Article

A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition

Huiyun Zhang [1,2], Heming Huang[1,2,*] and Henry Han [3,1]

[1] School of Computer Science, Qinghai Normal University, Xining, 810008, China;
[2] The State Key Laboratory of Tibetan Intelligent Information Processing and Application; Xining, 810008, China;
[3] Department of Computer and Information Science, Fordham University, New York/10023, USA

[*] Correspondence: E-mail: huanghm@qhnu.edu.cn; Tel.: +8613709727656

Abstract: Speech emotion recognition remains a heavy lifting in natural language processing. It has strict requirements to the effectiveness of feature extraction and that of acoustic model. With that in mind, a Heterogeneous Parallel Convolution Bi-LSTM model is proposed to address these challenges. It consists of two heterogeneous branches: the left one contains two dense layers and a Bi-LSTM layer, while the right one contains a dense layer, a convolution layer, and a Bi-LSTM layer. It can exploit the spatiotemporal information more effectively, and achieves 84.65%, 79.67%, and 56.50% unweighted average recall on the benchmark databases EMODB, CASIA, and SAVEE, respectively. Compared with the previous research results, the proposed model achieves better performance stably.

Keywords: Speech emotion recognition; Feature extraction; Heterogeneous parallel network; Spectral features; Prosodic features; Multi-feature fusion

1. Introduction

Emotion is a momentous element in human beings interactions, and speech contains a wealth of emotion information. People can perceive emotion from speech signals, and therefore they can capture emotional changes from speech. Speech Emotion Recognition (SER) aims to simulate emotion perception process of human beings to dig and decipher the emotional information contained in speech [1]. In the past decades, SER has attracted widespread concern of researchers, and many tremendous achievements have been made. For example, SER finds its applications in Human-Computer Interaction (HCI), robotics, mobile computing, and computer games [2-6]. With the fast development of Artificial Intelligence (AI), HCI becomes increasingly convenient and friendly by adding emotions to machines. To make human-computer interaction more harmonious, it is urgent to enable AI to recognize speech emotions so that machines or robots can act in a human-like manner. Hence, the SER research has strong academic and practical value.

2. Related Work

Generally, SER contains the undermentioned steps: corpus recording, signal preprocessing, emotion feature extraction, and classifier construction [7], etc. Among which emotion feature extraction is a principal step that extracts representative features for the downstream classification, and the classifier is the key part of a SER system that produces final SER results.

So far, a variety of Low-Level Descriptor (LLD) features have been used for SER [8], and MFCC is one of them [9]. There are other approaches (e.g., openSMILE) that extract the higher-level derivatives of the LLD features to seek deep feature extractions [9]. Besides, Chroma features can also represent emotion well. Compared to their peers, these features are more representative in capturing the affective information both from frequency and time domain for each frame in SER [7].

Traditionally, features are fed into acoustic models, and the recognition results are acquired through such machine learning based acoustic models as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), and so on [10-12]. These models usually achieve good performance on small-scale data rather than large-scale data.

With the development of deep learning technology, a variety of Artificial Neural Network (ANN) [13] is introduced to construct speech recognition classifiers. Compared with the early methods, when handling large-scale data, ANNs have better performances for their powerful capabilities in feature extraction and learning. Some representative deep acoustic models are proposed, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) [14-16].

The successful applications of the deep learning models have obtained exciting outcome in SER research, and it motivates us to develop more powerful models to recognize speech emotion. The recognition capability of a single network is usually limited. Therefore, the combination of different neural networks is suggested in quite a few previous works. Chen et al. proposed an ACRNN model that integrated CNN with LSTM, and 3D spectral features were used as the input of the acoustic model [17]. Trigeorgis et al. combined CNN with RNN, and they segmented the original audio data into equal-length speech fragments as the input of the classifier [18]. Sainath et al. proposed a CLDNN model consisted of a few convolution layers, LSTM layers, and fully connected layers in the respective order [19]. The CLDNN model, trained on the log-Mel filter bank energies [20] and on the raw waveform speech signals [21], outperformed both CNN and LSTM.

Inspired by the above research works, and to exploit the spatiotemporal information more effectively, a distinctive classifier, called Heterogeneous Parallel Convolution Bi-LSTM and abbreviated as HPCB hereafter, is proposed for SER. To exploit the spatiotemporal information more effectively, HPCB employs novel heterogeneous parallel learning structures. Furthermore, multi-features are used to dig and learn the complete emotional details in a more robust and effective way. HPCB demonstrates an advantage over the previous methods in literature on the benchmark databases EMODB, CASIS, and SAVEE [22-24].

The remainder of the study is arranged as follows. Section 2 describes the details of the proposed model HPCB. Section 3 presents the experimental and Section 4 concludes advantages of the proposed model and possible research directions.

3. Methods

Evolving from the preliminary models Bi-LSTM and CNN, the proposed model HPCB can process temporal coherence information in the spatial and time domains efficiently because of its well-designed heterogeneous parallel learning architecture that exploits the advantages of CNN and Bi-LSTM.

3.1. Heterogeneous Parallel Conv-BiLSTM

HPCB contains two heterogeneous branches, as shown in Figure 1. The purpose of designing the two heterogeneous branches is to project the original data into different transformation spaces for calculation, so as to better represent the original emotional speech.

The left one contains two dense layers and a Bi-LSTM layer, and it processes the temporal information of input data, the number of neurons in the two dense layers is 512, and the number of memory units in the Bi-LSTM layer is 256.

The right one contains a dense layer, a convolution layer, and a Bi-LSTM layer, and it handles the spatiotemporal information of input data. The number of neurons in the dense layer and one-dimensional convolution layer is 512, and the number of memory units in the Bi-LSTM layer is 256. 1D convolution is used to extract the spatial information of speech emotion signals in the time dimension, and Bi-LSTM is used to extract context information from the front and back ends of speech.

To represent emotional speech more completely, the features extracted from the left- and right-branches are fused through $Concatenate(\bullet)$ operation, where $Concatenate(\bullet)$ is the joint feature matrix. This operation increases the dimension of the features describing the original data, but the information corresponding to each dimension feature does not increase.

A $Softmax(\bullet)$ function is used to classify emotions according to emotional signals from concatenation layer that concatenate and fuse the information from the two heterogeneous branches. The number of neurons in $Softmax(\bullet)$ layer is equal to the number of emotion categories in the corresponding database.

The proposed parallel learning architecture speedups the convergence in deep learning, it also contributes to capture and retrieval spatiotemporal coherence information which plays an essential role in improving the learning performance of the model.

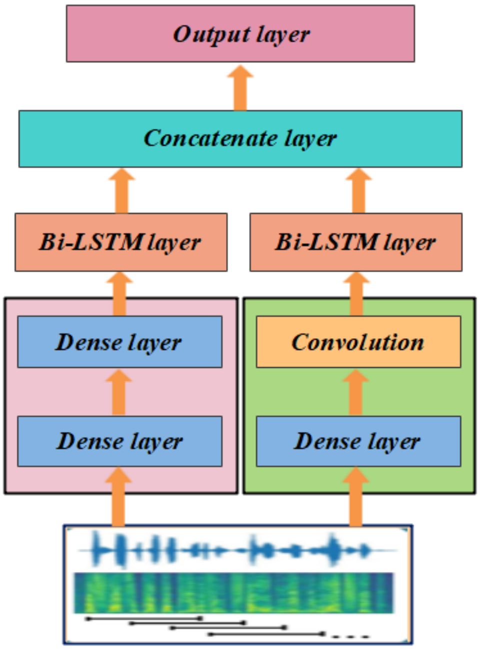


Figure 1. HPCB network topology

HPCB employs a valid convolution operation, and it performs convolution operation only for time dimensional tensor. This means that the convolution kernel moves inside the one-dimensional tensor. The output h of convolution is calculated as:

$$h = f\left(\frac{h^1 * F}{S} \times N\right). \tag{1}$$

Where h^1 denotes the output of dense layer, $F = [k_1, k_2, ..., k_{512}]$ denotes the convolution kernel, N denotes the number of filters and is set to 512. S denotes the stride and is set to 1 by default.

Bi-LSTM is adept at context modeling on time series data. Different from the traditional neural network, there is a connection between any two neurons in the same hidden layer. Bi-LSTM receives the input from the convolutional layer, and it helps the HPCB model to extract spatial and temporal time coherence emotion features more effectively.

The outputs y_L^B and y_R^B of the left and right branches are concatenated in the concatenate layer to merge information:

$$F_c = concatenate\left({y_L}^B ,\ {y_R}^B \right). \tag{2}$$

On the top of model HPCB, there is an output layer using $Softmax(\bullet)$ to classify emotion. It is noted that HPCB employs the Adam optimization in its learning procedure. Compared to the original Bi-LSTM or CNN, HPCB automatically extracts information both in the spatial and time domains in a parallel learning architecture by exploiting the pros of the two models.

4. Experiments

The proposed model HPCB outperforms its peers on three benchmark datasets described in subsection 4.1.

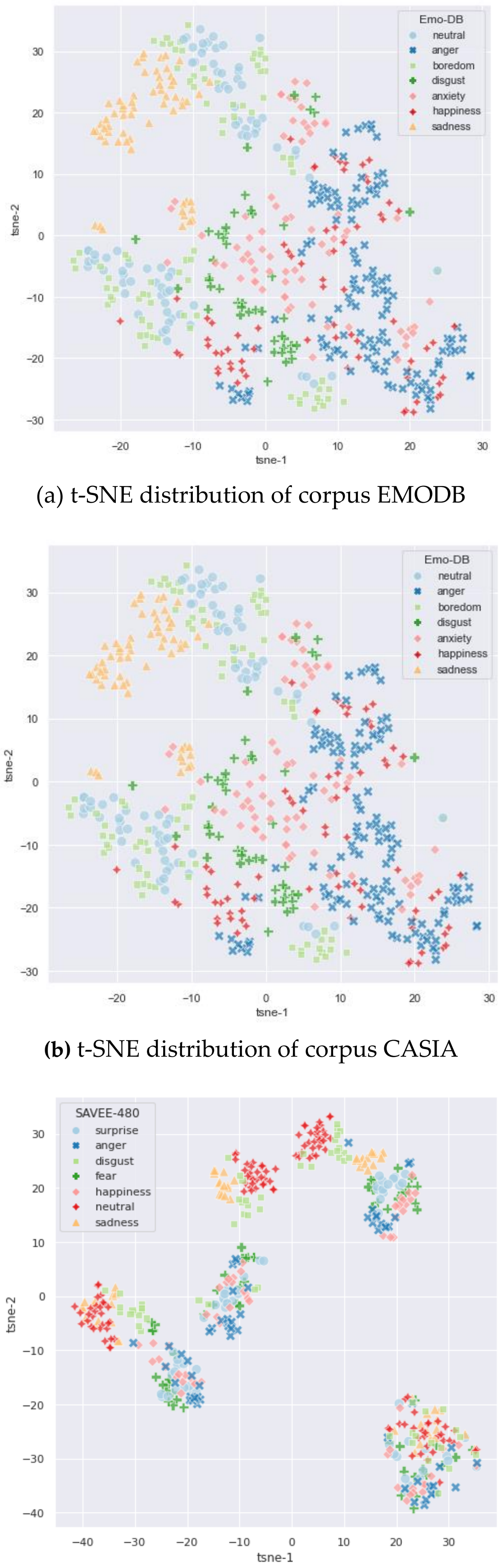
4.1. Databases

To query the effectiveness of HPCB in SER, it has been tested on three benchmark databases: EMO-DB [22], CASIA [23], and SAVEE [24]. EMO-DB is a German corpus and it contains 535 emotional sentences in total. It contains 10 speakers and 7 emotions, namely, boredom (B), anger (A), fear (F), sadness (S), disgust (D), happiness (H), and neutral (N). EMODB corpus

CASIA is a Chinese corpus, and it is constructed by the Institute of Automation, Chinese Academy of Sciences. The publicly CASIA corpus contains 1200 utterances and the average length of each audio is about 1.9s. There are 4 speakers and each speaker records 300 words in the same text. There are 6 emotions, namely, anger (A), fear (F), happy (H), neutral (N), sad (Sa), surprise (Su).

SAVEE is an English corpus [24], and it contains 4 speakers and 7 different emotions, namely, anger (A), disgust (D), fear (F), happiness (H), sadness (Sa), surprise (Su), and neutral (N). The number of samples of neutral is 120 while that of each remainder class is 60. Totally, there are 480 utterances.

Figure 2 shows t-SNE distribution of databases EMODB, CASIA, and SAVEE. It can be seen from Figure 2(a) that if the samples of database EMODB is projected into the two-dimensional space, the degree of confusion is large, and the data are inseparable. Figures 2(a) and 2(b) show us that the similar phenomena have also occurred to the samples of databases CASIA and SAVEE. This shows that the boundary of emotion classification is not clear, and it is difficult to identify.



(a) t-SNE distribution of corpus EMODB

(b) t-SNE distribution of corpus CASIA

(c) t-SNE distribution of corpus SAVEE

Figure 2. The t-SNE distribution of databases EMODB, CASIA, and SAVEE

4.2. Feature Extraction

Each speech is segmented into frames with a 25ms window and 10ms shifting step size. Each frame is Z-normalized. To each frame, 32D Low-Level Descriptor (LLD) features, including 12D Chroma [25] and 20D MFCC [26], are extracted. The High-Level Statistical Functions (HSF), such as the mean of Chroma and the mean, variance, and maximum of MFCC, are calculated. Totally, 72 D acoustic features are used as the input of the model.

4.3. Experimental Setup

All experiments are performed on a powerful PC with 64G RAM running under Windows 10. CPU speed is 2.10 GHz, core is 40, and logic processor is 80. 2 RTX 2080 Ti GPUs are used to accelerate computing. All models are implemented with TensorFlow toolkit [27].

To prevent possible overfitting, during the training stage, dropout is implemented in all layers. Dropout rate is 0.5, batch size is 32, and epoch is 100. In addition, Adam [28] is adopted as optimizer.

The datasets EMODB, CASIA, and SAVEE do not provide a separate training and testing set, therefore, speaker-independent (SI) strategy is employed to do train-test partition. The samples of each dataset are randomly divided into 5 equal parts, and 4 parts are used as the training data while the remaining one is used as the testing set. Experiments are repeated 10 times and the average value of all trials are computed. Confusion matrix and such evaluation measures as precision, unweighted average recall (UAR), accuracy, and F1-score are employed to evaluate the performance.

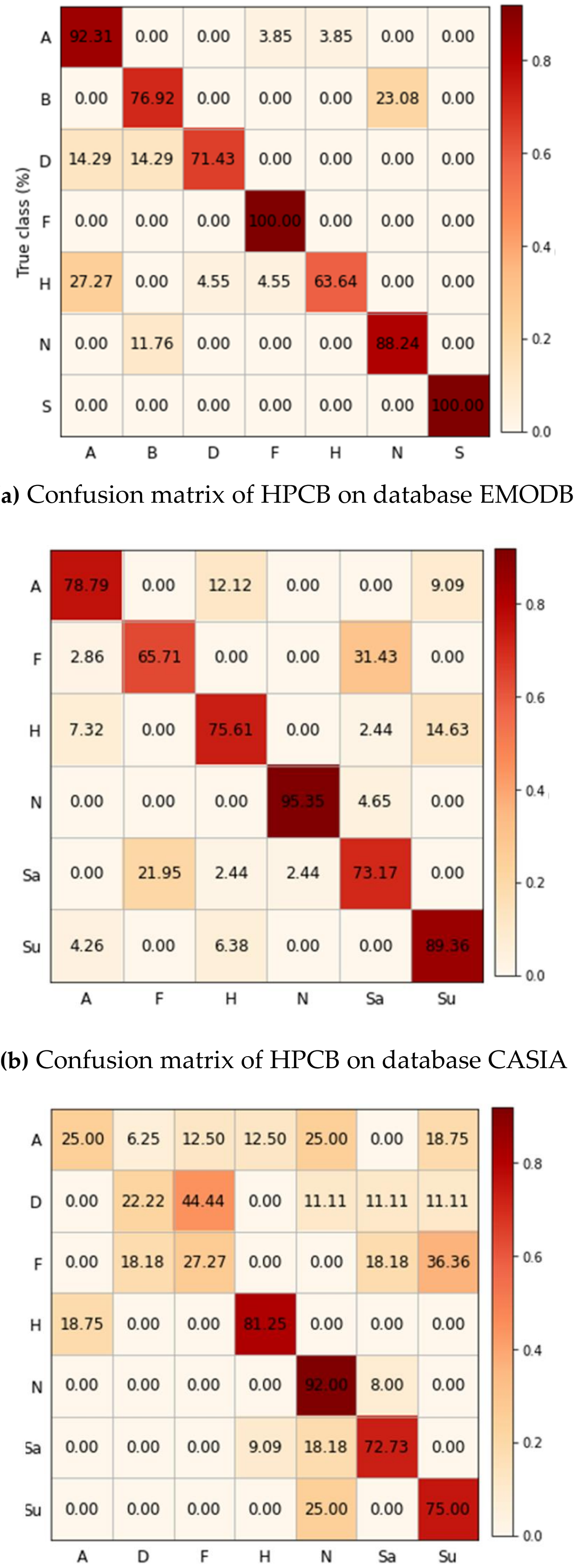
4.4. The Performance of HPCB and its peer methods

To analyze generalization ability, on the datasets EMODB, CASIA, and SAVEE, confusion matrices of the model HPCB are obtained by averaging 10 experimental results, as shown in Figures 3. The diagonal entry of each confusion matrix represents the recall rate. The prediction results of the three confusion matrices are summarized as follows.

First, on the test sets of databases EMODB, CASIA, and SAVEE, the average UARs of the model HPCB are 84.65%, 79.67%, and 56.50% respectively. Obviously, it achieves the best performance on the EMODB database.

Second, on the test set of the EMODB database, emotions Fear (F) and sadness (S) achieve 100.00% UAR, which is a very impressive recognition result because it has rarely achieved in the previous literature. Similarly, emotions Neutral (N) and Surprise (Su) achieve 95.35% and 89.36% UAR on the test set of the CASIA database, emotions Happiness (H) and Neutral (N) achieve 81.25% and 92.00% UAR on the test set of the SAVEE database.

Third, on the test set of the EMODB database, it is noted that the emotions Boredom (B) and Neutral (N) are easily confusing pairs, so do Happiness (H) and Anger (A). On the test set of the CASIA database, it is noted that the emotions Fear (F) and Sadness (Sa) are easily confusing pairs. On the test set of the SAVEE database, emotions Anger (A), Disgust (D), and Fear (F) have a low-level recognition performance.



	A	B	D	F	H	N	S
A	92.31	0.00	0.00	3.85	3.85	0.00	0.00
B	0.00	76.92	0.00	0.00	0.00	23.08	0.00
D	14.29	14.29	71.43	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	100.00	0.00	0.00	0.00
H	27.27	0.00	4.55	4.55	63.64	0.00	0.00
N	0.00	11.76	0.00	0.00	0.00	88.24	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(a) Confusion matrix of HPCB on database EMODB

	A	F	H	N	Sa	Su
A	78.79	0.00	12.12	0.00	0.00	9.09
F	2.86	65.71	0.00	0.00	31.43	0.00
H	7.32	0.00	75.61	0.00	2.44	14.63
N	0.00	0.00	0.00	95.35	4.65	0.00
Sa	0.00	21.95	2.44	2.44	73.17	0.00
Su	4.26	0.00	6.38	0.00	0.00	89.36

(b) Confusion matrix of HPCB on database CASIA

	A	D	F	H	N	Sa	Su
A	25.00	6.25	12.50	12.50	25.00	0.00	18.75
D	0.00	22.22	44.44	0.00	11.11	11.11	11.11
F	0.00	18.18	27.27	0.00	0.00	18.18	36.36
H	18.75	0.00	0.00	81.25	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	92.00	8.00	0.00
Sa	0.00	0.00	0.00	9.09	18.18	72.73	0.00
Su	0.00	0.00	0.00	0.00	25.00	0.00	75.00

(c) Confusion matrix of HPCB on database SAVEE

Figure 3. The confusion matrices of HPCB on the datasets EMODB, CASIA, and SAVEE

Table 1 to 3 summarizes the improvement of the performances of HPCB in terms of UAR to the related peer methods on databases CASIA, EMODB, and SAVEE. Among them, literatures [7], [29-30], [32] used the research results of previous researchers as the baseline, while literature [21] was originally proposed in the research of automatic speech recognition. When researchers in literature [34-37] applied it to speech emotion recognition, the database used was also inconsistent with the database used in this study. Therefore, this study adopts the model structure proposed in the literature [21], [34-37], and verifies the model performance on the four databases used in this study. Final results are shown in Table 1-Table 3.

Table 1. Performance comparisons (%) of the model HPCB to those of the peers in literature on CASIA database

Model	WAR	UAR
GA-BEL [29]	38.55	38.55
HuWSF [30]	43.50	43.50
RDBN [32]	48.50	48.50
PCRN [7]	58.25	58.25
Bi-LSTM [34]	/	75.00
Bi-GRU [35]	/	72.50
CNN [36]	/	76.67
CLDNN [21]	/	61.67
CapsNet [37]	/	63.33
HPCB(Ours)	/	79.67

Table 2. Performance comparisons (%) of the model HPCB to those of the peers in literature on EMODB database

Model	WAR	UAR
HuWSF [30]	81.74	/
RDBN [31]	82.32	/
LNCMSF [32]	/	74.46
ACRNN [33]	/	82.82
PCRN [7]	86.44	84.53
Bi-LSTM [34]	/	71.03
Bi-GRU [35]	/	70.09
CNN [36]	/	78.50
CLDNN [21]	/	56.07
CapsNet [37]	/	77.57
HPCB(Ours)	/	84.65

Table 3. Performance comparisons (%) of the model HPCB to those of the peers in literature on SAVEE database

Model	WAR	UAR
GA-BEL [29]	44.18	/
HuWSF [30]	50.00	/
RDBN [31]	53.60	/
PCRN [7]	62.49	59.40
Bi-LSTM [34]	/	44.79
Bi-GRU [35]	/	41.67
CNN [36]	/	54.17
CLDNN [21]	/	43.75
CapsNet [37]	/	56.25
HPCB (Ours)	/	56.50

The databases CASIA and EMODB, performances of model HPCB are better than the previous model structures. Among them, the model HPCB has achieved 79.67% and 84.65% recognition performance on databases CASIA and EMODB, respectively. On database SAVEE, the model HPCB has achieved 56.50% recognition performance, the UAR of HPCB on the SAVEE database is only 2.90% lower than that in the literature [7]. This suggests that the proposed model has good robustness and generalization.

5. Conclusion and discussion

In this study, a novel heterogeneous parallel acoustic model called HPCB is proposed for speech emotion recognition. It exploits the spatiotemporal information more effectively. It is characterized by its two heterogeneous branch structures: the left one is composed of two dense layers and a Bi-LSTM layer, while the right one is composed of a dense layer, a convolution layer, and a Bi-LSTM layer. The 72D high-level statistical functions (HSF) features are calculated to verify the robustness and generalization of the model HPCB. Experimental results on the databases EMO-DB and CASIA suggest that HPCB demonstrate stable advantages over the previous methods in the literature.

In the future, the effectiveness of HPCB will be further verified by applying it to other emotion databases and analyzing possible overfitting risks., HPCB can also be extended to other audio recognition or image classification problems for its superior learning capabilities. Furthermore, it will be compared with other deep learning models such as Generative adversarial network (GAN) in SER besides zero-shot learning techniques.

The proposed model demonstrates an impressive performance in SER by retrieving spatiotemporal information in deep learning. It suggests that the spatiotemporal signals extraction can be essential to achieve high-performance SER. On the other hand, how to decrease possible overfitting risk can be another interesting topic to explore further. This is because the proposed HPBC may face possible overfitting though the 0.5 dropout is employed in learning for its complicate learning architecture. We plan to evaluate the learning performance in comparison with other peer deep learning models to query whether the integration of different types of neural networks will lead to an increase of overfitting and how should we overcome it efficiently if it were to happen.

Acknowledgments: The authors acknowledge the National Natural Science Foundation of China (Grant: 62066039). This study is supported by the Key Laboratory of Tibetan Information Processing, Ministry of Education, Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province (Grant: 2020-ZJ-Y05), Tibetan Information Engineering Technology Research Center of Qinghai Province.

References

1. Tahon M, Devillers L. Towards a small set of robust acoustic features for emotion recognition: challenges. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(1), 16-28(2016).
2. Xu H H, Gao J, Yuan J. Application of speech emotion recognition in intelligent household robot// Proceedings of International Conference on Artificial Intelligence and Computational Intelligence (AICI). Sanya, China, 537-541(2010).
3. Schuller B W. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61(5), 90-99(2018).
4. Perez-Liebana D, Liu J L, Khalifa A, et al. General Video Game AI: A Multitrack Framework for Evaluating Agents, Games, and Content Generation Algorithms. IEEE Transactions on Games, 11(3), 195-214(2019).
5. Khaki H, Erzin E. Use of affect based interaction classification for continuous emotion tracking// Proceedings of International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA, 2881-2885(2017).
6. Tzirakis P, Zhang J H, Schuller B W. End-to-End Speech Emotion Recognition Using Deep Neural Networks// Proceedings of International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada, 5089-5093(2018).
7. Jiang P X, Fu H L, Tao H W, et al. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. IEEE Access, 9(1), 90368-90376(2019).
8. Xia X H, Liu J M, Yang T, et al. Video Emotion Recognition using Hand-Crafted and Deep Learning Features// Proceedings of Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). Beijing, China, 1-6(2018).

9. Koo H, Jeong S, Yoon S, et al. Development of Speech Emotion Recognition Algorithm using MFCC and Prosody// Proceedings of International Conference on Electronics, Information, and Communication. Barcelona, Spain, 1-4(2020).
10. Nwe T L, Foo S W, Silva L. Speech emotion recognition using hidden Markov models. Speech Communication, 41(4), 603-623(2003).
11. Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs// Proceedings of INTERSPEECH. Pittsburgh, USA, 809-812(2006).
12. Amol T K, Guddeti. Multiclass SVM-based language independent emotion recognition using selective speech features// Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI). New Delhi, India, 1069-1073(2014).
13. Fu L Q, Mao X, Chen L J. Relative Speech Emotion Recognition Based Artificial Neural Network// Proceedings of Pacific Asia Conference on Language, Information and Computing (PACLIC), Wuhan, China, 140-144(2008).
14. Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention// Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA, 2227-2231(2017).
15. Xie Y, Liang R Y, Liang Z L, et al. Speech Emotion Classification Using Attention-Based LSTM. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 27(11), 1675-1685(2019).
16. Mao Q, Dong M, Huang Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 16(8), 2203-2213(2014).
17. Chen M, He X, Yang J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 25(10), 1440-1444(2018).
18. Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network// Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, 5200-5204(2016).
19. Zhao Z P, Zheng Y, Zhang Z X, et al. Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition// Proceedings of INTERSPEECH. Hyderabad, India, 272-276(2018).
20. Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia, 4580-4584(2015).
21. Sainath T N, Weiss R J, Senior A, et al. Learning the speech front-end with raw waveform CLDNNs// Proceedings of International Speech Communication Association. Dresden, Germany, 1-5(2015).
22. Cirakman O, Gunsel1 B. Online Speaker Emotion Tracking with a Dynamic State Transition Model// Proceedings of International Conference on Pattern Recognition (ICPR). Cancún, México, 307-312(2016).
23. Wang K X, An N, Li B N, et al. Speech Emotion Recognition Using Fourier Parameters. IEEE Transactions on Affective Computing, 25(1), 69-75(2015).
24. Kim Y, Provost E M. ISLA: Temporal Segmentation and Labeling for Audio-Visual Emotion Recognition. IEEE Transactions on Affective Computing, 10(2), 196-208(2019).
25. Grag U, Agarwal S, Gupta S, et al. Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma// Proceedings of International Conference on Computational Intelligence and Communication Networks (CICN). Bhimtal, India, 1-5(2020).
26. Kumbhar H S, Bhandari S U. Speech Emotion Recognition using MFCC features and LSTM network// Proceedings of International Conference on Computing, Communication, Control and Automation (ICCUBEA). Pune, India, 1-3(2019).
27. Suen H Y, Hung K E, Lin C L. TensorFlow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews. IEEE Access, 7(1): 61018-61023(2019).
28. Zou F Y, Shen L, Jie Z Q, et al. A Sufficient Condition for Convergences of Adam and RMSProp// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 11119-11127(2019).
29. Liu Z T, Xie Q, Wu M, et al. Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing, 309(1), 145-156(2018).
30. Sun Y, Wen G, Wang J. Weighted spectral features based on local Hu moments for speech emotion recognition. Biomedical Signal Processing and Control, 18(1): 80-90(2015).
31. Wen G, Li H, Huang J, et al. Random deep belief networks for recognizing emotions from speech signals. Computational Intelligence and Neuroscience, 5(4): 1-9(2017).
32. Tao H, Liang R, Zha C, et al. Spectral features based on local Hu moments of Gabor spectrograms for speech emotion recognition. IEICE Transactions on Information and Systems, E99. D(8): 2186-2189(2016).
33. Chen M, He X, Yang J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 25(10), 1440-1444(2018).
34. Dai T, Zhu L, Wang Y X, et al. Attentive Stacked Denoising Autoencoder With Bi-LSTM for Personalized Context-Aware Citation Recommendation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28(1): 553-568(2020).
35. Wang H, Zhao D Q. Emotion analysis of microblog based on emotion dictionary and Bi-GRU// Proceedings of Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). Dalian, China, 197-200(2020).
36. Lee K H, Kim D H. Design of a Convolutional Neural Network for Speech Emotion Recognition // Proceedings of International Conference on Information and Communication Technology Convergence (ICTC). Jeju, Korea, 1332-1335(2020).

doi:10.20944/preprints202108.0433.v1

37. Sara S, Nicholas F, Geoffrey E H. Dynamic Routing Between Capsules// Proceedings of Neural Information Processing Systems (NIPS), Long Beach, USA, 1-11(2017).