IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 19, 2019, accepted August 23, 2019, date of publication August 28, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938007

Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network

HAO MENG, TIANHAO YAN, FEI YUAN, AND HONGWEI WEI
College of Automation, Institute of Robotics and Intelligent Control, Harbin Engineering University, Harbin 150001, China
Corresponding author: Tianhao Yan (1825109095@hrbeu.edu.cn)

This work was supported by the National Natural Science Foundation (NNSF) of China under Grant 51379044.

ABSTRACT Speech emotion recognition is a vital and challenging task that the feature extraction plays a significant role in the SER performance. With the development of deep learning, we put our eyes on the structure of end-to-end and authenticate the algorithm that is extraordinary effective. In this paper, we introduce a novel architecture ADRNN (dilated CNN with residual block and BiLSTM based on the attention mechanism) to apply for the speech emotion recognition which can take advantage of the strengths of diverse networks and overcome the shortcomings of utilizing alone, and are evaluated in the popular IEMOCAP database and Berlin EMODB corpus. Dilated CNN can assist the model to acquire more receptive fields than using the pooling layer. Then, the skip connection can keep more historic info from the shallow layer and BiLSTM layer are adopted to learn long-term dependencies from the learned local features. And we utilize the attention mechanism to enhance further extraction of speech features. Furthermore, we improve the loss function to apply softmax together with the center loss that achieves better classification performance. As emotional dialogues are transformed of the spectrograms, we pick up the values of the 3-D Log-Mel spectrums from raw signals and put them into our proposed algorithm and obtain a notable performance to get the 74.96% unweighted accuracy in the speaker-dependent and the 69.32% unweighted accuracy in the speaker-independent experiment. It is better than the 64.74% from previous state-of-the-art methods in the spontaneous emotional speech of the IEMOCAP database. In addition, we propose the networks that achieve recognition accuracies of 90.78% and 85.39% on Berlin EMODB of speaker-dependent and speaker-independent experiment respectively, which are better than the accuracy of 88.30% and 82.82% obtained by previous work. For validating the robustness and generalization, we also make an experiment for cross-corpus between above databases and get the preferable 63.84% recognition accuracy in final.

INDEX TERMS 3-D Log-Mel, dilated CNN, residual block, center loss, BiLSTM, attention mechanism.

I. INTRODUCTION

Affective content analysis has been an active research area in recent score years, and it's worth paying close attention that human emotional state is the most significant factor in human's communication. Human can run judgement on the emotions according to expression, speech, paralanguage and so on. However, speech is an efficient and essential bridge in human's communication that the speech signal has also become the latest and fastest system between Human-Machine Interfaces (HMI) which is bound to have emotional intelligence and more harmonious. Therefore, recognizing speech emotion is one of the crucial research directions in emotion detection and recognition naturally [1]. It is widely used in the field of education and healthy once the SER theory is earliest proposed [2].

Speech emotion recognition plays an important role in the HMI. As well known, thanks to recent advances in accurate speech recognition and replenishing wide availability of speech recognition devices, the speech emotion recognition (SER) has made great progress in the last score years since researchers are increasingly engaged on the SER experiments. In spite of that nevertheless, owing to taking account the speakers' sexuality, age even physical state in the process of the SER, it's very difficult to draw any affective information and specific emotion from voice of an individual.

In order to conquer various situation, researchers have run amounts of trials in the each stage of SER, such as the stage of signal preprocessing, features extracted and classification [2], [3]. In the course of picking up features like low-level

The associate editor coordinating the review of this article and approving it for publication was Xinyu Du.

description (LLDs) and high-level statistics functions (HSFs) that are raised and arranged from a great deal of conferences such as the INTERSPEECH 2013 ComParE feature set and so on [4]–[6]. In addition, many researchers develop some new features from speech signal that makes the network model obtain better accuracy in the recognition result. Turgut et.al. employed a novel feature selection method which was based on be given four definitions about feature set and emotion state [7]. Shaoling Jing proposed a novel type of features related to prominence together with traditional acoustic features which was used to classify emotional states [8]. Qirong Mao et.al. adopted a DA based method called Emotion-discriminative and Domain-invariant Feature Learning Method (EDFLM) for SER that both domain divergence and emotion discrimination were considered to learn emotion-discriminative and domain-invariant features [9]. However, it occurs to a considerable problem that not all of these features are effective for emotion recognition, so far researchers still haven't found the optimal feature set in the layer of frames or utterances.

Though there is still a long way to go before we really find the most suitable feature set, the innovation of the classification model renders the scientists a great surprise. They have invested their energy into the scientific experiment to enhance performance in the stage of classification. No matter the traditional algorithm like KNN, GMM, SVM or deep learning framework it is, people have made a great contribution on how to build the model. Han *et al.* [10] proposed to use the segments with highest energy to train a DNN model to extract effective emotional information. Trigeorgis *et al.* [11] directly utilized the raw audio signal to train a convolutional recurrent neural network (CRNN) to predict continuous arousal /valence space. Chen and He et al. [12] proposed a 3-D attention-based convolutional recurrent neural networks(ACRNN), which was drafted by the learning ability of relevant feature representations for specific tasks from attention mechanism that was raised firstly in networks by Bahdanau *et al.* [13]. Last year, temporal convolution networks was proposed for NLP, which show the better performance than LSTM by Bai *et al.* [14]. On the basis of TCN, we propose a novel architecture that was the ADRNN (dilated CNN with residual block and BiLSTM based on the attention mechanism) and we find numerous superiorities by means of using the ADRNN framework for speech emotion recognition. a) ADRNN is the best dominant position that makes convolution operation more convenient and works in parallel compared with RNN. b) ADRNN possesses more flexible ability by means of piling more convolution layers, using bigger dilated coefficient and magnifying the filters' specification. These advantages help model read more useful historic information that realizes what content should be remembered. c) ADRNN only needs to less memory in the process of training even long-term sequences.

In this paper, we will utilize a novel ADRNN framework to assist us classify the emotional state from speech signal that is a kind of time-varying signal which needs special processing to reflect the time-varying property. Therefore, we extract 3-D Log-Mel spectrum features from raw speech signal, and then we feed these features into the model to classify after postprocessing. The experimental results show that the designed ADRNN network does not only recognize the speech emotion effectively, but also has better generalization ability. High recognition accuracy and favorable generalization can provide a guarantee of designed networks for application in some familiar area such as health care, social education etc.

Our original contributions of the work are as follows: Section II firstly introduces the related methods or frameworks in the area of SER at large. Section III, we will explain the ADRNN framework clearly. Section IV presents the experimental results on the IEMOCAP database, Berlin EmoDB corpus and cross-corpus between two databases respectively. Section V, we conclude this paper and point out the future work in final.

II. RELATED WORKS

Speech emotion recognition has attracted a great deal of attention in human-centered signal processing research. In recent years, many researchers divert their sight line on how to extract features from speech signal and take advantage of these to get the state-of-the-art recognition accuracy. The model is built increasingly that it plays an important role in the process of emotional state. In the last decade years, the model development of SER has undergone grant changes from which based on the machine learning algorithm to deep learning framework. Among many classification methods, DNNs are widely used in speech emotion recognition. W. Zheng et al. constructed a CNN architecture to implement emotion recognition on labelled data, the ultimate experimental results showed that their proposed approach outperformed the SVM classification [15]. John et al. exploited the eGeMAPS features of the temporal information and put them into EmNet which was designed as the classifier, the final result achieves a state-of-the-art performance of 88.9% recognition rate in EMO-DB database [16]. J.Huang et al. applied the triplet loss in the model which utilized the LSTM to extract features and train, and then put those features into SVM to go to classify in the IMOCAP. It was a remarkable fact that researchers use the cycle mode that was chosen in the process of padding for gaining the isometric sequence [17]. Z. Zhao et al. combined attention-based BLSTM with fully convolutional networks and FCNs to apply for speech emotion recognition that showed more accurate predictions compared with other existing algorithms [18]. S. Mirasamadi used the relevant features that was automatically discovered emotionally from speech and adopted local attention along with DNN/RNN in order to focus on specific regions of a speech signal that were more emotionally salient [19].

D.T presented three modeling methods under the end-to-end learning framework and validated CRNN model to own the best accuracy into them, except that they added the data augmentation and balancing to further enhance the

system performance that achieves 48.8% UAR on the develop dataset in final [20]. W. Han et al. proposed to treat an utterance's label sequence that was used to put into the RNN based connectionist temporal classification modal to judge the emotional segments' labels, and then got the better results compared to other algorithm to demonstrate the effectiveness [21]. M.Sarma presented the TDNN-LSTM-Attention model based on the TDD-Statistics Pooling that was found to get better accuracy, furthermore they also made the data augmentation and achieved the 70.6% UA at last [22]. Jaebok Kim used a novel architecture that was the Convolution Highway Layer accomplished with identity skip-connections of ResNet based on fusing a great deal of databases such as FAU-AIBO, EMODB, and other common corpus to adopt the CNN LSTM to go to classification model [23]. J.Zhao utilized the 1D CNN LSTM to recognize speech emotion from audio clips and the 2D CNN LSTM to focus on learning global contextual information from the handcrafted features, and then applied the deep learning architecture to yield recognition accuracy 52.14% of speaker-independent [24]. Jaebok Kim also used the deep 3D CNNs to process the modelling spectrum-temporal dynamics with t-distributed stochastic neighbour embedding for SER, and then got 51.2% recognition accuracy [25]. Toktam Zoughi et al. adopted the gender-aware deep Boltzmann machine (GADBM) which was used for DNN pre-training, because the Boltzmann could exploit the additional information to improve the recognition accuracy possibly and the results showed that it outperformed DBN and DBM respectively [26]. Z. Huang proposed to learn affect-salient features for SER to use semi-CNN that includes the two stages, one was utilizing the unlabeled samples to learn candidate features, the other was using a novel objective function to encourage the feature saliency, orthogonality and discrimination [27]. Qirong Mao et al. exploited CNN based on the learning affect-salient features that were used to learn LIF and included unlabeled samples, and then the features were utilized a variant of sparse auto-encoder (SAE) with reconstruction penalization to continue doing salient discriminative feature analysis (SDFA) for SER [28].

Our work differs from the others' work mentioned above. The designed model is based on the attention mechanism DRNN networks that can learn more available information and draft some advantages of dilated convolution networks with residual block, BiLSTM networks, attention mechnism and loss function. Hence, we acquire high recognition accuracy in the speech emotion recognition at last.

III. PROPOSED WORK

In this section, we will introduce our proposed work for SER at large and we improve the X.J He's experiment [12] to acquire the better results. Firstly, we utilize the static of log-Mel spectrum, delta and deltas-deltas that are combined to make up the feature vector together from raw speech signal as proposed model's input. Then we will introduce our innovativeness architecture that adopts the dilated convolution and residual unit which is the skip connection trick attached with the tradition convolution networks, and then take advantage of the feature from DRN (dilated CNN with residual block) to be fed into the BiLSTM layer to extract further features, and make them pass through the attention mechanism at last. In additional, we optimize the loss function to use the center loss which helps our model distinguish the features more easily. According to our experiment, we get the better results compared with ACRNN model and other previous works in the area of SER. In this paper, our entirety framework picture is shown in the Figure 1.

A. 3-DIMENSION LOG-MEL SPECTRUM

In recent years, CNN (convolution networks) and their different variants acquire the tremendous achievement in the image classification. And we will get some successful experience from the image classification to apply for speech emotion recognition. We will take the static of Log-Mel spectrums as the one of our inputs that means a feature map if it's in the area of image, and the others are the Log-Mel spectrums' deltas and delta-deltas respectively. These three feature maps comprise the whole model's input that can be utilized to acquire a better performance.

The massive experiments show us the Mel spectrum is more suitable in human's auditory sense characteristic that presents the linear distribution under the 1000Hz and the logarithm growth above the 1000Hz, we utilize this point to obtain the Log-Mel spectrum static. The relationship between the Mel spectrum and the frequency is shown in (1).

$$f_{mel} = 2595 \cdot \lg(1 + \frac{f}{700Hz}). \tag{1}$$

We adopt the number of 40 filterbanks to process the raw signal under the control of the 16kHz sample rate and the length of FFT is set to 512. Furthermore, we choose the hamming window which is taken the window length of 25ms and the window shift of 10ms to add into the signal. Before gaining the 40 Mel-filterbank vectors, we also select the lower frequency of 300 and the upper frequency of 8000. Then we will take the signal to feed into the filterbanks to get the$H_m(k)$, which is shown in (2).

$$H_m(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 & others. \end{cases} \tag{2}$$

According the results of computing, we will get the outputs from the filterbanks, and then multiply the energy spectrum is used by the STFT (short-time Fourier transform) processed from the raw signal, which is shown in 3.

$$Log\_MelSpec(m) = \sum_{k=f(m-1)}^{f(m+1)} \log(H_m(k) * |X(k)|^2). \tag{3}$$

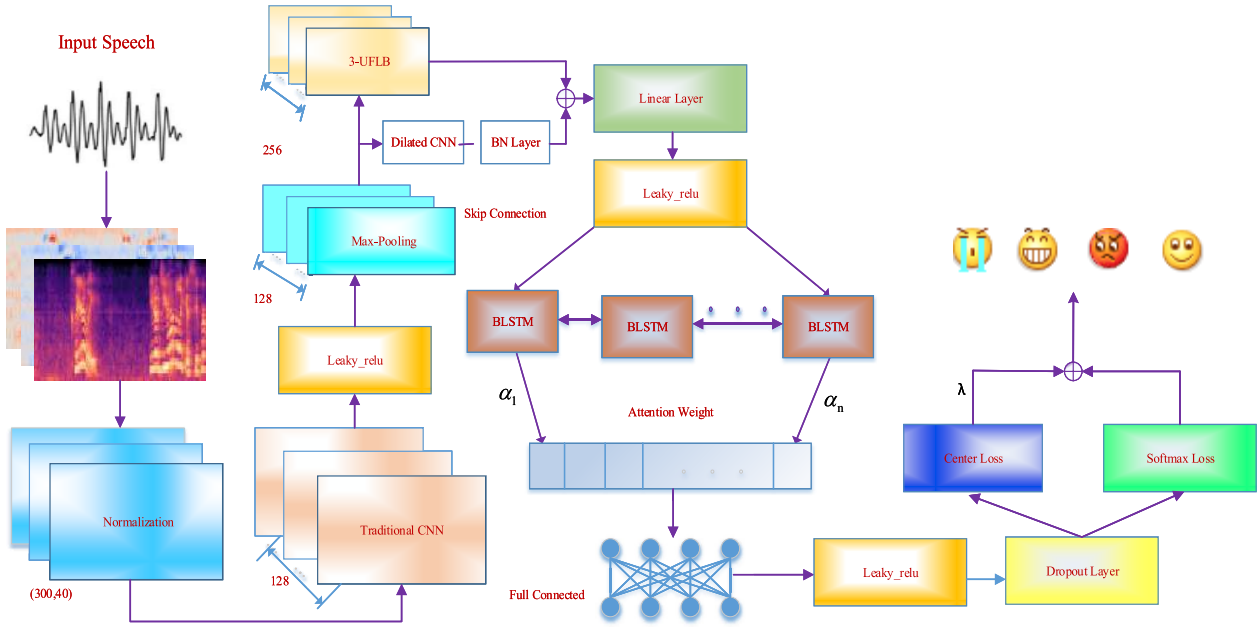

FIGURE 1. The whole structure of dilated CNN with residual block and BiLSTM based on the attention mechanism.

where the $|X(k)|^2$ describes the energy spectrum in the points of kth energy, m is the number of the filterbanks and k is the point of the FFTs.

We will obtain the 40 mel-filterbanks outputs that are the Log-Mel spectrums from this formula as the 1-D of inputs. In addition, we could extract the functional values of Log-Mel that are the deltas and the deltas-deltas. Then we could constitute the 3-D inputs with the statics like the three feature maps. The static only describes the energy spectrum envelope base on the frame level of speech, but the signal could own some dynamic information. Therefore, it is significant to extract the functional value in the classification model. We take the value according to the 4 and 5.

$$L\_M(m)^d = \frac{\sum_{n=1}^{N} n(L\_M(m)_{t+n} - L\_M(m)_{t-n})}{2\sum_{n=1}^{N} n^2}. \quad (4)$$

$$L\_M(m)^{dd} = \frac{\sum_{n=1}^{N} n(L\_M(m)^d{}_{t+n} - L\_M(m)^d{}_{t-n})}{2\sum_{n=1}^{N} n^2}. \quad (5)$$

where the t describes the number of frames, N is set to 2 based on the popular experience and L_M is the abbreviation of Log_MelSpec.

Similarly, the delta-deltas features are calculated by taking the time derivative of the deltas. Finally, we get three dimensions features $\mathbf{X} \in \mathbf{R}^{\mathbf{s,f,c}}$ that are used to be inputs in the convolution networks. The input expresses the time that is set changeless to 3 seconds by OpenEar [29] in every speech data that helps us align the tensor as the inputs. And the f is set to the number of 40 in this task that it means to utilize the 40 filterbanks to extract the features. And the c is set to 3 channels that denote the three feature maps which are represented the static, deltas and delta-deltas respectively. The waveform and spectrogram from the speech signal are shown in the Figure 2.

B. THE DILATED CONVOLUTION NETWORKS

In this SER mask, we adopt dilated convolution networks to instead of the tradition CNN, because the dilated CNN could acquire more comprehensive receptive field in the process of features extraction. The dilated CNN is applied for semantic segmentation to fetch a better performance. We use its characteristic of holding the interior information in the mask. In the area of the image classification, it's routine to add the pooling layer for reducing the size of image to raise the receptive field and to follow the upsampling to expand the size of image. However, in the process of changing the size of image, there is some info that could be ignored. Hence, the dilated CNN is developed to overcome this trouble. In this paper, we utilize the tradition CNN compared with 3*3 filter kernel, set the one stride at first. Then, we add the max pooling with 2*4 kernel size and 2*4 stride at the same time. The padding is chosen the "VALID". Furthermore, we adopt the trick of dropout as the last layer in first part.

In a subsequent part of this chapter, an upgrade feature learning block (UFLB), a substitute of dilated CNN, is designed to extract emotional features. Each UFLB consists of one dilated convolutional layer among the 3*3 dilated kernel size, one batch normalization (BN) layer and one leaky relu layer as illustrated in Figure 3. we utilize the three UFLBs structure to establish our CNN model that the most outstanding convolutional layer is not only spatially local connectivity and shared weights, but also could be avoided ignoring some information due to owning more receptive fields [31]–[33]. We can adjust the value of rate to be 2 and

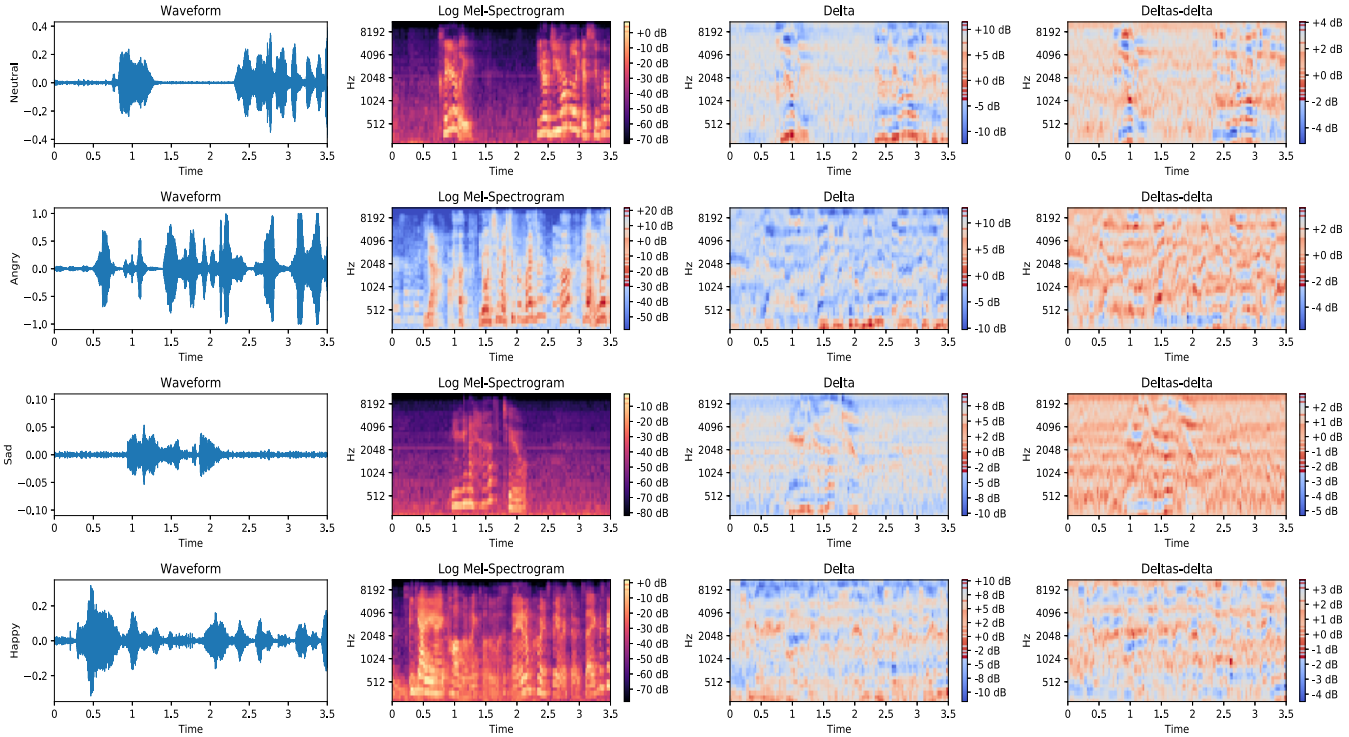

FIGURE 2. Waveforms and 3-D Log-Mel spectrograms of the Neutra, Angry, Sad, and Happy emotional state respectively in the IEMOCAP database.

the stride is fixed to one. These properties allow the feature extracted to restore more info and perform the function of the learning kernel. BN layer helps us process the features normalization and improves the performance and stability of deep structure, because we could utilize the bigger learning rate to train our model, raise the training speed and avoid the vanishing gradient problem. The leaky relu layer defines the output of the BN layer different from the relu activation. It could generate the phenomenon of dead neuron when the input value of relu is negative due to the fact that the output value is zero and the first-order derivative is zero at the same time. Therefore, we use the leaky relu to be our model's activation function to guarantee the nolinearism shown in 6.

$$y_i = \begin{cases} x_i & if\ x_i \geq 0 \\ \dfrac{x_i}{a_i} & if\ x_i < 0 \end{cases} \tag{6}$$

where a_i is the fixed value between 1 and positive infinity.

We propose the dilate convolution layer help us keep back a large of useful info instead of using the max-pooling layer, and the CNN plays the important role of a local feature extractor. Ibidem, our inputs consist of the data that is processed by 40 filterbanks in the length of 3 seconds which are substituted for 300 points as the height and width respectively. It means the data will be convolved with the convolution kernels across of the input volume. Then, the three feature maps include the Log-Mel static, deltas and delta-deltas are produced by computing the dot product between the input and the CNN's kernel respectively. The detailed layer parameters of our proposed network is illustrated as Table 1.

TABLE 1. The layer parameters of the whole network. The output dimension is represented as height * width * number. M * N is the size of the low-level features. The kernel size E of FCN is the number of the emotions. In the framework of UFLB, we set the dilation rate that is equal to two instead of stride.

Name	OutputDim	Kernel Size	Stride
Traditional CNN	M*N*128	3*3	1*1
Max-Pooling	M/2*N/2*128	2*4	2*4
1UFLB	M/2*N/2*128	3*3	–
2UFLB	M/2*N/2*128	3*3	–
3UFLB	M/2*N/2*128	3*3	–
Linear	–	–	–
BLSTM	–	256	–
Attention Mechanism	–	–	–
FCN	–	E	–

If the input of the convolution layer is $x(i, j)$, the result $z(i, j)$ can be acquired by convolving the signal $x(i, j)$ with the convolution kernel $w(i, j)$ of size c*d as shown in 7:

$$z(i,j) = x(i,j) * w(i,j) = \sum_{a=-c}^{c} \sum_{b=-d}^{d} x(a,b) \cdot w(i-a, j-b). \tag{7}$$

And then, we will get the value of $z(i, j)$ to put into the BN layer after the convolution layer processing shown in 8, which normalizes the activations of the previous layer at each batch [34]. BN layers apply a vital role that maintains the

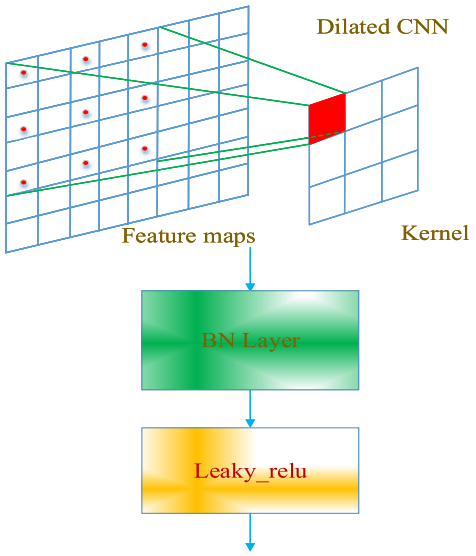

FIGURE 3. The structure of an upgrade feature learning block (UFLB) in feature extraction (rate=2).

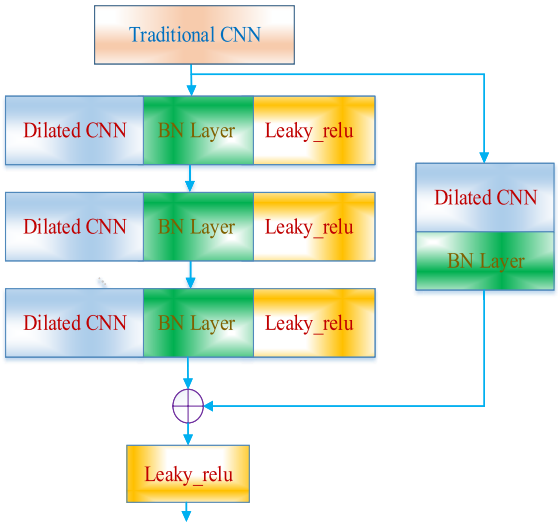

FIGURE 4. The structure of the skip connection with dilated convolution networks.

mean of the convolved features close to zero and the variance of the convolved features close to one based on each batch sample during the training process shown in 9.

$$z_i^l = b_i^l + \sum_j z_i^{l-1} \cdot w_{ij}^l. \tag{8}$$

$$Z_i^l = BN(z_i^l) = \gamma(\frac{z_i^l - \mu}{\sqrt{\sigma^2 + \varepsilon}}) + \beta. \tag{9}$$

where z_i^l and z_i^{l-1} denote the i-th output feature at the l-th layer and the j-th input feature at the (l-1)-th layer; w_{ij}^l represents the convolution kernel between the i-th and j-th feature. μ and σ^2 describe the i-th output feature's mean and valence at the l-th layer. ε and β explain the learnable parameters to highlight the expressive power of the networks. We can set these two parameters to adapt our model.

In final, we will put the Z_i^l learned by BN $(\cdot)$ normalization into the formula 6 that is the leaky relu activation function of network.

C. RESIDUAL BLOCK STRUCTURE

In the process of features extraction, we also add the trick of the skip connection to compose the Residual framework in the dilated convolution network. We could make the networks increase the depth rather than degeneration according to overlying a layer of "$\mathbf{y} = \mathbf{x}$" in the shallow networks, we call it "identity mappings". It means to gather the raw features and the features which learned by convolutional layer shown in 10 and 11. The reason for doing it is not only utilizing the current info, but also avoiding losing foregone info. It is shown the framework in Figure 4. After passing through the leaky relu layer, we get a new Z_i^l. Due to adopting the three dilated convolutional layers, we will use the trick of skip connection between the (l-2)th layer and l-th layer.

$$z_i^{l-1} = w_i^{L-1} \cdot z_i^{l-2} + b_i^{l-1}. \tag{10}$$

$$Z_i^{l-1} = \varphi(BN(z_i^{l-1})). \tag{11}$$

Similarly, we could generate the output of the residual block shown as formula 12.

$$Z_i^l = \varphi(z_i^l + Z_i^{l-2}). \tag{12}$$

where the z_i^{l-1} and z_i^{l-2} denote the i-th feature in the (l-1)-th layer, i-th feature in the (l-2)-th layer of dilated convolutional network. w and b express the corresponding weight and bias. φ represents the leaky relu activation function.

Finally, we get the Z_i^l that will be sent the BLSTM layer after the DRN structure which is fused the dilated CNN and residual block.

D. GLOBAL FEATURE LEARNING

RNN (recurrent neural network) helps to solve the current policy decision by means of keeping the historic info such as to utilize past speech frame to augment the present speech frame to make the whole utterance turn clearly to recognition emotion. The networks can take more advantage of net framework to build the model info that doesn't emerge in the tradition network. However, it could generate a serious problem that it happened to the long-term dependencies. Because every historic sequence point doesn't have effect on the current policy decision. Therefore, LSTM (long short-term memory) is designed for overcoming this problem and getting better result and efficiency in the performance [36]. It is a special model of RNN, which could learn the long-term info or sequence selectively. Compared with the recurrent structure of the single tanh, LSTM owns the three "door" that is an input gate, an output gate, a forget gate. The forget gate helps the model forget some useless info in the light of the input x_t and the previous time output h_{t-1} is to decide what to

forget [37], [38]. We assume that Z_t^{l-1} is a high-level feature vector obtained from the framework of CNN at time (t). i_t, f_t, Z_t^l, and C_t, represent input, forget, output gate and a cell with a self-recurrent connection respectively [23]. Then the updating of a LSTM unit is shown in (13)-(17):

$$i_t = \sigma(W_i Z_t^l + U_i h_{t-1} + b_i). \tag{13}$$
$$f_t = \sigma(W_f Z_t^l + U_f h_{t-1} + b_f). \tag{14}$$
$$o_t = \sigma(W_o Z_t^l + U_o h_{t-1} + b_o). \tag{15}$$
$$C_t = f_t \cdot C_{t-1} + \tanh(W_c Z_t^l + U_c h_{t-1} + b_c). \tag{16}$$
$$Z_t^l = o_t \cdot \tanh(C_t). \tag{17}$$

where σ denotes the function of sigmoid, the W_i, U_i, W_f, U_f, W_o, U_o are their weight matrices respectively and corresponding b represents the value of bias.

In this paper, we adopt the bidirectional LSTM that is upgraded in the LSTM. The bidirectional structure helps the model sequence that is not only learning from the previous sequence, but also the future sequence which could have an effect on the current state. It means that this framework could calculate to get and restore the output of hidden layer from first-time to t-th time in the forward layer, and continue the same operation to reverse computing from t-th time to first-time in the backward layer, and combine the output from forward and backward to get the final result. It could be shown in 18,19 :

$$i_t^* = \sigma(W_i^* Z_t^l + U_i^* h_{t+1} + b_i^*). \tag{18}$$
$$o_t^* = \sigma(V \cdot i_t + V^* \cdot i_t^*). \tag{19}$$

where the i_t^* denotes the input based on the future sequence, W_i^* and U_i^* represent the new weights aimed at the backward layer. o_t^* describes the new result. The others parameters from LSTM is similar with description like i_t^*. In the last, we get the new Z_t^l according to the BiLSTM layer.

After the performance of the recurrent networks, we will add the attention mechanism in the experiment. The attention-based model has been successfully used in plenty of sequence-to-sequence learning that is employed to concentrated on emotion relevant parts and produce discriminative utterance-level representations for SER [39]. It is just like the framework of Encoder-Decoder, and the part of encoder is the feature vector from Log-Mel according to the DRN. Then the part of decoder is from the output of BiLSTM. The attention mechanism is to select relevant encoded hidden vectors via attention weights during the decoding phase instead of simply performing a mean or a max pooling over time. In the attention structure of BiLSTM, we define the c_t that is the significant part to draw into the attention in 20.

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j. \tag{20}$$

where the h_j denotes the j-th hidden condition info of encoding input that is $h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}]$ at time step t. It could split joint the hidden info between forward and backward. There is also vital variable α that is described which sequence could generate the major output info shown in 21.

$$\alpha_{ij} = \frac{\exp(W \cdot h_t)}{\sum_{k=1}^{T} \exp(W \cdot h_t)}. \tag{21}$$

where α_{ij} represents the relationship probability of the j-th sequence to the current sequence of output. Therefore, we calculate the utterance-level representations c_i that will put into a fully connected layer(FCN) with output unit to obtain higher-level representation. In the last process of classification, we will adopt the softmax loss together with center loss to be our loss function that makes our model more perfect.

E. CENTER LOSS

After the feature vectors are set into the FCN, we choose the softmax loss and center loss to achieve the final classification collectively. In our task, not only do we learn the separable feature, but we also intend these features to be discriminative. That's why we need to utilize the fuse of two kinds of loss function. It could cause the issue that the robustness of training model become poorer because the distance of within-class turns too large when we only use the softmax loss that could make the model perform excessive confidence so that the classification result is basically not 1 or 0 especially in the boundary points. Compared with utilizing the softmax loss alone, center loss together with softmax loss are both taken advantage of classification that overcome the above shortcomings at the same time. We could adjust the distance of within-class and between-class by means of setting the value of λ in the process of exploiting the loss function. The center loss simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers [40]. Specifically, the softmax loss could guarantee the maximum distance between the class and the center loss could assure the minimum within-class that shown in 22 and 23.

$$\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \cdot x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \cdot x_i + b_j}}. \tag{22}$$
$$\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^{m} ||x_i - c_{y_i}||_2^2. \tag{23}$$

where m denotes the mini-batch that will be set the classifier every step and n represents the number of class. c_{y_i} describes the i-th sample corresponding to the class of y_i that belongs to the central of class '$\mathbf{c} \in \mathbf{R^d}$' and d is the dimension. In order to make the c_{y_i} update in real-time along with features to be learned and avoid the misclassification, we add a parameter λ to limit the value of center loss to achieve the balance between two kinds of loss shown in 24.

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C. \tag{24}$$

$\mathcal{L}$ consists of the final function of loss in our paper. According to the performance, the larger the hyperparameter of λ we

set, the smaller the spacing of within-class will be. It could generate a tiny improvement in the recognition result after adopting suitable λ and we will verify the superiority of center-loss function in the next section.

IV. EXPERIMENT

A. DATABASES

We employ the Interactive Emotional Dyadic Motion Capture Corpus (IEMOCAP) [41] and the Berlin Database of Emotional Speech(EMO-DB) [43] to evaluate the effectiveness of our proposed model with a given emotional state that is described by the static, deltas and delta-deltas of Log-Mel spectrums. The IEMOCAP corpus was collected following theatrical theory in order to simulate natural dyadic interactions between actors that include five sessions, and each session contains utterances from two speakers (one male and one female). It is seized of two emotion elicitation methods which include performance of theatrical scripts and improvisations of affective scenarios used in the dialog recording. Each wave file has segment level emotion category label annotated by three human annotators at least. There are 10039 utterances in the prototypical data (complete agreement on the affective state from evaluators) of improvisations from affective scenarios that cover the 4.5 seconds with an average duration in each Session (Angry, Excited, Frustrated, Happy, Neutral, Sad, and Surprise). It contains the seven emotional state and the sample rate is 16kHz. The Berlin EMO-DB is recorded in 2005 contains seven emotions, including seven emotions (neutral, fear, joy, angry, sadness, disgust and boredom), and each emotion consists of nearly the same number of utterances to evaluate the classification accuracy properly. There are 535 sentences of the utterances displayed by 5 female and 5 male professional actors, which are from the life every day. This database sampled at 44.1 kHz, and later downsample to 16 kHz at the same time.

In this paper, we only use the utterances that are the affective improvisational scenarios spontaneously and don't take account with the performance of theatrical scripts with labels from the four following emotions: Angry, Happy, Sad, and Neutral in IEMOCAP corpus. And then we adopt all the utterances in Berlin EMO-DB database from seven emotions from our experiments.

B. EXPERIMENTAL SETUP

As for the feature extraction, we dispose the length of data to render the centralized three seconds, because we must construct the same dimension for better parallel acceleration as the input of the Dilated Residual Networks. If the utterance doesn't reach the three seconds, we adopt the zero-padding to supply the part of absent sequence, and if the length of utterance is more times than three seconds, we split the utterance into equal-length segments. The pretreatment of raw speech signal is same with the Chen and He et al. [12]. However, we have a tremendous different from him in the constructional detail of the feature extraction and the classification in the model. Firstly, we apply the 3-D feature (the static, deltas and delta-deltas of Log-Mel spectrum from 40 filterbanks) to be normalized by the global mean and the standard deviation. Then, we put them into our proposed DRN structure to train our database. After that, we go a step further to extract features from BiLSTM and attention mechanism. Finally, we employ the FCN and the loss function of sofmax with center loss to achieve the ultimate classification. The whole experiment is built in the NVIDIA Graphics card 1080Ti, and the model is implemented with Tensorflow toolkit [42].

C. SPEAKER-INDEPENDENT EXPERIMENTS

In the speaker-independent experiments, we perform the spontaneous emotion data of the IEMOCAP corpus in five session (each of session contains 1 male and 1 female) and all of the EMODB database that both contain 10 speakers (5 males and 5 females) that employ a 10-fold cross-validation technique between speakers. When eight of ten speakers are divided into the training set, the one of remaining speaker is used to be validation set, and the other is selected to be test set according to the subjects. And we report the average value and standard deviation from tenth cross-validation. It is rather remarkable that we employ the unweighted accuracy (UA) to be the final result on the test set due to the imbalanced test class distribution when the value of UA is the optimum on the validation set with two databases respectively.

For our baseline in this paper, we choose the X.J He's experiment to be our baseline as the comparison. The confusion matrix of classification conducts on IEMOCAP corpus and EMODB database are shown in the Figure 5 and Figure 6 respectively.

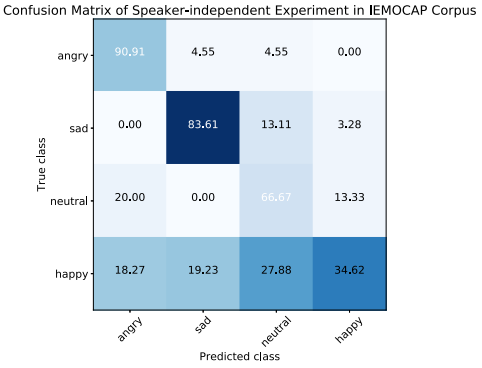

FIGURE 5. The Confusion Matrix in the Speaker-independent with unweighted accuracy in the IEMOCAP Corpus.

In the speaker-independent experiment of the IEMOCAP corpus with our architecture of networks, we acquire the obvious superiority on the angry emotion state of 92.00% with unweighted accuracy that will be close to the full percentage better than the 70.47% angry recognition accuracy from X.J He's model. And in the emotion state of the sad emotion and the neutral emotion, we gain almost equal performance result of 83.61% and 66.67% compared with them.

TABLE 2. SER Results (%) with Average and Standard Deviation based on Attention Mechanism and Function of Center Loss for our proposed networks in Ablation Study of two databases.

Features	Configuration of Network	SD (IEMOCAP)	SI(IEMOCAP)	SD(EMODB)	SI(EMODB)
Off-the-shelf	3×FCN+ELM	–	51.20±5.00	–	71.54±4.76
Raw Waveform	CNN+2LSTM	–	54.20±5.00	–	79.27±4.28
3-D Log-Mel Spectrum	1Traditional CNN+BiLSTM	67.34±4.21	64.20±5.73	86.83±2.68	81.53±4.76
	2Traditional CNN+BiLSTM	68.86±4.78	64.18±5.57	87.21±2.44	81.72±3.63
	3Traditional CNN+BiLSTM	69.37±4.47	64.11±5.23	88.02±2.29	81.84±3.07
	3Traditional CNN+BiLSTM+Residual Block	71.29±4.58	67.26±4.31	89.72±2.87	83.63±2.84
3-D Log-Mel Spectrum	1Dilated CNN+BiLSTM	68.56±4.21	64.33±4.76	86.97±2.59	81.83±3.29
	2Dilated CNN+BiLSTM	69.21±4.53	65.44±4.35	87.21±2.21	81.87±2.74
	3Dilated CNN+BiLSTM	69.67±4.64	65.46±4.63	88.67±2.04	82.93±2.12
	3Dilated CNN+BiLSTM+Residual Block	**74.96±4.27**	**69.32±3.76**	**90.78±1.62**	**85.39±1.86**

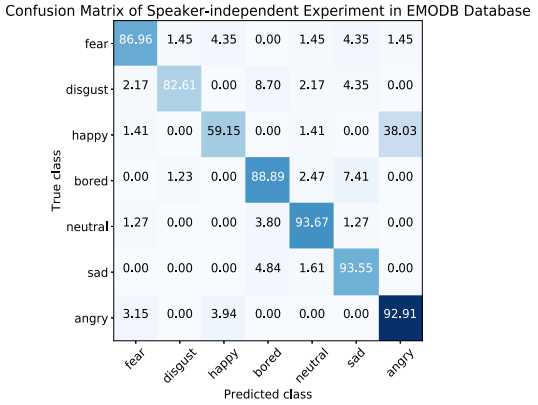

FIGURE 6. The Confusion Matrix in the Speaker-independent with unweighted accuracy in the EMODB database.

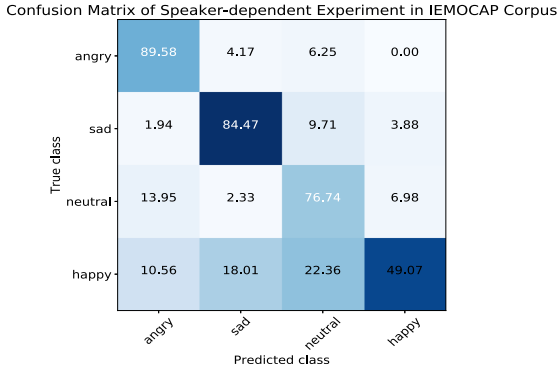

FIGURE 7. The Confusion Matrix in the Speaker-dependent with unweighted accuracy in the IEMOCAP corpus.

Despite of getting better performance, we still discover the happy emotional state engender the lower 34.62% recognition accuracy in the experiments and it has a 4.67% recognition accuracy improvement better than the result of X.J He, and the happy emotion is easily confused with other emotions like the neutral emotion in the classification.

From the speaker-independent experiment of the EMODB database, we can see that the emotions are recognized with high recognition accuracy in our proposed framework. Compared with contrast test, we acquire a 4.55% notable improvement in the angry emotion of 92.91% because our model could extract more valuable feature so that we can identify it better. And we gain the 88.89% and 86.96% recognition accuracy of the bored emotion and the fear emotion better than 83.41% and 81.76% respectively [12] compared with the contrast test. The neutral emotional of 93.67% recognition accuracy is equal to the contrast test and other emotional states are not diverse in comparison due to the less data volume in the database. It is worthwhile to note that we gain 12.29% larger promotion in the happy emotion of 59.15% accuracy the same as the situation of IEMOCAP corpus, but it is also extraordinary low recognition accuracy in total compared with other emotional states.

D. SPEAKER-DEPENDENT EXPERIMENT

In the speaker-dependent experiments, the division of the data set is not the same as the speaker-independent, we pick up all of the speech from the two databases to be a whole set respectively. We split the whole set into ten parts randomly with different random seeds, the training set takes 80% of ten parts, and the testing set takes the remaining 20% of the parts. Similarly, we report the average and standard deviation to get more reliable results from each model evaluation. In order to reduce overfitting and achieve the goal of our work which is to recognize speech emotion with high generalization performance and high accuracy, only the best predictive and fitted models are selected. Furthermore, we also report UA to be our final performance result with two databases.

The best experiment results of classification in the IEMOCAP corpus are illustrated in the Figure 7 and the experimental results conducted on EMODB database are shown in the Figure 8. For the baseline in this experiment, we perfect the whole performance and contrast with our other fundamental model with ablation study shown as the Table 2.

In the speaker-dependent experiment of the IEMOCAP corpus, we acquire the 74.96% recognition accuracy of emotion classification on average in the IEMOCAP database from confusion matrix. We take the previous work of P. Yenigalla et.al to be our contrast experiment [47]. From the Figure 7, we can see the 89.58% recognition result of the angry emotion better than the 55.9% accuracy of them and have the 33.68% improvement in the comparison. Then, the sad emotion and the neutral emotion achieve the 84.47% and 76.74% recognition accuracy to have a tiny promotion compared with

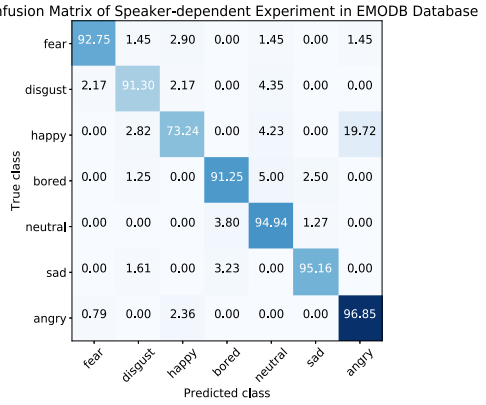

FIGURE 8. The Confusion Matrix in the Speaker-dependent with unweighted accuracy in the EMODB database.

the 83.50% and 75.50% result respectively. Moreover, it has a slight difference on the recognition rate of diverse emotions compared with them due to the lack of the database size of the happy emotion, we get the 49.07% experimental result in final. In other comprehensive contrast experiment, we raise the 1.2% accuracy rate in the four-class of emotional recognition compared with the 73.78% recognition rate of Kim *et al.* [44].

From the speaker-dependent experiment of the EMODB database, we find the recorded model does not only suit the experimental data well but also have the superior predictive performance to recognize different speech emotional state on average of 90.78% better than other previous work from the confusion matrix. We select the previous work of Y.M Huang to be our contrast work [43]. From the confusion matrix of Figure 8, we gain the obvious improvement on each emotional state like the sad emotion of 95.16%, the neutral emotion of 94.94%, the bored emotion of 91.25% and so on compared with them of 84.00%, 80.95%, 72.31% respectively. The 96.85% experimental result of the angry emotion is not much diverse and we both get better accuracy between our work. However, it can be seen that the 73.24% recognition accuracy of the happy emotion is still relatively low and easy to confuse with the angry emotion compared with other emotion, and the reason could be the similar as the speaker-independent experiment.

According to our analysis, we find the one point that the data size of corpus is less than other pattern database like text or image, and the other is that features of the happy emotion are easy to be misrecognized as the other emotion without the help of linguistic information from our practical experience so that the designed model are confused with them. This is probably due to the difference of the culture, environment and education of speakers [24]. Hence, recording a speech emotion database with a large number of utterances like the happy emotion is very significant to promote the development of this field.

E. CROSS-CORPUS EXPERIMENT

With the high-speed development of single-corpus in SER, humans have achieved a superior recognition accuracy in a specific experimental environment. However, when it simulates speech emotion recognition in a natural environment, there are many aspects of actual factors that could generate some passive effects such as the differences of language, surrounding noise, etc. It is a tremendous challenge for humans in the SER of cross-corpus because there are some influencing factors that include huge difference of speaker, language and culture. Then, there is a situation of context-dependent in the speech samples from the database that it doesn't match between the expressed emotional meaning and acoustic characteristics. Moreover, it also emerges the reverberation, background noise and other factors that have an effect on the acoustic conditions in the recording of diverse databases. Hence it is extremely vital for model to possess the property of robustness and generalization aimed at above some elements, and the model could be suitable for the cross-corpus in the SER.

In this paper, we also make a cross-corpus experiment between the database of IEMOCAP and EMODB. we select four emotions (sad, happy, angry, neutral) to complete our cross-corpus classification and utilize our proposed novel networks to apply for the corpus that we adopt the IEMOCAP to treat as the training set and test the model in the EMODB database. The experimental confusion matrix is shown in the Figure 9.

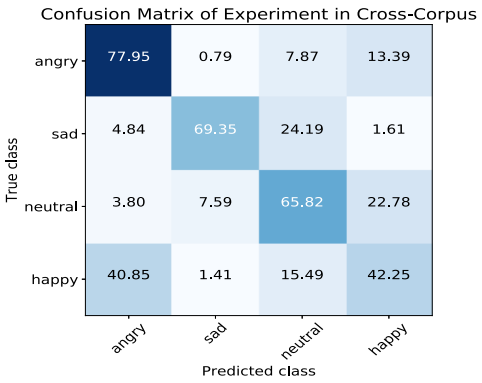

FIGURE 9. The Confusion Matrix in the Speaker-dependent with unweighted accuracy in the Cross-Corpus.

From the Figure 9, we utilize our proposed model to make progress in the cross-corpus experiments and we gain the 63.84% recognition accuracy rate on average of the four-class emotional classification. Both of the angry emotion and the sad emotion are obtained the 77.95% and 69.35% recognition rate respectively according to training our model. Compared to the previous works like Latif *et al.* [45], although we adopt the diverse databases to accomplish our cross-corpus research respectively, it can be confirmed that our model possesses strong robustness and generalization ability.

F. PARAMETERS OPTIMIZATION

In the process of setting the parameter, we make a comprehensive model to try our best to make it optimal. Firstly, we set the batch-size 60 that more fit to the limit of memory. Then, we choose the Adam optimizer that is renovated to add bias-correction and momentum based on the RMSprop to actualize the best effect. After that, we set the learning rate is 10^{-4} and the moment is set 0.9 for batch normalization layer. Moreover, we make a comparative experiment to choose the optimum number of λ from two databases as shown in the Table 3 and the Table 4. And we also compare the influence from utilizing different loss function for the results. Moreover, when we add the function of center loss together with softmax loss on the ADRNN, it outperforms other recorded models to fit the experimental data well and obtain better accuracy in the recognition of cross-corpus.

TABLE 3. The Recognition Results (%) of Speaker-dependent (SD) and Speaker-independent (SI) based on diverse value of λ in the IEMOCAP Corpus.

Model	λ	SD (%)	SI (%)
ADRNN + Softmax Loss (alone)	-	74.54	68.92
ADRNN + Softmax Loss (with Center Loss)	0.001	74.73	69.31
	0.01	**74.96**	**69.32**
	0.1	74.65	69.11

TABLE 4. The Recognition Results of Speaker-dependent (SD) and Speaker-independent (SI) based on diverse value of λ in the EMODB Database.

Model	λ	SD (%)	SI (%)
ADRNN + Softmax Loss (alone)	-	90.26	84.96
ADRNN + Softmax Loss (with Center Loss)	0.001	90.63	85.27
	0.01	**90.78**	**85.39**
	0.1	90.45	85.03

TABLE 5. The Recognition Results (%)of Cross-Corpus based on diverse Loss Function.

Model	UA (%)
ACRNN + Softmax Loss (alone)	58.68
ADRNN + Softmax Loss (alone)	63.37
ADRNN + Softmax Loss (with Center Loss)	**63.84**

From the Table 3-4, we can see that it has an improvement about 0.4% recognition accuracy in the speaker-dependent and speaker-independent experiments from IEMOCAP corpus respectively when we utilize the softmax loss function together with center loss compared with employing the softmax loss alone. And we enhance about 0.5% recognition rate in the EMODB database based on the same condition from two experiments. In addition, we compare the diverse value of λ that has effect on the experiment and the results illustrate that we obtain the best recognition accuracy when the value of λ is set to the 0.01 in the two databases. Furthermore, we also make an experiment to validate the effect on the cross-corpus and the result shows about 0.5% promotion that the center loss function is equally applicable to the cross-corpus experiment in the Table 5. The center loss could be beneficial to the distance of between-class so that model is robust to speech emotion recognition. It draws on this account to gain more classification accuracy with exploiting the trick of center loss function. Finally, the value of λ in the function of center loss is to set 0.01.

Furthermore, we also compare processing the total time from training and testing in diverse operation modes with the X.J He works in the model classification for proving the feasibility of our model. Before we take the measurement of time in the two databases respectively, we also confirm some hyper-parameters to maintain the consistence. In the speaker-independent experiment of IEMOCAP corpus, we set the iterations to be 2000, batch size to be 60, and learning rate to be 10^{-4}. It means that we merely compare the total time in the architecture of model between ACRNN and ADRNN and make other qualifications equally. Besides in the EMODB corpus, the experiments are conducted using the same approach and condition as the IEMOCAP, and it just changes some hyper-parameters to adapt the corpus and keep invariant between the two models. Therefore, we also set the iterations to be 500, batch size to be 5, and learning rate 10^{-4} from the speaker-independent of EMODB database. It is shown in the Table 6.

TABLE 6. The Processing Time in Training and Testing of speaker-independent Experiment compared with previous works.

Networks	IEMOCAP(s)	EMODB(s)
ACRNN [12]	13487	6811
ADRNN + Softmax Loss (alone)	13975	7233
ADRNN + Softmax Loss (with Center Loss)	13887	7187

From the Table 6, we can see that our operation mode spends more time in the training and testing, because it could be the framework of ADRNN that is more sophisticated than previous works. Absolutely, we sacrifice the training time to exchange for the recognition rate of SER duo to using higher model complexity. In addition, we also find that the function of center loss benefits to reducing the time than employing the softmax loss function alone. The reason could be determined by the advantage of the center loss function that it owns the property of spending feature space to return the tiny time in the process of training the model.

On the whole, we test and verify the robustness and stability of our model according to calculating the total time of training and testing in the operation mode compared with previous work, and it also evaluates our model that adapts the research of the speech emotion recognition. Though we find the model could take more time in the training and testing process, with the development of the computing hardware in future, it could be narrowing the gap between our model and previous works.

G. DISCUSSION

In the above subchapters, we achieve two kinds of classification experiment and acquire better performance with spontaneous speech data from IEMOCAP corpus and EMODB database respectively. Besides, we still contrast the tradition CNN structure and DNN-ELM [43] compared with our proposed architectures for validating our networks' effectiveness, and we make a great deal of ablation study about each of our innovative point with unweighted accuracy shown as the Table 2. We discover that the dilated CNN is nearly equal to the tradition CNN in the only one layer, because we consider the networks which gets loss receptive field in one layer, but with the number of layers increasingly, we utilize the three layers of dilated CNN gain less improvement. Although we don't get ideal promotion, we save the number of layers because there is no pooling layer in the structure and we still achieve about 1.3% improvement in the IEMOCAP database and 1% enhancement in the EMODB of speaker-independent on average. Then, we catch sight of the prominent enhancement when we adopted the residual block into the dilated CNN and BiLSTM with attention mechanism and we think the skip connection could obtain more useful info. In final, we find that it is optimum to use three dilated CNN with residual block and BiLSTM with the attention mechanism based on the loss function of softmax with center loss.

In addition, we also expand an experiment of cross-corpus that gain a preferable result that shows 63.84% recognition accuracy rate in the emotional state of four-class for confirming the robustness and generalization of our model. However, the neutral-happy emotion and the angry-happy emotion are also confused with each other at the same as the single-corpus experiment respectively. We consider the reason that the same speech emotion of different speakers from different cultures has not the same probability of being recognized or misrecognized, and the data size of the happy emotion is less than other emotions' data. At the same time, we acquire the best results in the process of comparing with other research works in the two databases respectively whatever speaker-dependent or speaker-independent shown as the Table 7-8.

TABLE 7. The comparison of average recognition accuracy (%) of our proposed methods in IEMOCAP Corpus with other well-established methods. The best performances are indicated in boldface.

Research Works	Speaker-dep (UA/%)	Speaker-indep (UA/%)
W.Q Zheng et al.[15]	/	40.02
K.Han et al.[10]	/	51.24
X.J He et al.[12]	/	64.74
Our Work	**74.96**	**69.32**

When compared with other well-established feature representations and methods on average recognition accuracy, the designed ADRNN network with center loss function also performs satisfactorily. Table 7 indicates that our proposed network conducted on the 3-D Log-Mel Spectrograms of IEMOCAP database also performs well. Table 8 shows that the average accuracy of our network conducted on the Log-Mel Spectrograms of Berlin EMODB achieves the highest accuracy.

TABLE 8. The comparison of average recognition accuracy (%)of our proposed methods in EMODB database with other well-established methods. The best performances are indicated in boldface.

Research Works	Speaker-dep (UA/%)	Speaker-indep (UA/%)
Z.W Huang et al.[27]	88.3	85.2
Y.M Huang et al.[46]	75.5	/
X.J He et al.[12]	/	82.82
Our Work	**90.37**	**84.99**

From our work, 3-D Log-Mel spectrograms to apply in the ADRNN networks, which consists of three UFLBs and BiLSTM layer based on the attention mechanism are built to learn log-mel spectrum static and its functional of emotion-related features. As a result of the time-varying of speech signals, we need more sophisticated analysis to reflect characters. The designed networks with the strength of dilated CNN, BiLSTM with attention mechanism and the loss function of softmax together with center loss are utilized to recognize the speaker's emotional state. The experiments have accomplished the task of learning more emotional info from the two given experimental databases and recognized the emotions with high accuracies in experiments. Besides, we evaluate the processing time of diverse operation modes in training and testing with previous work. Though it could spend more time due to the complexity of our model, we acquire about 4.58% improvement in the recognition accuracy and the performances which could hold the influence of the processing time. Moreover, We have also done a great deal of work to prove the generalization ability and robustness of the designed deep networks which are used to apply in the cross-corpus performance. The similar prediction performances of the designed networks in the cross-corpus experiment show that the designed networks are effective approaches for recognizing speech emotion.

V. CONCLUSION

In this paper, we propose a novel architecture ADRNN networks to recognize speech emotion. The method of how to learn local correlations and global contextual information from log-mel (static, deltas, delta-deltas) spectrograms of raw audio signal is investigated. UFLB which consists of one dilated convoltional layer, one BN layer and one leaky relu layer is designed to learn local features. When local features learned by UFLBs are reshaped, they are inputted into a BiLSTM layer. The BiLSTM layer can learn contextual dependencies from inputted local features. Then, we utilize the attention mechanism to further extract the useful features. In addition, we adopt the loss function of softmax together with center loss that benefits the global classification and acquire a tiny enhancement based on the ADRNN from the performance. Therefore, the features learned by the designed networks contain local information and long-term contextual dependencies. The performances of the networks are tested on two benchmark databases in

the speaker-dependent and the speaker-independent respectively. And we also perform the cross-corpus experiment between two databases for confirming our proposed model. The results show that our designed networks can learn distinguishing features and high-level abstractions of the emotional information. The comparisons of the experimental results show that the ADRNN networks have a certain advantage over ACRNN networks in overall performance. When compared with other well-established feature representations and methods, ADRNN networks also have the edge on average accuracy.

Though the deep networks proposed in this paper have obtained better performance in speech emotion recognition, there are many aspects still need to be promoted. In the future, we will still put eyes on the model diversity and find a flexible structure that could adapts to other speech corpus. Then, our proposed SER system can be integrated with other signals, like employing joint knowledge of the linguistic and paralinguistic components of speech to achieve a unified model for speech processing. We believe that we can acquire better performance for different emotional analysis task using the multimodal signal.

REFERENCES

[1] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 385–394, Oct./Dec. 2015.

[2] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[4] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proc. ISCA INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 3201–3204.

[5] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, Rob van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. ISCA INTERSPEECH*, Portland, OR, USA, Sep. 2012, pp. 1–4.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. ISCA INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[7] T. Özseven, "A novel feature selection method for speech emotion recognition," *Appl. Acoust.*, vol. 146, pp. 320–326, Mar. 2019.

[8] S. Jing, L. Chen, and X. Mao, "Prominence features: Effective emotional features for speech emotion recognition," *Digit. Signal Process., Rev. J.*, vol. 72, pp. 216–231, Jan. 2018.

[9] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition," *Speech Commun.*, vol. 93, pp. 1–10, Oct. 2017.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2014, pp. 223–227.

[11] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 5200–5204.

[12] M. Chen, J. Yang, H. Zhang, and X. He, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018. doi: 10.1109/LSP.2018.2860246.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*. [Online]. Available: https://arxiv.org/abs/1803.01271

[15] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact.*, Sep. 2015, pp. 827–831.

[16] J. W. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2018, pp. 937–940.

[17] J. Huang, J. Tao, Z. Lian, and Y. Li, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2018, pp. 3673–3677.

[18] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2018, pp. 272–276.

[19] S. Mirsamadi, C. Zhang, and E. Barsoum, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. 42nd IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[20] D. Tang, M. Li, and J. Zeng, "An end-to-end deep learning framework with speech emotion recognition of atypical individuals," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 162–166.

[21] W. Han, X. Chen, Z. Wang, H. Li, B. Schuller, and H. Ruan, "Towards temporal modelling of categorical speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 932–936.

[22] M. Sarma, D. Povey, N. K. Goel, K. K. Sarma, N. Dehak, and P. Ghahremani, "Emotion identification from raw speech signals using DNNs," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 1–5.

[23] J. Kim, K. P. Truong, V. Evers, and G. Englebienne, "Deep temporal models using identity skip-connections for speech emotion recognition," in *Proc. ACM MM*, Mountain View, CA, USA, Oct. 2017, pp. 1006–1013.

[24] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.

[25] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, 2017, pp. 383–388.

[26] T. Zoughi and M. M. Homayounpour, "Gender aware deep Boltzmann Machines for phone recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–5.

[27] Z. Huang, Q. Mao, Y. Zhan, and M. Dong, "Speech emotion recognition using CNN," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 801–804.

[28] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—Introducing the munich open-source emotion and affect recognition toolkit," in *Proc. Affect. Comput. Intell. Interact. Workshops*, 2009, pp. 1–6.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[31] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 4, Jul. 2003, pp. 2758–2763.

[32] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 1766–1770.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–10.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[35] K. He, S. Ren, J. Sun, and X. Zhang, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: https://arxiv.org/abs/1512.03385

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780,1997.
[37] Y. Zhang, G. Chen, K. Yao, S. Khudanpur, J. Glass, and D. Yu, "Highway long short-term memory RNNs for distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jan. 2016, pp. 5755–5759.
[38] J. G. Zilly, R. K. Srivastava, J. Koutník, and J. Schmidhuber, "Recurrent highway networks," 2016, *arXiv:1607.03474*. [Online]. Available: https://arxiv.org/abs/1607.03474
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.
[40] Y. Wen, Z. Li, Y. Qiao, and K. Zhang, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, 2016, pp. 499–515.
[41] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.
[42] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467
[43] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, vol. 5, 2005, pp. 1517–1520.
[44] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Speech Signal Process. (ICASSP)*, vol. 32, May 2013, pp. 3687–3691.
[45] S. Latif, M. Usman, J. Qadir, and A. Qayyum, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, 2018, pp. 88–93.
[46] Y. Huang, A. Wu, G. Zhang, and Y. Li, "Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition," *IET Signal Process.*, vol. 9, no. 4, pp. 341–348, 2015.
[47] P. Yenigalla and A. Kumar, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 3688–3692.

HAO MENG was a Visiting Scholar with the Robotics Laboratory, Dalhousie University, Canada, from 2006 to 2007. He is currently a Professor with the School of Automation, Institute of Robotics and Intelligent Control, Harbin Engineering University. He participated in many scientific research projects, such as the International Cooperation Project of the Ministry of Science and Technology, the National Defense Science Foundation. At present, as the Project Leader, he has presided over two military horizontal projects and one municipal science and technology innovation talent special fund project. His research interests include advanced control theory and applications, robotics and intelligent control, machine vision inspection technology, and underwater high-speed vehicle motion control technology. He also participated in several international academic conferences.

TIANHAO YAN received the bachelor's degree from the Northeast Forestry University, Harbin, China, in 2016, and he makes continuous academic programme that involves both postgraduate and doctoral study from Harbin Engineering University, Harbin. He is currently pursuing the Ph.D. degree. He is currently with the Institute of Robotics and Intelligent Control. His research interests include spoken signal processing, pattern recognition, machine learning, and affective computing.

FEI YUAN is currently pursuing the Ph.D. degree with the School of Automation, Institute of Robotics and Intelligent Control, Harbin Engineering University. Her major is in control science and engineering. She helped her teacher to do several projects, mainly using deep neutral networks. Her main research interests include deep neural networks, image processing, and expression recognition.

HONGWEI WEI is currently pursuing the M.D. degree with the College of Automation, Institute of Robotics and Intelligent Control, Harbin Engineering University. His major is control engineering. He is mainly involved in the research of brain–computer interface technology, deep learning theory, data structure budget algorithm, and finite element analysis. At present, he is mainly involved in the development of a mechanical modeling and simulation systems.