with the state-of-the-art methods; Q Mao et al [24] proposed learning affect-salient features for SER using CNN, which leads to stable and robust recognition performance in complex scenes; Z Huang et al [25] trained a semi-CNN model, which is stable and robust in complex scenes, and outperforms several well-established SER features. And the accuracy is 78% on SAVEE database, 84% on EMODB database; K Han et al [26] proposed a DNN-ELM model, which leads to 20% relative accuracy improvement compared to the HMM model; Sathit Prasomphan [27] detected the emotional by using information inside the spectrogram, then using the Neural Network to classify the emotion of EMODB database, and got the accuracy is up to 83.28% of five emotions; W Zheng [28] also used the spectrogram with DCNNs, and achieves about 40% accuracy on IEMOCAP database; H. M Fayek [29] provided a method to augment training data, but the accuracy is less than 61% on ENTERFACE database and SAVEE database; Jinkyu Lee [30] extracted high-level features and used recurrent neural network (RNN) to predict emotions on IEMOCAP database and got about 62% accuracy, which is higher than the DNN model; S Zhang et al[31] proposed multimodal DCNNs, which fuses the audio and visual cues in a deep model, This is an early work fusing audio and visual cues in DCNNs; George Trigeorgis et al [32] combined CNN with LSTM networks, which can automatically learn the best representation of the speech signal from the raw time representation. Haytham M. Fayek [33] and Wootaek Lim [34] proposed a SER system to empiraclly explore feed-forward and RNN architectures and their variants. Seyedmahdad Mirsamadi [35] extracted features and used RNN with local attention to predict emotions on IEMOCAP database and got good predictions.

Though previous studies have achieved some results, the accuracy of recognition remains relatively low, and it is far from the practical application. In order to address the problems of small training data and low accuracy, this paper proposes a data preprocessing algorithm, we called Data Processing Algorithm Based on Retinal Imaging Principle (DPARIP), using the principle of retinal and convex lens imaging, we get more training data by changing the size of the spectrogram.

2. PROPOSED ALGORITHMS

As we all know, the closer we get to the object, the bigger we see it. In other words, what we see in our retina is different because of the different distance. But it doesn't affect our recognition. Since our brains have learned high-level features of the object, we can accurately identify the same thing of different sizes.

Figure 1. Single Lens Reflex (SLR) camera is used to simulate people's retina. And it is used to simulated the same thing from different distances on the retina. The closer to the girl, the bigger the image is, and vice versa.

In Figure 1 we use the SLR camera to simulate the same thing from different distances on the retina. We can find that the closer we get to the girl, the bigger the image is, and vice versa. However, it doesn't affect our judgment. Similarly, the DCNNs is constructed from the simulation of human neurons. Hence they could also make accurate judgments about the same thing in different sizes.

However, the training of deep neural network needs a large amount of data, while the data provided by current common speech emotion database is very limited. This leads to the problem that the deep neural network can't be fully trained. Referring to the imaging principle of the retinal and the convex lens, we propose DPARIP algorithm. The algorithm process is shown in Figure 2. And our work consisting of two parts.

1) Data Processing Algorithm Based on Retinal Imaging Principle (DPARIP). As shown in Table 1 and Fig.3.

2) Deep Convolution Neural Networks (DCNNs) [36]. We refer to the AlexNet [37] in the experiment and change the output of the 'fc8' fully connected layer to the number of emotions we want to classify. As shown in Figure 4. It has 5 convolution layers, 3 pooling layers and 3 fully connected layers. The processed spectrograms are the input of DCNNs. After training, the DCNNs can classify and predict the emotions.

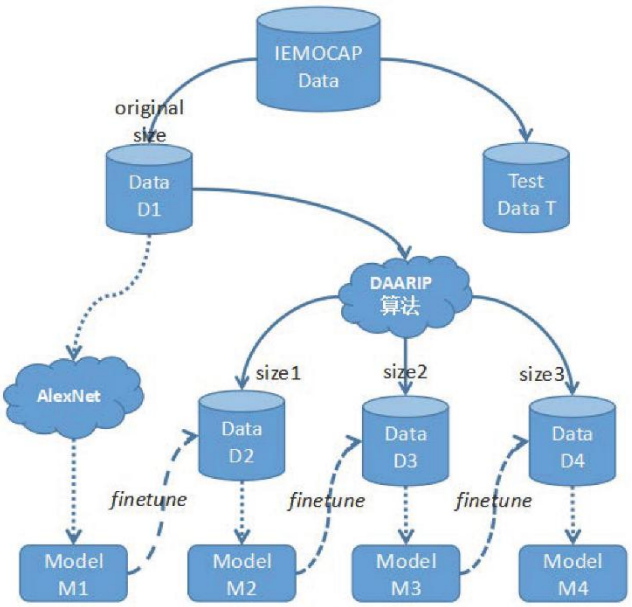

Figure 2. The working process of our work. Firstly, we randomly select 90% original data as training data D1, 10% original data as test data T. Secondly, the training data D1 is processed by DPARIP, and we can get some different size training data, such as training data D2, D3 and D4. Finally, we train Alexnet in the train data D1, and we can get model M1, then finetune the model M1 in the train data D2. Similarity, we can get model M2, M3, and M4, then use M4 to predict emotions.

Table 1. PSEUDO-CODE OF DPARIP ALGORITHM

DPARIP	
Input	Original training audio data.
Output	Spectrograms in different size.
Step1	Read audio data from file.
Step2	The speech spectrogram is obtained by short time Fourier transform. (nfft = 512, window = 512, numoverlap = 384)
Step3	According to the principle of retinal imaging and convex lens imaging, take x point at location L_1 (F<L_1<$2F$) and attain x images bigger than original. So we can get x training data.
Step4	Take y point at L_3 (L_3>$2F$) and attain y images smaller than original. So we can get y training data.
Step5	After Step3 and Step4, we can get $x+y+1$ times training data
Step5	Convert all images to size 256 * 256

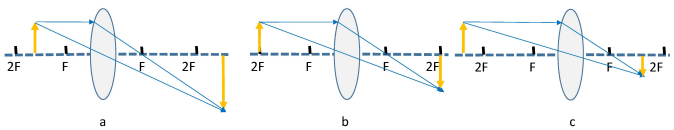

Figure 3. Using convex lens to simulate our eyes.

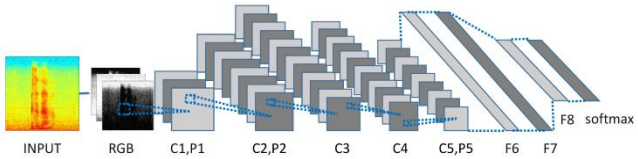

Figure 4. The DCNNs architecture for SER using the spectrogram as input, which has 5 convolution layers (C1~C5), 3 pooling layers (P1, P2, P5) and 3 fully connected layers (F6, F7, F8).

3. EXPERIMENTAL RESULT AND ANALYSIS

3.1 IEMOCAP Database Description

IEMOCAP database [38] contains audio and label data from 10 actors, including anger, happiness, sadness, neutral, frustration, excitement, fear, surprise, disgust and other. Each utterance is labeled by 3 annotators. If their feedbacks are inconsistent with one another, the data shall be invalid. In this paper, we select 6 kinds of emotions without regard to the influence of gender. These emotions are angry, excited, frustrated, happy, neutral and sad, and the detail of the dataset description is shown in table 2.

Table 2. Data Description of IEMOCAP DATABASE

Emotion	Train data	Validation data	Test data
Frustration	1490	166	193
Neutral	1370	152	186
Sad	886	98	123
Angry	882	98	117
Excitement	832	92	100
Happy	479	53	63

3.2 Data Pre-processing

Before the experiment, data pre-processing is an important step. Firstly, we count the distribution of the size of each emotion spectrogram, and we get that, each spectrogram's height is fixed, width is different. Secondly, based on the width, we split each emotion data into four parts, then we select 10% of the data of each part as the test data T. the other data as train data D1. As shown in Figure 5 and Figure 6.

3.3 First Experiment

In the experiment 1, we randomly select 10% data D1 as validation data V1, and the other as train data T1, after 65 epochs training on the train data T1, and the highest accuracy of the validation data V1 is achieved to 45.33% in epoch 63. The highest accuracy of the test data T is achieved to 43.73% in epoch 65. So we can get model M1 in epoch 65. As shown in Figure 7, the parameter of first experiment is shown in table 3.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

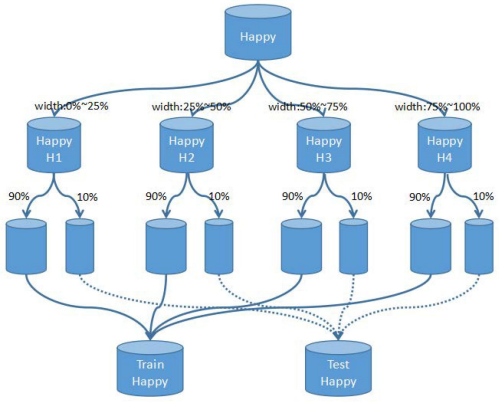

Figure 5. The flowchart of happy emotion data pre-processing. Similarity, we process angry, excitement, frustrated, neutral and sad emotion.

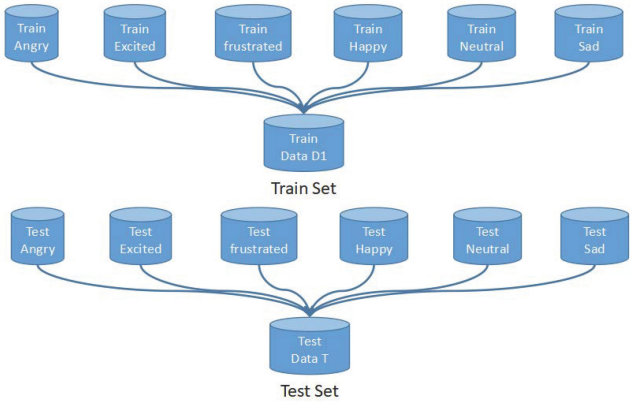

Figure 6. The experiment's train data D1 and test data T.

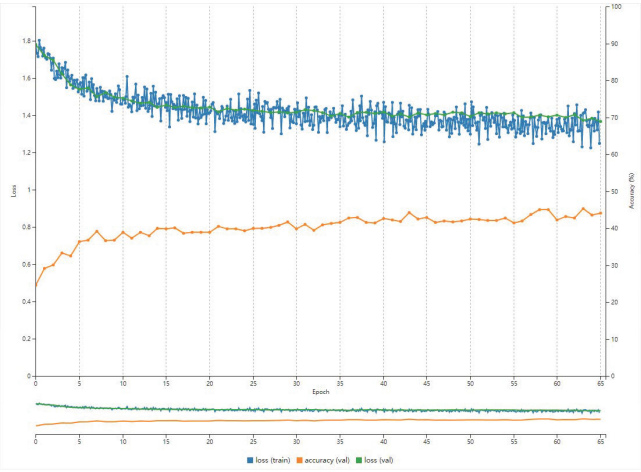

Figure 7. The experiment on the data D1 (original size).

Table 3. main parameter of the first experiment

Parameter name	**Parameter value**
base learning rate	0.01
learning rate policy	multistep
gamma	0.1

momentum	0.9
step values	20%
weight decay	1e-05
solver type	SGD

After the first experiment, the data D1 is processed by the DPARIP algorithm, then we can get data D2, D3, D4, and these data sizes are different to the data D1, as shown in Figure 8.

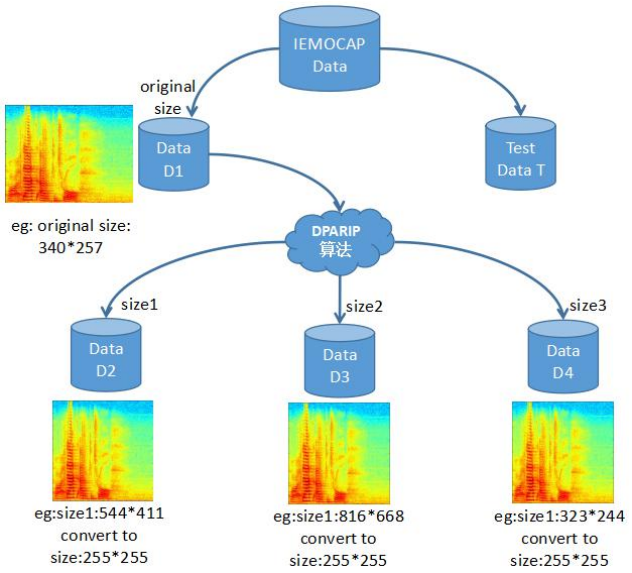

Figure 8. The data D1 is processed by the DPARIP algorithm, we can get data D2, D3, and D4.

3.4 Second Experiment

In the experiment 2, we select model M1 to finetune the data D2. Similarity, we choose 10% data D2 as validation data V2, and the other as train data T2. After 65 epochs training on the train data T2, and the highest accuracy of the validation data V2 is achieved to 55.07% in epoch 56. The highest accuracy of the test data T is achieved to 46.93% in epoch 60. So we can get model M2 in epoch 60. As shown in Figure 9. The main parameters of second experiment is shown in table 4.

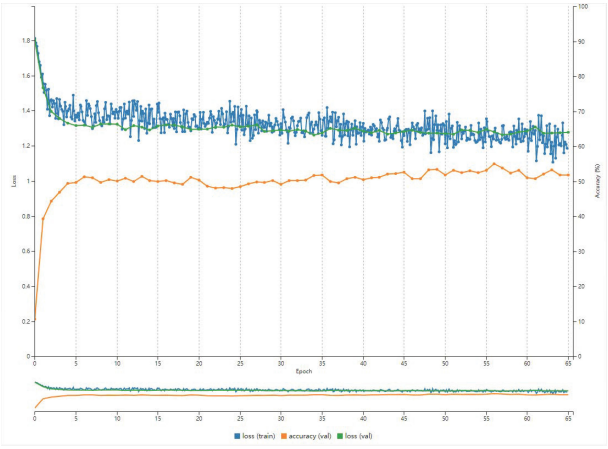

Figure 9. The experiment on the data D2.

Table 4. main parameter of the second experiment

Parameter name	Parameter value
base learning rate	0.001
learning rate policy	fixed
momentum	0.9
weight decay	1e-05
solver type	SGD

3.5 Third Experiment

In the experiment 3, we select model M2 to finetune the data D3, Similarly, we choose 10% data D3 as validation data V3, and the other as train data T3. After 65 epochs training on the train data T3, and the highest accuracy of the validation data V3 is achieved to 50.13% in epoch 46. The highest accuracy of the test data T is achieved to 48.08% in the epoch 40. So we get model M3 in epoch 40. We also find that the 46 epoch later, the model tends to be overfitting. As shown in Figure 10. The main parameters of third experiment is same to the second experiment.

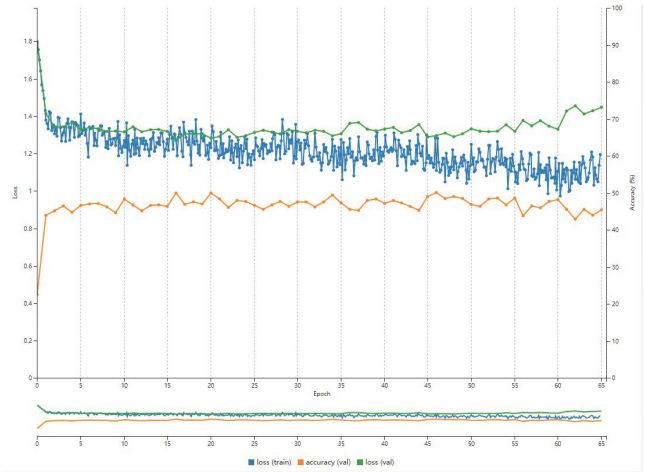

Figure 10. The experiment on the data D3, the 46 epoch later, the model tends to be overfitting.

3.6 Forth Experiment

In the experiment 4, we select model M3 to finetune the data D4. Similarity, we choose 10% data D4 as validation data V4, and the other as train data T4. After 40 epochs training on the train data T4, and the highest accuracy of the validation data V4 is achieved to 58.93% in epoch 26. The highest accuracy of the test data T is achieved to 48.85% in epoch 12. So we can get model M4 in epoch 12. As shown in Figure 11. The main parameter of the forth experiment is shown in table 5.

Table 5. main parameter of the forth experiment

Parameter name	Parameter value
base learning rate	0.001
learning rate policy	multistep
gamma	0.1
momentum	0.9
step values	50%
weight decay	1e-05
solver type	SGD

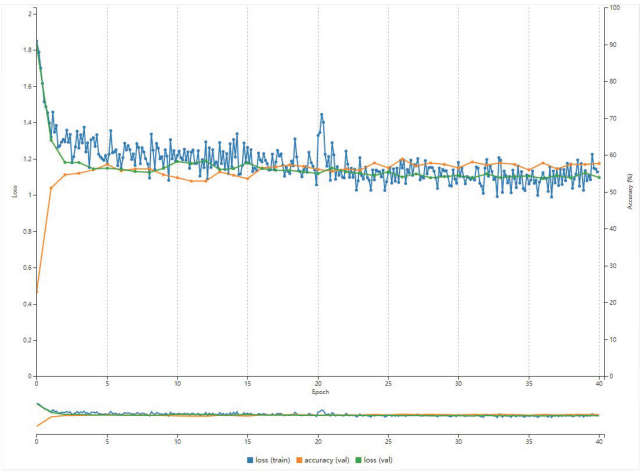

Figure 11. The experiment on the data D4.

Compared with the recent related studies, we can find that our results are better from both the number of emotions and accuracy, the detail is shown in Table 6. And Generally speaking, the emotion classification accuracy is unable to achieve better on IEMOCAP, since the data distribution of each emotion from the database is imbalanced.

Table 6. Compared with other studies on IEMOCAP database.

Method	Emotions	Accuracy
Ref[28]	Excitement, Happiness, Frustration, Neutral, Surprise	40.02%
Experiment 1	Excitement, Happiness, Frustration, Neutral, Surprise, Anger	43.73%
Experiment 2	Excitement, Happiness, Frustration, Neutral, Surprise, Anger	46.93%
Experiment 3	Excitement, Happiness, Frustration, Neutral, Surprise, Anger	48.08%
Experiment 4	Excitement, Happiness, Frustration, Neutral, Surprise, Anger	48.85%

4. CONCLUSION AND FUTURE WORK

SER is particularly useful for enhancing naturalness in speech based on human machine interaction. SER system has extensive applications in day-to-day life. For example, emotion analysis of telephone conversation between criminals would help criminal investigation department to detect cases. Conversation with robotic pets and humanoid partners will become more realistic and enjoyable if they are capable of understanding and expressing emotions like humans do. Automatic emotion analysis may be applied to automatic speech to speech translation systems, where speech in one language is translated into another language by the machine.

In this paper, we propose a novel method called DPARIP, addressing the problem of small training data, and we use AlexNet to train model. Finally, we obtain about 48.85% classification accuracy. Our future work is to extend the proposed method and evaluate its performance on multilingual speech emotion database.

5. ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (No. 61309013) and Chongqing Basic and frontier research projects. (No. CSTC2014JCYJA40042).

6. REFERENCES

[1] Luo, Q. "Speech emotion recognition in E-learning system by using general regression neural network." Nature 153.3888(2014):542-543.

[2] Koolagudi, Shashidhar G., and K. S. Rao. "Emotion recognition from speech: a review." International Journal of Speech Technology 15.2(2012):99-117.

[3] Ayadi, Moataz El, M. S. Kamel, and F. Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3(2011):572-587.

[4] Wang, Sheguo, et al. "Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network." Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on 2010:437-440.

[5] Bhatti, M. W., Y. Wang, and L. Guan. "A neural network approach for human emotion recognition in speech." International Symposium on Circuits and Systems IEEE Xplore, 2004: II-181-4 Vol.2.

[6] Fragopanagos, N., and J. G. Taylor. "2005 Special Issue: Emotion recognition in human-computer interaction." Neural Networks 18.4(2005):389-405.

[7] Nicholson, J., K. Takahashi, and R. Nakatsu. "Emotion Recognition in Speech Using Neural Networks." International Conference on Neural Information Processing, 1999. Proceedings. ICONIP IEEE, 1999:495-501 vol.2.

[8] Ververidis, Dimitrios, and C. Kotropoulos. "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition." Signal Processing 88.12(2008):2956-2970.

[9] Mao, Xia, L. Chen, and L. Fu. "Multi-level Speech Emotion Recognition Based on HMM and ANN." Computer Science and Information Engineering, 2009 WRI World Congress on 2009:225-229.

[10] Nwe, Tin Lay, S. W. Foo, and L. C. D. Silva. "Speech emotion recognition using hidden Markov models." Speech Communication 41.4(2003):603-623.

[11] Schuller, B., G. Rigoll, and M. Lang. "Hidden Markov model-based speech emotion recognition." International Conference on Multimedia and Expo, 2003. ICME '03. Proceedings IEEE Xplore, 2003: I-401-4 vol.1.

[12] Ntalampiras, Stavros, and N. Fakotakis. "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition." IEEE Transactions on Affective Computing 3.99(2012):116-125.

[13] Zhou, Jian, et al. "Speech Emotion Recognition Based on Rough Set and SVM." International Conference on Machine Learning and Cybernetics 2005:53-61.

[14] Hu, Hao, M. X. Xu, and W. Wu. "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition." IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP IEEE, 2007: IV-413-IV-416.

[15] Neiberg, Daniel, K. Laskowski, and K. Elenius. "Emotion Recognition in Spontaneous Speech Using GMMs."

INTERSPEECH 2006- ICSLP (2006):1 - 4.
[16] Wu, Chung Hsien, and W. B. Liang. "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels." IEEE Transactions on Affective Computing 2.1(2011):10-21.
[17] Schuller, B., et al. "Speaker Independent Speech Emotion Recognition by Ensemble Classification." IEEE International Conference on Multimedia & Expo IEEE, 2005:864-867.
[18] Bengio, Yoshua, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives." IEEE Transactions on Pattern Analysis & Machine Intelligence 35.8(2013):1798-828.
[19] Lecun, Y, Y. Bengio, and G. Hinton. "Deep learning. " Nature 521.7553 (2015):436-44.
[20] Mnih, V, et al. "Human-level control through deep reinforcement learning." Nature 518.7540(2015):529-533.
[21] Silver, D, et al. "Mastering the game of Go with deep neural networks and tree search." Nature 529.7587(2016):484.
[22] Kim, Yelin, H. Lee, and E. M. Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2013:3687-3691.
[23] Zheng, Wei Long, et al. "EEG-based emotion classification using deep belief networks." IEEE International Conference on Multimedia & Expo IEEE, 2014:1-6.
[24] Mao, Q., et al. "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks." Multimedia IEEE Transactions on 16.8(2014):2203-2213.
[25] Huang, Zhengwei, et al. "Speech Emotion Recognition Using CNN." the ACM International Conference 2014:801-804.
[26] Han, Kun, D. Yu, and I. Tashev. "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine." INTERSPEECH 2014.
[27] Prasomphan, S. "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram." International Conference on Systems, Signals and Image Processing IEEE, 2015:73-76.
[28] Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." International Conference on Affective Computing and Intelligent Interaction 2015:827-831.
[29] Fayek, H. M., M. Lech, and L. Cavedon. "Towards real-time Speech Emotion Recognition using deep neural networks." *International Conference on Signal Processing and Communication Systems* 2015:1-5.
[30] Lee, Jinkyu, and I. Tashev. "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition." INTERSPEECH 2015.
[31] Zhang, Shiqing, et al. "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition." ACM, 2016:281-284.
[32] Trigeorgis, George, et al. "Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network." ICASSP 2016.
[33] Fayek H M, Lech M, Cavedon L. Evaluating deep learning architectures for Speech Emotion Recognition. [J]. Neural Networks, 2017.
[34] Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and Recurrent Neural Networks[C]// Signal and Information Processing Association Summit and Conference. IEEE, 2016:1-4.
[35] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017.
[36] Esteva, A, et al. "Dermatologist-level classification of skin cancer with deep neural networks." Nature 542.7639(2017):115.
[37] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." Advances in Neural Information Processing Systems 25.2(2012):2012.
[38] Busso, Carlos, et al. "IEMOCAP: interactive emotional dyadic motion capture database." Language Resources and Evaluation 42.4(2008):335-359.

Columns on Last Page Should Be Made As Close As Possible to Equal Length

Authors' background

Your Name	Title*	Research Field	Personal website
Yafeng Niu	**master student**	machine learning, deep learning, affective computing	
Dongsheng Zou	**assistant professor**	machine learning, data mining, pattern recognition	
Yadong Niu	**phd candidate**	machine learning, deep learning, signal processing strategies	
Zhongshi He	**full professor**	machine learning, data mining, natural language computing, image processing	
Hua Tan	**master student**	machine learning, deep learning, affective computing	

***This form helps us to understand your paper better, the form itself will not be published.**

***Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor**