applied sciences

MDPI

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] School of Computer Science, Qinghai Normal University, Xining 810008, China; zhy1406043513@163.com
[2] The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810008, China
[3] Department of Computer Science, School of Engineering and Computer Science, Baylor University, One Bear Place #97141, Waco, TX 76798, USA; xhan9@fordham.edu
* Correspondence: huanghm@qhnu.edu.cn; Tel.: +86-13709727656

Abstract: Speech emotion recognition is a substantial component of natural language processing (NLP). It has strict requirements for the effectiveness of feature extraction and that of the acoustic model. With that in mind, a Heterogeneous Parallel Convolution Bi-LSTM model is proposed to address the 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 recalls 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

check for updates

Citation: Zhang, H.; Huang, H.; Han, H. A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition. *Appl. Sci.* **2021**, *11*, 9897. https://doi.org/10.3390/app11219897

Academic Editor: Doroteo Torre Toledano

Received: 15 August 2021
Accepted: 11 October 2021
Published: 22 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, rapid progress of technology makes smart devices more attractive in our daily life. Intelligent services such as chatbots, psychological diagnosis assistants, intelligent healthcare, sales advertising and intelligent entertainment consider not only the completion of services, but also the humanization of communication between human and computer [1]. How to implement an intelligent human–machine interface becomes an important issue. For the applications of the spoken dialog systems, the leading organizations use chatbots to improve their customer service and generate good business results for the organization [2]. In contrast to customer engagement, empathy, which is highly related to emotion, has been incorporated into the design of a dialogue system for improving user experience in human–computer interaction (HCI). More importantly, being empathetic is a necessary step for the dialogue system to be perceived as a social character by users. The realization of humanized HCI based on the above emotion motivation will be research of far-reaching significance.

Emotion plays an important role in perception, attention, memory, and decision-making processing of human beings, and human speech contains a wealth of emotional information [3]. People can perceive emotion from different speech signals, and therefore they can capture emotional changes from speech. As a vital process for human-to-human communication, speech emotion recognition is automatically and subconsciously performed by humans [4]. Thus, to achieve better HCI, speech emotion recognition must be handled smoothly so that machines can detect emotional information from human speech in real time.

Speech Emotion Recognition (SER) aims to simulate the emotional perception process of human beings finding and deciphering the emotional information contained in speech [5]. In the past decades, SER has attracted the widespread concern from researchers, and many

tremendous achievements have been made. For example, Reeves et al. found that people tend to treat computers as if they are intelligent and emotion-aware [6]. This demonstrates a growing need for agents with proper affective behavior and affective understanding in areas such as interactive robots and story-telling agents [7,8]. With the fast development of Artificial Intelligence (AI), HCI becomes increasingly convenient and friendly by adding emotions to machines. To make HCI more harmonious and intelligent, 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

As the fundamental research of affective computing that integrates emotion with AI, SER has become an active research area for its wide applications in other fields that range from NLP, HCI, psychology, and deep learning. Generally, SER contains the undermentioned steps: corpus recording, signal preprocessing, emotion feature extraction, and classifier construction [9], etc. The emotion feature extraction is a principal step that extracts representative features from input data for the sake of downstream classification. As the key component of an SER system that produces final SER results, the classifier can be learning machines from ensemble learning, kernel-based learning, deep learning, or their integration [7,8].

Conventionally, speech emotion recognition systems are trained with supervised learning models or their variants [7,8]. The generalization of the models is often emphasized by training on a variety of samples with diverse labels [10]. Generally, labels for emotion recognition tasks are collected with perceptual evaluations from multiple evaluators. The raters annotate samples by listening to or watching the stimulus. This evaluation procedure is cognitively intense and expensive [11]. Therefore, standard benchmark SER datasets usually have a limited number of sentences with emotional labels, often collected from a limited number of evaluators.

2.1. Emotion Feature Extraction

The feature extraction in SER can be challenging because of the high nonlinearity of input speech data that contains noise from various sources ranging from data collection to data preprocessing. The two most challenging problems in SER feature extraction are probably the extraction of frame-based high-level feature representations and the construction of utterance-level features [12,13]. Since speech signals are considered to be approximately stationary in small frames, those acoustic features extracted from short frames, which include pitch, mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), and prosodic features, are believed to be influenced by emotions. The frame-based features, often referred to as low-level features, can provide detailed emotionally relevant local information [14]. So far, a variety of Low-Level Descriptor (LLD) features have been employed for SER. For example, Schmitt et al. proposed a bag-of-audio-words (BoAW) approach, which was created from MFCCs and energy Low Level Descriptors (LLDs), to build the feature vectors in SER and employed support vector regressions (SVR) to predict the emotion [15].

Furthermore, based on the frame-based low-level features, neural networks are utilized to extract neural representations frame by frame, which are referred to as frame-based high-level feature representations [16]. Recently, SER has made great progress by introducing neural networks to extract high-level neural hidden representations. For example, the authors in [16,17] applied neural networks on the low-level features, e.g., pitch, energy, to learn high-level features, i.e., the neural network outputs. Trigeorgis et al. proposed an end-to-end model composed of a convolution neural network (CNN) architecture used to extract features before feeding a bi-directional long short-term memory (BLSTM) to model the temporal dynamics in the data [18]. Neumann et al. proposed an attentive convolution neural network (ACNN) that combines CNNs with the attention mechanism [19].

However, the emotion recognition at utterance-level requires a global feature representation, which contains both detailed local information and global characteristics related to emotion. Based on the frame-based high-level features learned, various methods are used to construct the utterance-level features. Han et al. proposed to use extreme learning machines (ELM) upon the utterance-level statistical features [17]. The utterance-level features are the statistics of segment-level deep neural network (DNN) output probabilities, where each segment is a stack of neighboring frames. Satt et al. introduced recurrent neural networks to increase the ability of the model to capture temporal information [20]. However, since only the final states of recurrent layers were used for classification, it would lead to the loss of detailed information in the following emotion classification, because all information is stored in the fixed-size final states. Li et al. explored pooling utterance-level features from the high-level features output by CNNs with attention weights, because not all regions in the spectrogram contain information useful for emotion recognition [21], in which using pooling can also avoid squeezing all of the information into one fixed-size vector.

In this study, we propose a new multi-feature fusion representation method for discrete emotion recognition. Unlike most of the previous studies in the literature, the method of our feature fusion was inspired by the way that conventional speech features (e.g., Mel-Frequency Cepstral Coefficients (MFCCs)) are computed. That is, 32D Low-Level Descriptor (LLD) features, including 12D Chroma [22] and 20D MFCC [23], are extracted. The High-Level Statistical Functions (HSF), such as the mean of Chroma and the mean, variance, and maximum of MFCC, are calculated accordingly. Totally, 72 D acoustic features are used as the input for the following machine learning model for speech emotion recognition. Compared to their peers, these features are more representative in capturing the affective information both from frequency and time domains for each frame in SER [9]. Finally, our SER model outperforms the state-of-the-art studies for the benchmark databases EMODB [24], CASIA [25], and SAVEE [26]. We review state-of-the-art speech emotion models in the following section before introducing more details of our work.

2.2. Speech Emotion Recognition Models

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 [27–29]. These models usually achieve good performance on small-scale data rather than large-scale data.

In recent years, with the development of deep learning technology, a variety of Artificial Neural Networks (ANN) [30] are introduced to construct SER classifiers. A number of studies in the literature have focused on predicting emotions from speech using deep neural networks (DNN) For example, Xu et al. [31] aim to perform categorized recognition of five speech emotions as represented by joy, grief, anger, fear and surprise by means of algorithm with the combination of HMM and SOFMNN models so as to apply speech emotion recognition methods presented by integrated HMM/SOFMNN model to the platform of intelligent household robots. Stuhlsatz et al. [32] employed Restricted Boltzmann Machines (RBM) to extract discriminative features from the raw signal and developed a Generalized Discriminant Analysis (GerDA). Sainath et al. [4,33] proposed a convolutional long short-term memory deep neural network (CLDNN) model able to reduce temporal and frequency variations in speech emotion recognition.

Compared with the early methods, ANNs have better performance in handling large-scale data for their built-in 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) [14,34,35]. Among the mentioned state-of-the-art deep learning models, the CNN-based models show more powerful performance in representation learning for their advantages in spatial information extraction and learning. The structure of CNN usually

consists of pooling layers and convolutional layers, in which, the max-pooling layers can drop 'not the most important information' in learning and benefit the speech emotion recognition procedures. The convolutional layer efficiently handles the information in the receptive field and extracts features in the local regions. However, CNN may ignore time information in the emotional speech. Bi-LSTM is introduced to process time-series information to overcome this weakness of CNN [18].

The successful applications of the deep learning models with exciting results in SER motivate researchers to develop more powerful and efficient models. Since the recognition capability of a single network is usually limited, the combination of different neural networks is suggested in many works. Chen et al. proposed an ACRNN model that integrated CNN with LSTM, in which the 3D spectral features were used as the input of the acoustic model [36]. Trigeorgis et al. combined CNN with RNN besides segmenting the original audio data into equal-length speech fragments as classifier input [18]. Sainath et al. proposed a CLDNN model consisting of a few convolution layers, LSTM layers, and fully connected layers in the respective order [37]. The CLDNN model, trained on the log-Mel filter bank energies and on the waveform speech signals, outperformed both CNN and LSTM [38].

2.3. Our Contributions

Inspired by the above research works, we proposed a novel deep learning model called Heterogeneous Parallel Convolution Bi-LSTM (HPCB) to exploit spatiotemporal information more effectively in SER. HPCB employs novel heterogeneous parallel learning structures to exploit the spatiotemporal information. Furthermore, multi-features are used to unveil the complete emotional details in a more robust and effective way. Moreover, HPCB demonstrates an advantage over the previous methods in literature on the benchmark databases that include EMODB, CASIS, and SAVEE.

The model with Heterogeneous Parallel achieves improvements over the baselines in most cases. We make the following contributions in this study:

(1) 32D LLD features at the frame-level are extracted, and its HSF features at the sentence-level are computed. Totally, 72D acoustic features are used as the input of the model.
(2) The proposed HPCB architecture can be trained with features extracted at the sentence-level (high-level descriptors), or at the frame-level (low-level descriptors) simultaneously.
(3) We provide a comprehensive analysis of Heterogeneous Parallel network training for the sake of SER, demonstrate its advantages within corpus evaluations, and observe performance gains.

The rest of the paper is organized as follows. Section 3 presents the proposed Heterogeneous Parallel architectures. Section 4 gives details on the experimental setup including the databases and features used in this study. Section 5 presents the exhaustive experimental evaluations, showing the benefits of the proposed architecture. Finally, Section 6 provides the concluding remarks, discussing potential areas of improvements.

3. Methods

Evolving from the preliminary models Bi-LSTM and CNN [9,18], the proposed deep learning 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.

Heterogeneous Parallel Conv-BiLSTM (HPCB)

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.

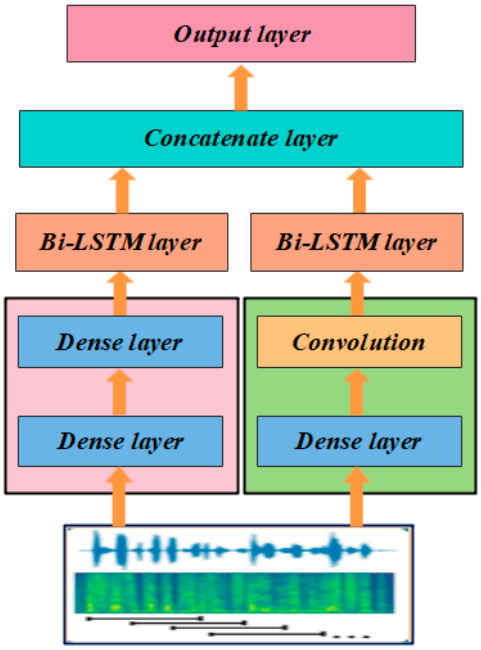


Figure 1. The topology of the proposed deep learning model HPCB.

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 cells 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 cells 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

$Concatenate(\cdot)$ operation, where $Concatenate(\cdot)$ 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 $Softmax(\cdot)$ function is used to classify emotions according to emotional signals from the concatenation layer that concatenates and fuses the information from the two heterogeneous branches. The number of neurons in the $Softmax(\cdot)$ layer is equal to the number of emotion categories in the corresponding database.

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

The proposed 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:

$h = f(h)$

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

where h^1 denotes the output of dense layer, $F = [k_1, k_2, \ldots, k_{512}]$ 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 convolution layer, and it helps the HPCB model to extract spatial and temporal 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:

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

On the top of model HPCB, there is an output layer using $Softmax(\cdot)$ 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. Experimental Evaluations

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

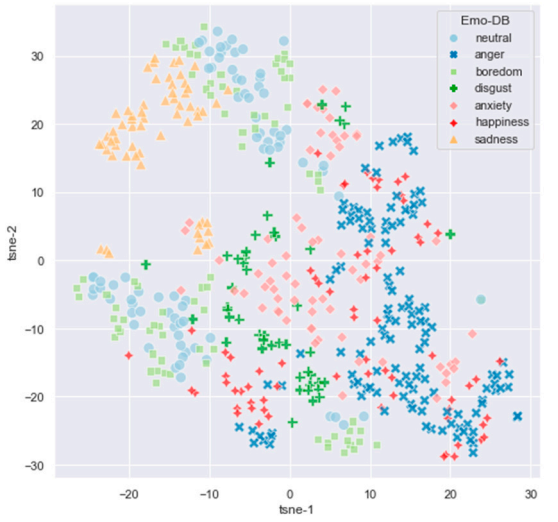
4.1. Databases

To query the effectiveness of HPCB in SER, it was tested on three benchmark databases EMO-DB [24], CASIA [25], and SAVEE [26]. 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).

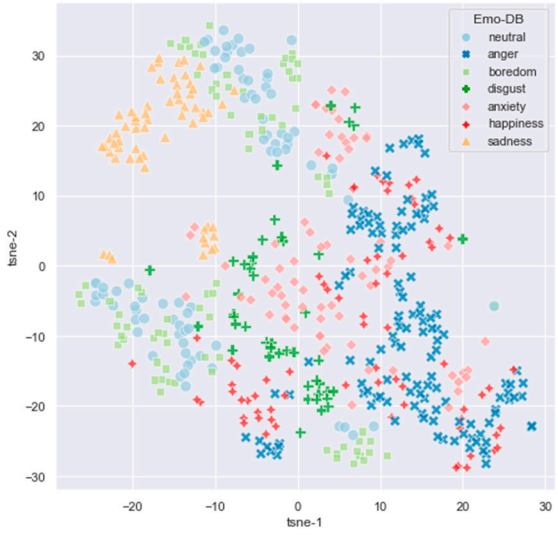
CASIA is a Chinese corpus, and it is constructed by the Institute of Automation, Chinese Academy of Sciences. The public CASIA corpus contains 1200 utterances and the average length of each audio is about 1.9 s. There are 4 speakers and each speaker records 300 words in the same text. There are 6 emotions,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

SAVEE is an English corpus, 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.

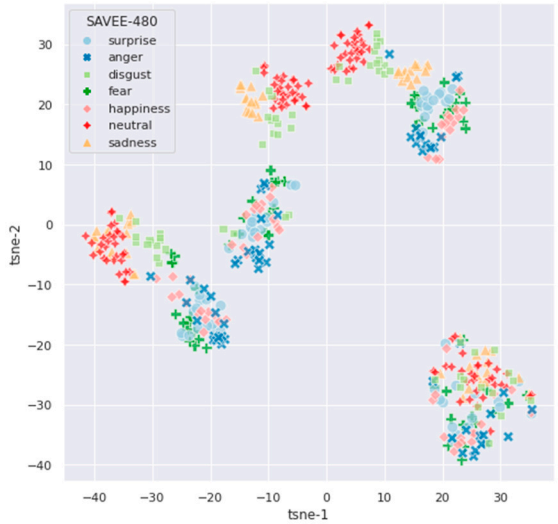
Figure 2 shows t-SNE visualizations of the databases EMODB, CASIA, and SAVEE. It can be seen from Figure 2a that even if the samples of database EMODB are projected into the two-dimensional space spanned by t-SNE bases, the degree of confusion can be still large. The t-SNE visualization implies that the dataset shows high nonlinearity and the corresponding SER classification seems to be a nonlinear inseparable problem. Figure 2b,c shows the similar situations for the samples of databases CASIA and SAVEE. They suggest that they both are nonlinear inseparable problems because there are no clear boundaries for each class and different types of samples are even wired together sometimes.



(**a**)



(**b**)



(**c**)

Figure 2. The t-SNE visualizations of the benchmark databases: EMODB, CASIA, and SAVEE. (**a**) The t-SNE visualization of corpus EMODB. (**b**) The t-SNE visualization of corpus CASIA. (**c**) The t-SNE visualization of corpus SAVEE.

4.2. Feature Extraction

We conducted the following feature extraction for the three databases in this study. Each speech was segmented into frames with a 25 ms window and 10 ms shifting step size. Each frame was Z-normalized. To each frame, 32D Low-Level Descriptor (LLD) features, including 12D Chroma [23] and 20D MFCC [24], were extracted. The High-Level Statistical Functions (HSF), such as the mean of Chroma and the mean, variance, and maximum of MFCC, were calculated. Totally, 72 D acoustic features were used as the input of the model.

4.3. Experimental Setup

All experiments were performed on a powerful PC with 64G RAM running under Windows 10. CPU speed was 2.10 GHz, core was 40, and logic processor was 80. To accelerate computing, 2 RTX 2080 Ti GPUs were used. All models were implemented with TensorFlow toolkit [39].

To prevent possible overfitting, during the training stage, dropout was implemented in all layers. Dropout rate was 0.5, batch size was 32, and epoch was 100. In addition, Adam [40] was adopted as optimizer.

The datasets EMODB, CASIA, and SAVEE do not provide a separate training and testing set, therefore, speaker-independent (SI) strategy was employed to do train–test partition. The samples of each database were randomly divided into 5 equal parts, and 4 parts were used as the training data while the remaining one was used as the testing set. Experiments were repeated 10 times and the average value of all trials was computed. Confusion matrix and such evaluation measures as precision, unweighted average recall (UAR), accuracy, and F1-score were 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 were obtained by averaging 10 experimental results, as shown in Figure 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.

Tables 1–3 summarize the performance improvements of HPCB in terms of UAR with respect to the related peer methods on the databases CASIA, EMODB, and SAVEE. Among them, the authors of [9,41–44] used the research results of previous researchers as the baseline, while the study of [45] was originally proposed in the research of automatic speech recognition. When researchers in [46–49] applied it to speech emotion recognition, the database used was also inconsistent with the database used in this study. Therefore, this study adopted the model structure proposed in [45–49] and verified the model performance on the three databases used in this study. Final results are shown in Tables 1–3.

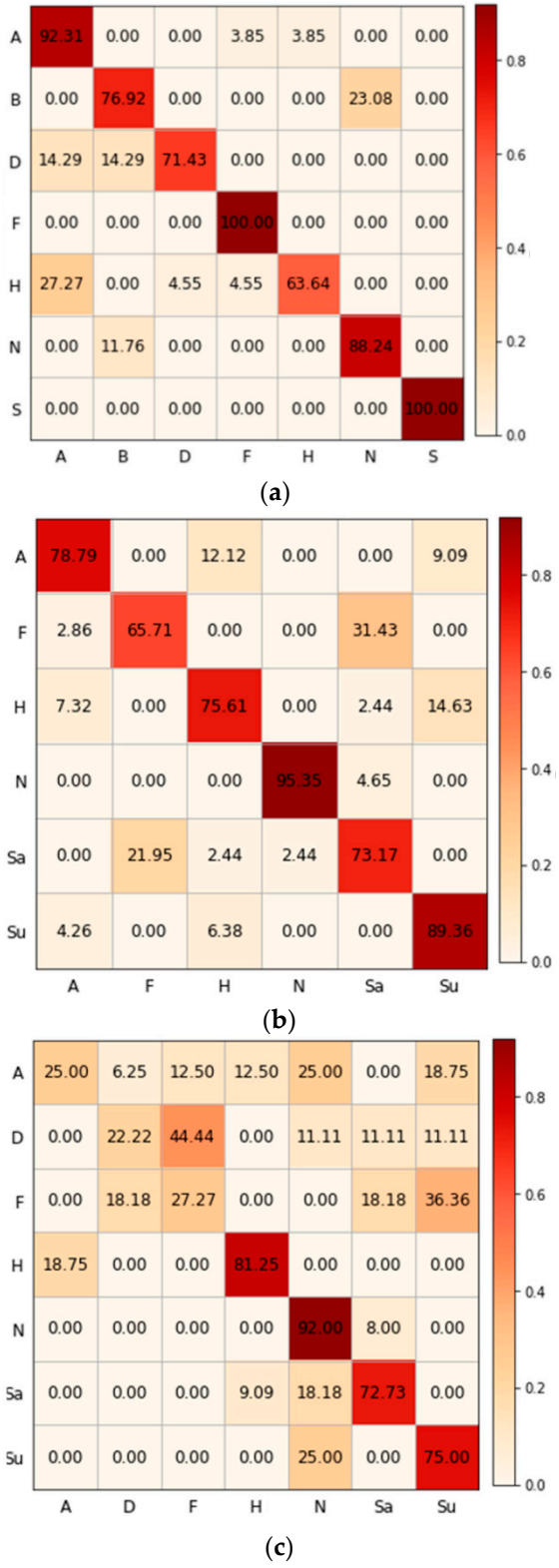


Figure 3. The confusion matrices of HPCB on the datasets EMODB, CASIA, and SAVEE. (**a**) Confusion matrix of HPCB on database EMODB. (**b**) Confusion matrix of HPCB on database CASIA. (**c**) Confusion matrix of HPCB on database SAVEE.

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

Model	WAR	UAR
GA-BEL [41]	38.55	38.55
HuWSF [42]	43.50	43.50
RDBN [44]	48.50	48.50
PCRN [9]	58.25	58.25
Bi-LSTM [46]	/	75.00
Bi-GRU [47]	/	72.50
CNN [48]	/	76.67
CLDNN [45]	/	61.67
CapsNet [49]	/	63.33
HPCB (Ours)	/	79.67

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

Model	WAR	UAR
HuWSF [42]	81.74	/
RDBN [43]	82.32	/
LNCMSF [44]	/	74.46
ACRNN [36]	/	82.82
PCRN [9]	86.44	84.53
Bi-LSTM [46]	/	71.03
Bi-GRU [47]	/	70.09
CNN [48]	/	78.50
CLDNN [45]	/	56.07
CapsNet [49]	/	77.57
HPCB (Ours)	/	84.65

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

Model	WAR	UAR
GA-BEL [41]	44.18	/
HuWSF [42]	50.00	/
RDBN [43]	53.60	/
PCRN [9]	62.49	59.40
Bi-LSTM [46]	/	44.79
Bi-GRU [47]	/	41.67
CNN [48]	/	54.17
CLDNN [45]	/	43.75
CapsNet [49]	/	56.25
HPCB (Ours)	/	56.50

The proposed HPCB model achieves much better performance on the CASIA and EMODB databases than the previous models. Among them, HPCB achieved 79.67% and 84.65% recognition performance on databases CASIA and EMODB, respectively. On database SAVEE, HPCB achieved 56.50% recognition accuracy, the UAR of HPCB on the SAVEE database is only 2.90% lower than that of the literature [9]. This suggests that the proposed model has good robustness and generalization.

5. Discussion

In this section, we further analyzed the effectiveness and robustness of the proposed system over the databases CASIA, EMODB, and SAVEE. We used the common evaluation matrices, such as weighted, unweighted, and F1-score to estimate the class level and the overall accuracy. In order to measure the model prediction performance among the actual

and the predicted labels, the confusion matrix of each dataset is shown. The confusion among the actual and the predicted labels of each class is shown in certain rows and columns in the confusion matrix. We conducted comprehensive experimentation for three datasets to show the model prediction performance in terms of precision, recall, F1-score, weighted, and unweighted results. We chose an optimal model combination for an efficient SER system.

The experimental results show that, for SER in different datasets, HPCB achieves higher performance compared with the other methods. The advantage of the CNN network is that it shares convolution kernels and automatically performs feature extraction, making it suitable for the high-dimensional data. However, at the same time, the pooling layer loses a lot of valuable information by ignoring the correlation between the local and the whole. This makes CNN fail to obtain high accuracy in learning time series. When dealing with tasks related to time series, LSTM is usually more appropriate. However, in terms of classification (including SER), LSTM has an obvious disadvantage in performance. Therefore, HPCB gets an obvious advantage in the variable mapping, to preserve the more valuable information. Additionally, it also performs well on the task sensitive to time series. It can be concluded that the proposed model has excellent generalization ability for SER tasks.

6. Conclusions

In this study, a novel heterogeneous parallel acoustic model called HPCB was 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 were calculated to verify the robustness and generalization of the model HPCB. Experimental results on the databases EMO-DB and CASIA suggest that HPCB demonstrate stably leading 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 could 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 HPCB may face possible overfitting though the 0.5 dropout ratio is employed in learning for its complicated 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.

Author Contributions: Conceptualization, H.H. (Henry Han) and H.H. (Heming Huang); methodology, H.Z.; software, Zhang. H.Y; validation, H.H. (Heming Huang) and H.Z.; formal analysis, H.H. (Henry Han) and H.H. (Heming Huang); investigation, H.H. (Henry Han) and H.H. (Heming Huang); resources, H.Z.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, Zhang. H.Y, H.H. (Henry Han) and H.H. (Heming Huang); visualization, H.H. (Henry Han); supervision, H.H. (Heming Huang); project administration, H.H. (Heming Huang); funding acquisition, H.H. (Heming Huang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by RESEARCH ON SPEECH RECOGNITION OF TIBETAN AMDO DIALECT BASED ON DEEP TRANSFER LEARNING, grant number 62066039. This research 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.

Informed Consent Statement: Not applicable.

Acknowledgments: Thank you for your support from the computer school of Qinghai Normal University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hsu, J.H.; Su, M.H.; Wu, C.H.; Chen, Y.H. Speech emotion recognition considering nonverbal vocalization in affective conversations. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1675–1686. [CrossRef]
2. Zhou, L.; Gao, J.; Li, D.; Shum, H.-Y. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguist.* **2020**, *46*, 1–62. [CrossRef]
3. Brosch, T.; Scherer, K.R.; Grandjean, D.; Sander, D. The impact of emotion on perception, attention, memory, and decision making. *Swiss. Med. Wkly.* **2013**, *143*, w13786. [CrossRef] [PubMed]
4. Tzirakis, P.; Zhang, J.H.; Schuller, B.W. End-to-end speech emotion recognition using deep neural networks. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093.
5. Tahon, M.; Devillers, L. Towards a small set of robust acoustic features for emotion recognition: Challenges. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 16–28. [CrossRef]
6. Reeves, B.; Nass, C.I. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*; Cambridge University Press: Cambridge, UK, 1996.
7. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. *Int. J. Speech Technol.* **2012**, *15*, 99–117. [CrossRef]
8. Vogt, T.; Andre, E.; Wagner, J. Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realization. *Affect Emot. Hum. Comput. Interact.* **2008**, *48*, 75–91.
9. Jiang, P.; Fu, H.; Tao, H.; Lei, P.; Zhao, L. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* **2019**, *9*, 90368–90376. [CrossRef]
10. Zhang, Y.; Liu, Y.; Weninger, F.; Schuller, B. Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4490–4494.
11. Lotfian, R.; Busso, C. Formulating emotion perception as a probabilistic model with application to categorical emotion classification. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 415–420.
12. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [CrossRef]
13. Ayadi, M.E.; Kamel, M.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
14. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
15. Schmitt, M.; Ringeval, F.; Schuller, B. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In Proceedings of the INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016; pp. 495–499.
16. Lee, J.; Tashev, I. High-level feature representation using recurrent neural network for speech emotion recognition. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1–4.
17. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the INTERSPEECH, Singapore, 7–10 September 2014; pp. 223–227.
18. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
19. Neumann, M.; Vu, N.T. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv* **2017**, arXiv:1706.00612.
20. Satt, A.; Rozenberg, S.; Hoory, R. Efficient emotion recognition from speech using deep learning on spectrograms. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 24–27 August 2017; pp. 1089–1093.
21. Li, P.; Song, Y.; McLoughlin, I.V.; Guo, W.; Dai, L.R. An attention pooling based representation learning method for speech emotion recognition. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 3087–3091.
22. Garg, U.; Agarwal, S.; Gupta, S.; Dutt, R.; Singh, D. Prediction of emotions from the audio speech signals using MFCC, MEL and Chroma. In Proceedings of the International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 25–26 September 2020; pp. 1–5.

23. Kumbhar, H.S.; Bhandari, S.U. Speech emotion recognition using MFCC features and LSTM network. In Proceedings of the International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 19–21 September 2019; pp. 1–3.
24. Cirakman, O.; Gunsel, B. Online speaker emotion tracking with a dynamic state transition model. In Proceedings of the International Conference on Pattern Recognition (ICPR), Cancún, Mexico, 4–8 December 2016; pp. 307–312.
25. Wang, K.; An, N.; Li, B.N.; Zhang, Y.; Li, L. Speech emotion recognition using Fourier parameters. *IEEE Trans. Affect. Comput.* **2015**, *25*, 69–75. [CrossRef]
26. Kim, Y.; Provost, E.M. ISLA: Temporal segmentation and labeling for audio-visual emotion recognition. *IEEE Trans. Affect. Comput.* **2019**, *10*, 196–208. [CrossRef]
27. New, T.L.; Foo, S.W.; Silva, L. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623.
28. Neiberg, D.; Elenius, K.; Laskowski, K. Emotion recognition in spontaneous speech using GMMs. In Proceedings of the INTERSPEECH, Pittsburgh, PA, USA, 17–21 September 2006; pp. 809–812.
29. Kokane, A.; Ram Mohana Reddy, G. Multiclass SVM-based language independent emotion recognition using selective speech features. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, India, 24–27 September 2014; pp. 1069–1073.
30. Fu, L.Q.; Mao, X.; Chen, L.J. Relative speech emotion recognition based artificial neural network. In Proceedings of the Pacific Asia Conference on Language, Information and Computing (PACLIC), Wuhan, China, 19–20 December 2018; pp. 140–144.
31. Xu, H.H.; Gao, J.; Yuan, J. Application of speech emotion recognition in intelligent household robot. In Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence (AICI), Sanya, China, 23–24 October 2010; pp. 537–541.
32. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
33. Khaki, H.; Erzin, E. Use of affect based interaction classification for continuous emotion tracking. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2881–2885.
34. Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B. Speech emotion classification using attention-based LSTM. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1675–1685. [CrossRef]
35. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
36. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
37. Zhao, Z.; Zheng, Y.; Zhang, Z.; Wang, H.; Zhao, Y.; Li, C. Exploring spatio-temporal representations by integrating attention-based Bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 272–276.
38. Sainath, T.N.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, long short-term memory, fully connected deep neural networks. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 4580–4584.
39. Suen, H.Y.; Hung, K.E.; Lin, C.L. TensorFlow-based automatic personality recognition used in asynchronous video interviews. *IEEE Access* **2019**, *7*, 61018–61023. [CrossRef]
40. Zou, F.; Shen, L.; Jie, Z.; Zhang, W.; Liu, W. A sufficient condition for convergences of Adam and RMSProp. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11119–11127.
41. Liu, Z.-T.; Xie, Q.; Wu, M.; Cao, W.-H.; Mei, Y.; Mao, J.-W. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing* **2018**, *309*, 145–156. [CrossRef]
42. Sun, Y.; Wen, G.; Wang, J. Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomed. Signal Process. Control* **2015**, *18*, 80–90. [CrossRef]
43. Wen, G.; Li, H.; Huang, J.; Li, D.; Xun, E. Random deep belief networks for recognizing emotions from speech signals. *Comput. Intell. Neurosci.* **2017**, *5*, 1–9. [CrossRef]
44. Tao, H.; Liang, R.; Zha, C.; Zhang, X.; Zhao, L. Spectral features based on local Hu moments of Gabor spectrograms for speech emotion recognition. *IEICE Trans. Inf. Syst.* **2016**, *99*, 2186–2189. [CrossRef]
45. Sainath, T.N.; Weiss, R.J.; Senior, A.; Wilson, K.W.; Vinyals, O. Learning the speech front-end with raw waveform CLDNNs. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1–5.
46. Dai, T.; Zhu, L.; Wang, Y.; Carley, K.M. Attentive stacked denoising autoencoder with Bi-LSTM for personalized context-aware citation recommendation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 553–568. [CrossRef]
47. Wang, H.; Zhao, D.Q. Emotion analysis of Microblog based on emotion dictionary and Bi-GRU. In Proceedings of the Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2020; pp. 197–200.

48. Lee, K.H.; Kim, D.H. Design of a convolutional neural network for speech emotion recognition. In Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 21–23 October 2020; pp. 1332–1335.
49. Sara, S.; Nicholas, F.; Geoffrey, E.H. Dynamic routing between capsules. In Proceedings of the Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1–11.