Journal of Robotics, Networking and Artificial Life
Vol. **8**(**3**); December (2021), *pp.* 218–223
DOI: https://doi.org/10.2991/jrnal.k.210922.013; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

Research on LSTM+Attention Model of Infant Cry Classification

Tianye Jian, Yizhun Peng*, Wanlong Peng, Zhou Yang

College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300222, China

ARTICLE INFO

Article History

Received 16 November 2020
Accepted 28 July 2021

Keywords

Children's emotional
time–sequence relationship
frame-level speech feature
deep attention gate
long- and short-term memory (LSTM)

ABSTRACT

According to the different emotional needs of infants, the effective acquisition of frame-level speech features is realized, and the infant speech emotion recognition model based on the improved Long- and Short-Term Memory (LSTM) network is established. The frame-level speech features are used instead of the traditional statistical features to preserve the temporal relationships in the original speech, and the traditional forgetting and input gates are transformed into attention gates by introducing an attention mechanism, to improve the performance of speech emotion recognition, the depth attention gate is calculated according to the self-defined depth strategy. The results show that, in Fau Aibo Children's emotional data corpus and baby crying emotional needs database, compared with the traditional LSTM based model, the recall rate and *F*1-score of this model are 3.14%, 5.50%, 1.84% and 5.49% higher, respectively, compared with the traditional model based on LSTM and gated recurrent unit, the training time is shorter and the speech emotion recognition rate of baby is higher.

1. INTRODUCTION

An important part of affective computing includes children's emotion recognition [1]. Children are far less able than adults to act rationally in emotional outbursts and in response to different emotions, which can lead to emotional disorders if children are not able to act rationally and are directed in a timely manner, which can lead to anxiety disorders and other mental health problems. Therefore, it is of great significance to use appropriate algorithms to judge children's emotions.

Researchers have conducted in-depth research on children's emotion recognition using methods such as acoustic features, machine learning, and deep learning [2]. It is proposed to use Support Vector Machine (SVM) and Convolutional Neural Network (CNN) to construct a system for detecting children's secondary emotional states [3] real-time emotional state of children is defined by multi-agent based interaction system [4] to establish the children's dual-modal emotion database and use the dual-modal emotion recognition method to measure the proportion of children's emotion contribution, and to point out that infants' (or young children's) emotion is more difficult to judge than older children's, babies usually cry to express their needs to their parents or guardians [5] the Mel-Frequency Cepstrum Coefficients (MFCC) of infant cries were extracted and classified based on Hidden Markov Model (HMM) to identify whether the infant cries were healthy or not [6]. The Spectrogram was used as the feature vector, and the CNN was selected as the classification model to classify and recognize the crying of infants in pain, hunger and sleepiness [7]. The SVM was used as the classifier to classify the crying sounds of infants in the condition of hunger, pain and sleepiness, and the recognition effect was better than that of the SVM.

The above algorithms have been successfully applied in the field of children's emotion recognition, but traditional machine learning algorithms, as well as autoencoders and CNNs in deep learning, can only accept fixed-dimensional data as input. This is in contradiction with the fact that the effective length of speech is constantly changing. To solve this problem, reference extracts emotion-related features (hereinafter referred to as frame-level features) from short-term speech frames, and applies static statistical functions (such as mean, variance, maximum, linear regression coefficient) to frame-level features, finally, the feature vectors with fixed dimensions are formed to represent the features of the frame speech [8–10]. Although this method solves the problem of model input, the time sequence information of the original speech is lost through the statistical analysis of the speech features.

We propose an improved Long- and Short-Term Memory (LSTM) based children's speech emotion recognition model. Based on the LSTM network structure, the frame-level speech features are substituted for the traditional statistical features, in order to obtain better recognition performance, attention gates were used to replace the traditional forgetting gates and input gates, and to construct the deep attention gates by weighting attention in multiple cellular states.

*Corresponding author. Email: pengyizhun@tust.edu.cn

2. RELATED WORK

2.1. LSTM Network

Long- and short-term memory network is mainly used to process sequence information with a large time difference. It is a variant of the Recurrent Neural Network (RNN). The LSTM network can solve the problem that the long-term information is difficult to store because of the gradient disappearance of RNN in the reverse propagation [11–13]. The LSTM network has been successfully applied in Natural Language Processing (NLP) [14–16]. To enhance the ability of LSTM network to process data in specific tasks, the researchers further optimized the internal structure of LSTM network. The fusion of the input gate and the forgetting gate of the LSTM network by the Gated Recurrent Unit (GRU) reduces the model parameters [17], the performance of LSTM network is better than that of GRU [18] in all machine translation tasks [19]. By using CONVLSTM network structure, the computing method of gate structure of LSTM is improved from Matrix multiplication to convolution [20]. Infinite Impulse Response (IIR) Filter Memory block of RNN is improved to Finite Impulse Response (FIR) Filter Memory block by Feedforward Sequential Memory Network (FSMN) [21], but FSMN usually needs to stack very deep layers, so FSMN has delay compared with one-way LSTM network [22]. Advanced LSTM network, which uses attention mechanism to weight multiple cell states, can be effectively used for emotion recognition. However, Xie et al. [23] indicates that this method does not change the gate structure in the LSTM network and requires more training time. In addition, researchers are exploring how to stack LSTM structures to achieve more reliable emotion recognition [24]. End-to-end emotion recognition is achieved by extracting multi-channel speech features from a 6s-long speech waveform via a CNN as input to the LSTM Network [25]. In this paper, 1280 kinds of abstract features were extracted from 6s-long speech waveform by CNN, and then fused with facial features as input of LSTM network [26].

The formula used in the traditional LSTM network is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$C_t = \tan h(W_C \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \tag{4}$$

$$O_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{6}$$

σ is the sigmoid activation function, h_{t-1} is the hidden layer output at $t-1$, x_t is the input at t, C_{t-1} is the cell state at $t-1$, C_t is the candidate value of the cell state at t, f_t, i_t and o_t are respectively forgetting gate, input gate and output gate.

2.2. Deep Attention Gate

Time series information will increase over time. Therefore, the calculation of the LSTM model at a certain time (that is, update the cell state C and hidden layer output h) is only based on the external input and the previous unit state and hidden layer output. Before the attention mechanism is put forward, if each moment is considered several times before, it will lead to too much information and the loss of important information, and increase the calculation amount and lead to the gradient explosion [27]. However, the information of cell state at t-time is not only related to the information at $t-1$, but also closely related to the information at $t-2$, which is selectively forgotten at $t-1$. In this paper, the concept of deep forgetting gate is proposed and the corresponding input gate is designed.

The deep forgetting gate looks not only at the information about the state of the cell at the last moment (depth length = 1), but also at the information about the state of the cell at the time $t-2$, $t-3$, …, $t-n$ (depth length = n), which constructs the Deep-Attention-LSTM Cell structure, as shown in Figure 1.

It is worth noting that the introduction of "depth" will lead to an increase in training time. This is because in addition to forward increasing the number of attention gates in the loop for computing the state of multiple cells, reverse propagation also increases the number of chain derivations [28,29]. From the point of view of the parameters of the model, although the depth will cause the increase of the training time, it will not cause the increase of the parameters of the model because the attention gate weights of each layer are shared by V and W.

In this paper, the aim of depth is to improve the performance of speech emotion recognition [30]. To study the performance improvement, the following experiments were carried out:

In experiment 1 investigated the effect of depth performance on children's emotion recognition rate [31]. The attention-gate-based LSTM model (hereafter referred to as the attention-gate LSTM model) at depths 1–3 is used for comparison.

In experiment 2, the effects of decreasing the number of parameters and training time on speech emotion recognition performance were investigated [32]. The attention gate LSTM model with depth 1 was compared with the traditional GRU model and LSTM model.

2.3. Training Framework

The training framework for the deep attention gate LSTM model is shown in Figure 2. Among them, LSTM0 represents the LSTM model of the first-level deep attentional gate, and LSTM1 represents the LSTM model of the second-level deep attentional gate. X_t extracts speech features [8–10] from INTERSPEECH in frame t after framing and windowing, h_t and C_t are the hidden layer output

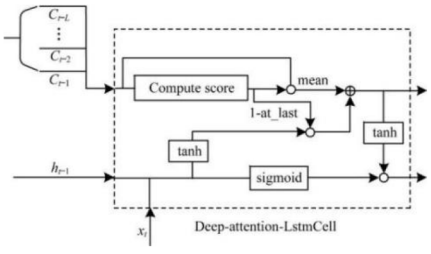

Figure 1 | Deep-attention-LSTM cell internal structure sketch.

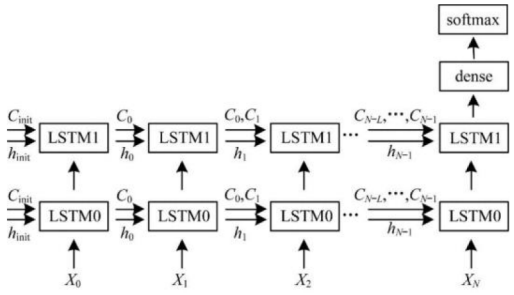

Figure 2 | Training framework of depth of attention gate LSTM model.

and cell state of its corresponding LSTM model output. As can be seen from Figure 2, the input state of the traditional LSTM model at time t is (h_{t-1}, C_{t-1}), while in the training of this paper, the input state at each time is expanded to $(h_{t-1}, \{C_{t-1}, C_{t-2}, \ldots, C_{t-L}\})$, where L is the depth of the attention gate. The last state of the last layer of LSTM, which contains all the temporal information of the preorder, is input into the subsequent classification network for the identification of children's emotions [33–35].

3. EXPERIMENTAL SETUP AND ANALYSIS

3.1. Experimental Setup

In our experiment, we used two databases of different emotion representations to verify the effectiveness of the algorithm in this paper. To study the performance of this algorithm in dealing with other types of emotion recognition problems, and whether reducing the number of parameters can optimize the time or reduce the performance [36–39], Fau Aibo Children's affective Corpus, infant crying affective needs Corpus and Chinese Academy of Sciences' Institute of Automation (CASIA) Chinese affective Corpus were used to verify the results.

3.2. Frame Level Feature Selection

In this experiment, some frame-level features are selected based on INTERSPEECH's speech emotion features [39]. In Schuller et al. [8], 16 Low-Level Descriptors (LLD, zero-crossing rate, root-mean-square Frame Energy, pitch frequency and MFCC 1–12) and their difference coefficients are extracted. For each descriptor, 12 statistical functions are calculated, so the total eigenvector has $16 \times 2 \times 12 = 384$ features. On this basis, the speech emotion characteristics of INTERSPEECH 2010 (IS2010) have been increased to 38 kinds of LLD, so the total feature dimension has been expanded to 1582 dimensions. The feature dimension of the speech feature set is increased to 6373 dimensions.

3.3. Setting of Experimental Parameters

The original data is divided into training set and test set according to the ratio of 4:1. The experiments all adopt unidirectional two-layer LSTM stack structure and use one full connection layer and one Softmax layer as the training model. During the training, a small batch of gradient descent is used and Tanh is used as the activation function, as shown in Table 1. To ensure the validity of experimental comparison, the same Corpus and model experimental parameters are identical.

4. ALGORITHM PERFORMANCE ANALYSIS

The traditional LSTM model removes redundant information through forgetting gates, and obtains new information through input gates. In this paper, we use the self-attention and the basic structure of LSTM to do self-attention to the cell state, so as to compare the forgetting gate and the input gate of LSTM. At the same time, considering the correlation of time series information, the depth-based self-attention Gates are proposed and compared at the depth of 1–3. Four kinds of models were compared: Traditional model, LSTM + deepf_1 model, LSTM + deepf_2 model, and LSTM + deepf_3 model. The depth distributions of these models are 0–3, as shown in Figure 3. As can be seen from Figures 3 and 4, after replacing the forgetting door and the

Table 1 | Experimental parameters

Name of parameter	Baby crying needs corpus parameter values	Fau Aibo	CASIA
Eta	$1e^{-3}$ beta2 = 0.7	$1e^{-4}$ beta2 = 0.9	$1e^{-4}$ beta2 = 0.9
Batch size	64	128	128
Epochs	1500	1200	1200
LSTM cells	[512, 256]	[512, 256]	[512, 256]
Dense layers	[256, 128]	[256, 128]	[256, 128]
Softmax layers	[128, 5]	[128, 5]	[128, 6]
L2	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$

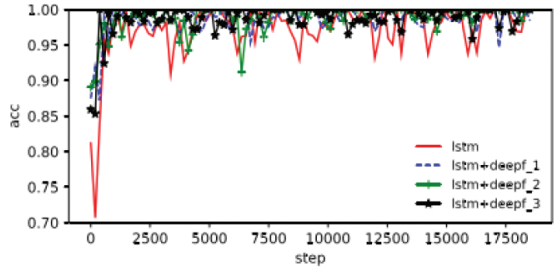

Figure 3 | The performance of infant crying emotional needs corpus in different LSTM models.

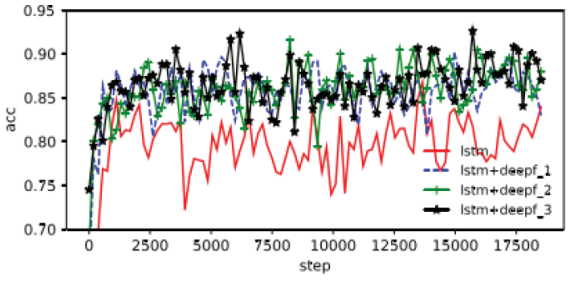

Figure 4 | Performance of different LSTM models using Fau Aibo children's emotion corpus database.

output door of the traditional LSTM model with the attention door proposed by using the infant crying affective needs corpus and Fau Aibo children affective Corpus, the convergence rate of the attention gate LSTM model on the training set and the test set is much higher than that of the traditional LSTM model, when Fau Aibo is used in children's affective Corpus, the traditional LSTM model starts to converge at about 30,000, while the attention gate model starts to converge at about 17,000 when the model converges, the attention-gate LSTM model outperforms the traditional LSTM model in the average recognition rate of children's emotion.

According to the analysis above, the performance of attention gate LSTM model is improved because it modifies the forgetting gate and input gate of traditional LSTM model, it makes the LSTM model leave important information by self-attention to the cell state at the last moment, and supplement the unimportant information as the new input in the corresponding position, thus improving the performance of the LSTM model. The attention gate LSTM model introduces the concept of depth so that each forgetting operation is determined by multiple cell states rather than one of them.

To compare the performance of different models in the test set for each kind of emotion, the performance index of the model with the highest recognition rate from the beginning of training to the end of the test set was Quantitative analysis, the performance indicators obtained from the infant crying affective needs Corpus and Fau Aibo children affective corpus are shown in Tables 2 and 3. It can be seen that for the test set, the performance index of the attention gate LSTM model is better than the traditional LSTM model.

From Table 2, it can be seen that the recall rate of attention-gate LSTM model is better than that of traditional LSTM model except that the items of drowsiness and sleepiness are close to each other The *F*1-scores of attention gate LSTM model were better than those of traditional LSTM model in five kinds of emotion. In the aspect of depth, the performance of the attention-gate LSTM model of depth 3 and 2 is close to that of the attention-gate LSTM model of depth 1, except "sad", the recall rate and *F*1-score of the other four items of the model are better than those of depth 1.

From Table 3, it can be seen that when FAU AIBO is used in children's affective corpus, the recall rate and *F*1-score of attention gate LSTM model are lower than those of traditional LSTM model except that the class E is lower than that of traditional LSTM model, and the other four items are better than traditional LSTM model. In the aspect of depth, the performance of the attentional gate LSTM model of depth 3 and 2 is close to that of the attentional gate LSTM model of depth 1, except for class R, the recall rate and F1 score of the other four items of the attentional gate LSTM model are better than those of depth 1.

It should be noted that the sample size of Fau Aibo Children's affective corpus is uneven, with a maximum of 5376 samples in *n* category and only 215 samples in *P* category. From the above analysis, with the increase of depth, the model can enhance the learning of a small number of samples. Compared with the traditional LSTM model, the recall rate of LSTM + deepf_1model was increased by 5.50%, and the *F*1-score was increased by 5.49%, and the recall rate of LSTM + deepf_2 model was increased by 3.14% when Fau Aibo was used, the *F*1-score of LSTM + deepf_3 model was increased by 1.84%.

Table 2 | Performance indicators of different LSTM models using emotional needs corpus of infant crying database

Model	Measure	Angry	Hungry	Pain	Sad	Tired	AVG
Sample size	—	40	45	34	35	46	200
LSTM	Recall	0.875	0.911	0.824	0.829	0.957	0.885
	*F*1-score	0.875	0.901	0.848	0.906	0.889	0.885
LSTM+deepf_1	Recall	0.875	0.889	0.882	1.000	0.935	0.915
	*F*1-score	0.909	0.941	0.822	0.946	0.945	0.916
LSTM+deepf_2	Recall	0.925	0.978	0.882	0.943	0.957	0.940
	*F*1-score	0.949	0.957	0.923	0.930	0.936	0.940
LSTM+deepf_3	Recall	0.925	0.978	0.824	0.943	0.957	0.930
	*F*1-score	0.949	0.926	0.875	0.943	0.946	0.930

Table 3 | Performance indicators of different LSTM models using Fau Aibo children's emotion corpus database

Model	Measure	A	E	N	P	R	AVG
Sample size	—	611	1508	5376	215	546	8256
LSTM	Recall	0.352	0.373	0.755	0.088	0.081	0.594
	*F*1-score	0.339	0.373	0.743	0.110	0.093	0.586
LSTM+deepf_1	Recall	0.326	0.300	0.814	0.153	0.119	0.621
	*F*1-score	0.358	0.338	0.770	0.173	0.133	0.603
LSTM+deepf_2	Recall	0.339	0.289	0.826	0.158	0.081	0.625
	*F*1-score	0.342	0.340	0.772	0.173	0.104	0.601
LSTM+deepf_3	Recall	0.360	0.327	0.800	0.191	0.095	0.619
	*F*1-score	0.371	0.363	0.765	0.192	0.112	0.604

5. CONCLUSION

In this paper, a child speech emotion recognition model based on improved LSTM network is proposed frame-level speech features are used instead of traditional speech features, the attention mechanism is introduced into the forgetting gate and the input gate of the internal structure of the LSTM network model to form the attention gate. The experimental results show that the recognition rate of this model is significantly higher than that of traditional LSTM model, and the recognition rate of depth model is higher than that of shallow model. In CASIA database with other emotions, the training time of this model is shorter than that of LSTM model, and the recognition rate is higher than that of LSTM model and GRU model. The next step is to introduce this model into the fields of speech recognition, machine translation and lie detection, to test and study the continuous affective Corpus and to improve the model for calculating attention scores, to further improve children's speech emotion recognition rate.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] L. Fang, C. Guopeng, An overview on the development of children's emotion regulation, Psychol. Sci. 26 (2003), 928–929 (in Chinese).
[2] Y. Gong, C. Poellabauer, Continuous Assessment of Children's Emotional States using Acoustic Analysis, 2017 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Park City, UT, USA, 2017, pp. 171–178.
[3] P.R. De Silva, A.P. Madurapperuma, A. Marasinghe, M. Osano, A Multi-agent Based Interactive System Towards Child's Emotion Performances Quantified Through Affective Body Gestures, 18th International Conference on Pattern Recognition (ICPR'06), IEEE, Hong Kong, China, 2006, pp. 1236–1239.
[4] W. Dai, Research on Expression and Speech Bimodal Emotion Recognition of Children, Southeast University, Nanjing, 2016 (in Chinese).
[5] D. Lederman, A. Cohen, E. Zmora, K. Wermke, S. Hauschildt, A. Stellzig-Eisenhauer, On the use of hidden Markov models in infants' cry classification, The 22nd Convention on Electrical and Electronics Engineers in Israel, IEEE, Tel-Aviv, Israel, 2002, pp. 350–352.
[6] C. Chuan-Yu, J.J. Li, Application of deep learning for recognizing infant cries, 2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), IEEE, Nantou, Taiwan, 2016, pp. 1–2.
[7] C.Y. Chang, C.W. Chang, S. Kathiravan, C. Lin, S.T. Chen, DAG-SVM based infant cry classification system using sequential forward floating feature selection, Multidimen. Syst. Sig. Process. 28 (2017), 961–976.
[8] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2009 emotion challenge, Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009, IEEE, Brighton, UK, 2009, pp. 312–315.
[9] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, et al., The INTERSPEECH 2010 paralinguistic challenge, Proceedings of the 11th Annual Conference of the International Speech Communication Association 2010, IEEE, Makuhari, Japan, 2010, pp. 2794–2797.
[10] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird, et al., The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language, Proceedings of Interspeech 2016, ISCA, San Francisco, USA, 2016, pp. 2001–2005.
[11] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997), 1735–1780.
[12] W.J. Han, H.F. Li, A brief review on emotional speech databases, Intell. Comput. Appl. 3 (2013), 5–7.
[13] B. Athiwaratkun, J.W. Stokes, Malware classification with LSTM and GRU language models and a character-level CNN, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, New Orleans, LA, USA, 2017, pp. 2482–2486.
[14] S. Merity, N.S. Keskar, R. Socher, Regularizing and Optimizing LSTM Language Models, [2019-10-20].
[15] W. Li, B. Mak, Derivation of document vectors from adaptation of LSTM language model, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Valencia, Spain, 2017, 456–461.
[16] Y. Liu, D. Zhai, Q. Ren, News text classification based on CNLSTM model with attention mechanism, Comput. Eng. 45 (2019), 303–308, 314.
[17] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734.
[18] D. Britz, A. Goldie, M.T. Luong, Q. Le, Massive Exploration of Neural Machine Translation Architectures, [2019-10-20].
[19] X. Shi, Z. Chen, H. Wang, D.Y. Yeung, W.K. Wong, W.C. Woo, Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, [2019-10-20].
[20] S. Zhang, H. Jiang, S. Wei, L. Dai, Feedforward Sequential Memory Neural Networks without Recurrent Feedback, [2019-10-20].
[21] S. Zhang, M. Lei, Z. Yan, L. Dai, Deep-FSMN for Large Vocabulary Continuous Speech Recognition, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, AB, Canada, 2018, pp. 5869–5873.
[22] F. Tao, G. Liu, Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, AB, Canada, 2018, pp. 2906–2910.
[23] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, B. Schuller, Speech emotion classification using attention-based LSTM, IEEE/ACM Trans. Audio Speech Language Process. 27 (2019), 1675–1685.
[24] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, et al., Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Shanghai, China, 2016, pp. 5200–5204.
[25] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using

deep neural networks, IEEE J. Select. Topics Sig. Process. 11 (2017), 1301–1309.
[26] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, [2019-10-20].
[27] J. Chorowski, D. Bahdanau, K. Cho, Y. Bengio, End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results, [2019-10-20].
[28] T. Luong, H. Pham, C.D. Manning, Effective Approaches to Attention-based Neural Machine Translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1412–1421.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need, 31st Conference on Neural Information Processing Systems (NIPS 2017), IEEE, Long Beach, CA, USA, 2017, pp. 5998–6008.
[30] Z.H. Lin, M.W. Feng, C.N. dos Santos, M. Yu, B. Xiang, B. Zhou, et al., A Structured Self-attentive Sentence Embedding, [2019-10-20].
[31] Y. Xie, R. Liang, Z. Liang, L. Zhao, Attention-based dense LSTM for speech emotion recognition, IEICE Trans. Inform. Syst. E102.D (2019), 1426–1429.
[32] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, New Orleans, LA, USA, 2017, pp. 2227–2231.
[33] F.A. Gers, J. Schmidhuber, Recurrent nets that time and count, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, IEEE, Como, Italy, 2000, pp. 189–194.
[34] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network, [2019-10-20].
[35] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, et al., Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, 2018, pp. 2704–2713.
[36] S. Han, H.Z. Mao, W.J. Dally, Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, [2019-10-20].
[37] D. Dai, L. Yu, H. Wei, Parameters sharing in residual neural networks, Neural Process. Lett. 51 (2020), 1393–1410.
[38] S. Pan, J. Tao, Y. Li, The CASIA Audio Emotion Recognition Method for Audio/Visual Emotion Challenge 2011, in: S. D'Mello, A. Graesser, B. Schuller, J.C. Martin (Eds.), Affective Computing and Intelligent Interaction, ACII 2011, Springer, Berlin, Heidelberg, 2011, pp. 388–395.
[39] Jaitly N, Hinton G, Learning a better representation of speech soundwaves using restricted boltzmann machines, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Prague, Czech Republic, 2011, pp. 5884–5887.

AUTHORS INTRODUCTION

Mr. Tianye Jian

He has obtained a double bachelor degree in engineering and management in 2018. Now he is a master student in Tianjin University of Science and Technology. The major is control engineering.

Mr. Wanlong Peng

He is currently the master course student in Tianjin University of Science & Technology. Principle of automatic control, his research field is about intelligent robot. During his study, he has participated in and won awards as university robot competition.

Dr. Yizhun Peng

He is an Associate Professor in the School of Electronic Information and Automation, Tianjin University of Science and Technology. In 2002, he graduated from Nankai University with a master's degree in pattern recognition and intelligent systems. He graduated from the Chinese Academy of Sciences in 2006 with a PhD in Control Theory and Control Engineering.

Mr. Zhou Yang

He is a graduate student majoring in Instrument Science and Technology at Tianjin University of Science and Technology. His research direction is industrial robot.