A Method upon Deep Learning for Speech Emotion Recognition

Nhat Truong PHAM[1], *Duc Ngoc Minh DANG*[1,3], *and Sy Dzung NGUYEN*[2,1,*]

[1]Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[2]Division of Computational Mechatronics, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[3]School of Graduate Studies, Ton Duc Thang University, Ho Chi Minh City, Vietnam

*Corresponding Author: Sy Dzung NGUYEN (email: nguyensydung@tdtu.edu.vn)

(Received: 9-Dec-2020; accepted: 7-Jan-2021; published: 10-Jan-2021)
DOI: http://dx.doi.org/10.25073/jaec.202044.311

Abstract. *Feature extraction and emotional classification are significant roles in speech emotion recognition. It is hard to extract and select the optimal features, researchers can not be sure what the features should be. With deep learning approaches, features could be extracted by using hierarchical abstraction layers, but it requires high computational resources and a large number of data. In this article, we choose static, differential, and acceleration coefficients of log Mel-spectrogram as inputs for the deep learning model. To avoid performance degradation, we also add a skip connection with dilated convolution network integration. All representatives are fed into a self-attention mechanism with bidirectional recurrent neural networks to learn long-term global features and exploit context for each time step. Finally, we investigate contrastive-center loss with softmax loss as loss function to improve the accuracy of the emotion recognition. For validating robustness and effectiveness, we tested the proposed method on the Emo-DB and ERC2019 datasets. Experimental results show that the performance of the proposed method is strongly comparable with the existing state-of-the-art methods on the Emo-DB and ERC2019 with 88% and 67%, respectively.*

Keywords

Speech emotion recognition, 3-D log Mel-spectrogram, contrastive-center loss, attention mechanism.

1. Introduction

Speech emotion recognition (SER) has an important role in Human-Computer Interaction (HCI) and also the most significant one in human communication. It has widely applied in health, education, robotics, and customer service systems. Yoon *et al.* [1] proposed the SER agent for mobile communication service. Huahu *et al.* [2] integrated the SER model into an intelligent household robot platform. Cen *et al.* [3] explored emotional recognition of continuous speech and developed a real-time SER system that can be applied to an online learning system.

There are two important roles in the SER system: (1) feature extraction that extracts the features from raw audio/speech data, and (2) emotional classification that decides the emotional state of speech. Feature extraction and selection are some of the key points in the SER systems, but nobody is quite sure what the features should be. Researchers have been used a variety of feature sets, such as Mel-frequency cepstral coefficients (MFCC) [4–7], Linear predictive coding (LPC) [7], signal energy, pitch, and zero-crossing rate [8–11]. The best approach is using deep learning as it will extract features into hierarchical abstraction layers. With the development of deep learning and neural networks, convolutional neural network (CNN) can perform better than traditional techniques in SER challenge [12–14]. Zhao *et al.* [15] found that 1-D CNN with long short-term memory (LSTM) and 2-D CNN with LSTM could explore the local and global features from raw audio/speech and the log Mel-spectrogram, respectively. Chan *et al.* [16] proved that the 2-D CNN is better than 1-D CNN without more data, and both time and frequency domains are of the same importance. To compare 2-D and 3-D convolution, researchers have tried to explore 3-D inputs for CNN to extract more features for the SER [17–19]. Recently, some studies have investigated attention mechanism to exploit the context for each time step or select emotional relevant frames for the SER [17, 18, 20–23].

For classifying the emotional states, a lot of classifier schemes have been used for the SER, such as hidden Markov model (HMM), Gaussian mixture model (GMM), support vector machine (SVM), artificial neural network (ANN), k-nearest neighbors (k-NN) and the others. HMM classifier has been applied widely in speech applications and emotional classification. Schuller *et al.* [24] used continuous HMM for the SER, Li *et al.* [25] investigated a hybrid deep neural network (DNN) HMM with discriminative pre-training for the SER, and Nwe *et al.* [4] proposed a method using short time log frequency power coefficients (LFPC) feature and classify the emotional states by a discrete HMM. As a special continuous HMM, GMM has also used to classify the emotional states from speech with global features extracted from training utterances. Tashev *et al.* [26] combined the GMM-based with DNN to extract both low-level and high-level features. Navyasri *et al.* [27] employed the GMM to classify the emotional states from speech features extracted by MFCC, spectral centroid, spectral skewness, and spectral pitch chroma. Shahin *et al.* [28] proposed a novel hybrid sequential GMM-DNN based classifier that gave significantly better accuracy than the SVM and multiplayer perceptron (MLP) classifiers. Lanjewar *et al.* [29] investigated and compared the GMM and k-NN classifiers to recognize six emotional states from speech features extracted by the MFCC, wavelet, and the pitch of vocal traces. Other researchers have optimized loss functions to train a state-of-the-art DNN for the SER. Tripathi and Zhu proposed a Focal loss to improve the accuracy of the emotion recognition system [30, 31]. Meng and Dai proposed a novel approach to discriminate emotional states from speech features by combining center loss with softmax loss as loss function [18, 32].

This research is motivated by previous works. Meng *et al.* [18] proposed a novel architecture ADRNN (dilated CNN with residual block and bidirectional long short-term memory (Bi-LSTM) based on the attention mechanism) which applied the dilated CNN to extract the features from the 3-D log Mel-spectrogram and combined the softmax loss with the center loss to improve the accuracy of emotion recognition. The ADRNN outperforms Chen's ACRNN (3-D attention-based convolutional recurrent neural networks) architecture with the softmax loss in [17] and the center loss in [32]. However, due to the weakness of the center loss, Qi *et al.* [33] proved that the contrastive-center loss outperforms the center loss for deep neural networks. Therefore, the combination of the contrastive-center loss with the softmax loss is investigated to improve the accuracy of emotion recognition.

In this article, we choose the 3-D static, differential, and acceleration coefficients of the log Mel-spectrogram extracted from the raw signal as inputs for the proposed model. Then all features are fed into an architecture AD-CRNN (attention-based dilated convolution and bidirectional recurrent neural networks) to extract high-level features. Finally, we use the contrastive-center loss with softmax loss to classify the emotional states. Furthermore, we also adopt a trick of dropout and batch normalization (BN) to normalize the features and improve the performance of deep neural network because of avoiding vanishing gradient problem in the training process. Our proposed method tested on the benchmark Emo-DB and validated on the ERC2019. Our contributions in this research are summarized below:

- We utilize the ADCRNN to extract spatial local features, learn the sequential global features and exploit the context for each time step from spectrogram-based inputs included static, differential, and acceleration coefficients.
- We also validate and compare the proposed loss and the center loss with our ADCRNN and ACRNN [17], respectively.
- Experimental results show that the proposed method outperforms the existing state-of-the-art methods on the Emo-dB and ERC2019 by 88% and 67%, respectively.

This article is organized as follows: Section 2 describes the methodology in detail, Section 3 shows the experimental results and comparison, and Section 4 describes the conclusion.

2. Methodology

2.1. Generate 3-D log Mel-spectrogram

The 3-D log Mel-spectrograms (static, differential, and acceleration coefficients) are used as the inputs for our proposed method. Given a raw audio/speech signal, we compute the log Mel-filterbank energy features under the sample rate of 16 kHz, the number of 40 filters in the filterbank, and the FFT size is chosen to 512. Besides, we choose the length of the analysis window of 0.025 sec and the step between successive windows of 0.01 sec. Furthermore, to obtain the 40 Mel-filterbank, we also choose the lowest band edge of Mel filters, the highest band edge of Mel filters, and the pre-emphasis filer with preempt as coefficients of 300 Hz, 8,000 Hz, and 0.97, respectively.

The static coefficient of the log Mel-spectrogram is obtained following six steps as below:

- Firstly, the Mel scale frequency analysis [34] was computed as below:

$$M(freq) = 1125 \times \ln\left(1 + \frac{freq}{700}\right), \quad (1)$$

where $M(freq)$ is the Mel scale converted from the frequency $freq$. The lowest and highest frequencies were converted to 401.25 Mels and 2,835.00 Mels, respectively.

- Secondly, we need at least 42 points to get the 40 filterbanks. So, we added 40 points spaced linearly between the lowest and highest Mel points.
- Thirdly, we inverted the Mel scale back to frequency as in Eq. 2 so that there were 42 frequency points between 300 Hz and 8,000 Hz as mentioned before:

$$M^{-1}(m) = 700 \times \left[\exp\left(\frac{m}{1125}\right) - 1\right], \quad (2)$$

where $M^{-1}(m)$ is the frequency inverted from the Mel scale m.

- Next, we had to round those frequency points to the nearest FFT bin numbers because we could not have the exact frequency resolution as calculated above. The FFT bin numbers can be computed as follows:

$$f(n) = floor\left[\left(nFFT + 1\right) \times \frac{h(n)}{sp}\right], \quad (3)$$

where the $nFFT$ is the FFT size, the sp is the sample rate, the $h(n)$ is the frequency points in Hert, and the $floor$ is the function that gives the greatest integer output less than or equal to the real number input.

- Then, the filterbanks can be defined as follows:

$$H_m(k) = \begin{cases} 0 & k < f(m\text{-}1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m\text{-}1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases}, \quad (4)$$

where k is the point of the FFT bin numbers, m is the number of filterbanks we wanted, and $f()$ is a list of $m+2$ Mel points. After that, we passed the power spectrum calculated by short-time Fourier transform (STFT) through the Mel filterbanks to get the Mel-spectrogram.

- Finally, we computed the log Mel-spectrogram ($logM$) by taking the logarithm of the Mel-spectrogram, which is the first dimension of the inputs.

After computing the static coefficient of the log Mel-spectrogram, we obtained the differential (second dimension) and acceleration (third dimension) coefficients of the inputs as below:

$$\Delta(M) = \frac{\sum_{n=1}^{N} n\big(logM_{t+n} - logM_{t-n}\big)}{2\sum_{n=1}^{N} n^2}, \qquad (5)$$

where Δ(M) is the differential coefficient (deltas) computed by taking the time derivative t of static coefficient from $logM_{t+N}$ to $logM_{t-N}$, and N is set to 2 in this equation. The acceleration coefficient (delta-deltas) is computed likely in Eq. 2, but the input from the deltas (differential) coefficient.

We combined the 3-D log Mel-spectrogram features $X \in \Re^{t,f,c}$ as the inputs of the proposed model. In which, t is set to 0.3 sec as the chunk size of the audio/speech duration, f is set to the number of 40 filters in the filterbank, and c is set to 3 channels or dimensions represented the static, differential, and acceleration coefficients, respectively. The 3-D log Mel-spectrogram corresponding to the wave data from the audio/speech signal is shown in Fig. 1.

2.2. Attention-based dilated convolution and bidirectional recurrent neural networks

In this section, we utilize the ADCRNN by combining dilated convolution neural network (DCNN) and Bi-LSTM with the attention mechanism to extract the features from the 3-D log Mel-spectrogram for SER. Firstly, we perform the CNN and the DCNN to extract spatial local features from the 3-D log Mel-spectrogram inputs. Then, we fit all feature maps into Bi-LSTM to learn long-term sequential global features and exploit the context for each time step by attention mechanism. Next, all features are passed to a fully connected network layer to obtain high-level features. Finally, we integrate the contrastive-center loss with the softmax loss as a loss function to classify the emotional states. Our proposed method is described as in Fig. 2 and the ADCRNN architecture is presented in detail as in Tab. 1, where ES is the number of emotional states.

Tab. 1: The layers and parameters of the proposed model.

Layer	Kernel/ Size	Output	Stride
Input	—	300×40×3	—
CNN	3×3	298×38×128	1×1
Max-Pool	2×4	149×9×128	2×4
DCNN1	3×3	149×9×256	—
DCNN2	3×3	149×9×256	—
DCNN3	3×3	149×9×256	—
DCNN4	3×3	149×9×256	—
Skip-Net	—	149×9×256	—
Linear	—	512	—
Bi-LSTM	256	—	—
Attention	—	512	—
FC1	—	64	—
FC2	—	ES	—

1) Dilated convolution neural network

Dilated convolutions are investigated by Yu *et al.* [35] to design a new convolutional network module for dense prediction. Based on the dilated convolutions, the receptive field exponentially expands without loss of resolution or coverage while the number of parameters grows linearly. The dilated convolutions with the height P and the width Q is defined as follows:

$$y(h,\ w) = \sum_{i=1}^{P}\sum_{j=1}^{Q} x(h + dr * i,\ w + dr * j)F(i,\ j) \qquad (6)$$

where $y(h,\ w)$ is the receptive field output from the input $x(h,\ w)$ when applied a dilation rate

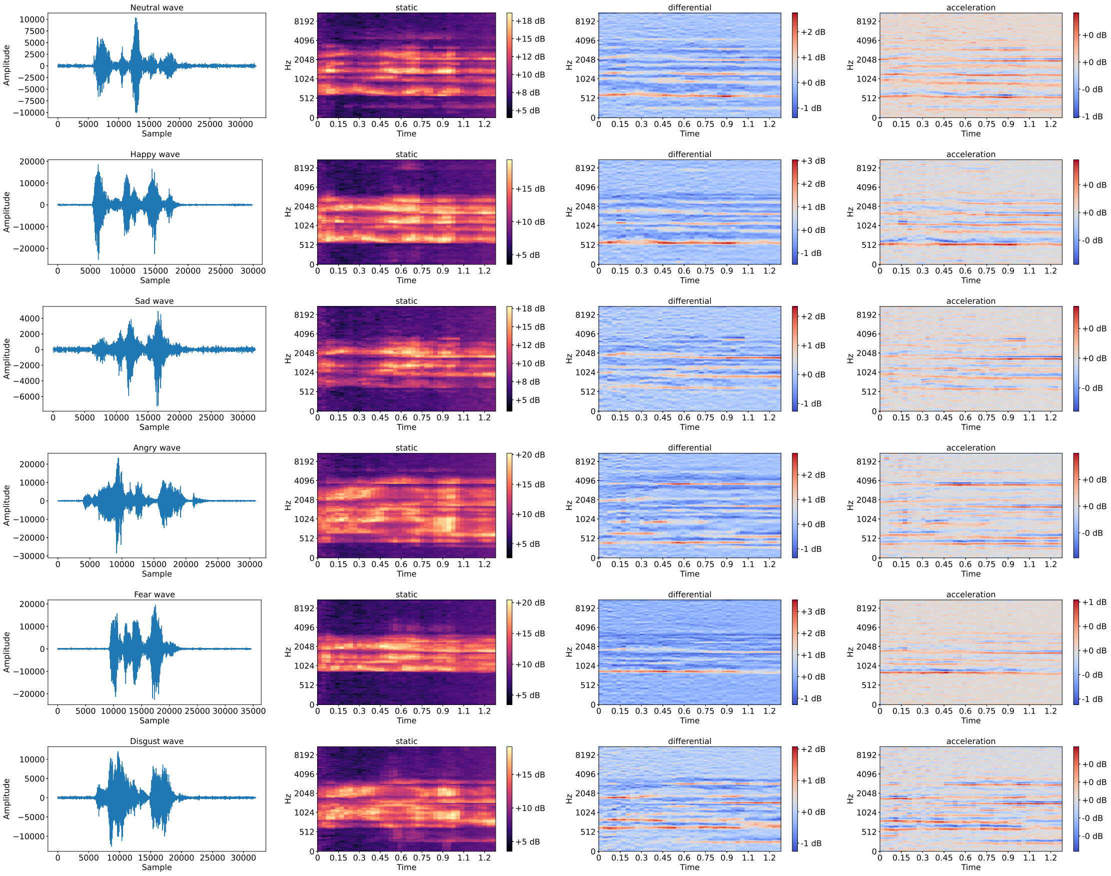

Fig. 1: Visualization of the 3-D log Mel-spectrogram in the ERC2019 dataset.

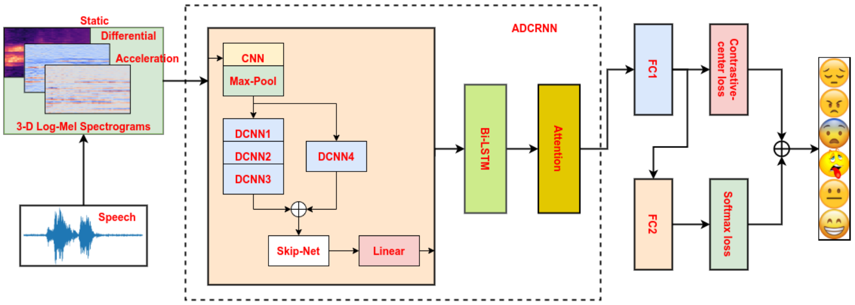

Fig. 2: The proposed method architecture.

dr to the filter $F(i,\ j)$. Therefore, the original convolution is the dilated one with the $dr = 1$. The illustration of 3×3 kernel size with the $dr = 2$ is shown in Fig. 3.

In this subsection, we design three DCNN layers after performing by one original CNN and Max-Pool layer. We utilize the CNN layer with 3×3 kernel size and stride at 1. To down-sample representation, the Max-Pool layer with 2×4 kernel size and 2×4 stride is added. Each DCNN layer has 3×3 kernel size, the stride of 1, and the dilation rate of 2. Furthermore, we add a skip

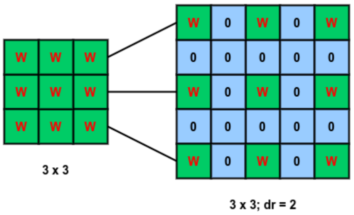

Fig. 3: Illustration of the dilated convolution with the dr = 2.

connection with dilated convolution (Skip-Net) to avoid performance degradation. The padding is set to the "VALID" for both the CNN and Max-Pool layers while it is the "SAME" in all the DCNN layers. Instead of using ReLU, we use the Leaky ReLU activation function to solve the vanishing gradient descent and ensure the non-linearity of deep neural networks when the input value is negative. The Leaky ReLU activation function $g(x)$ can produce a non-zero output for a negative input as below:

$$g(x) = \begin{cases} x, & if\ x \geq 0 \\ \alpha x, & if\ x < 0 \end{cases}, \tag{7}$$

where α is a constant in range (0, 1). We choose the $\alpha = 0.01$ for the Leaky ReLU activation function in this experiment.

2) Recurrent neural networks

Recurrent neural networks (RNN) has solved many sequential problems by learning historical features. It has taken more advantage of natural language processing, video processing, and time series prediction. However, it has a limitation with long-term dependencies. Hochreiter *et al.* [36] proposed a novel LSTM method to deal with complex, artificial long-term tasks. In this study, the effectiveness of BiLSTM proposed by Schuster *et al.* [37] for the SER is investigated. The BiLSTM can learn the sequential features in both forward and backward directions by splitting the neurons of regular RNN. Therefore, the BiLSTM can use the input information from the future and past of the current time step for prediction. Each LSTM cell is updated as in Eqs. 8-12:

$$i_t = \sigma_g\left(W_i z_t + U_i h_{t-1} + b_i\right), \tag{8}$$

$$f_t = \sigma_g\left(W_f z_t + U_f h_{t-1} + b_f\right), \tag{9}$$

$$o_t = \sigma_g\left(W_o z_t + U_o h_{t-1} + b_o\right), \tag{10}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tau_g\left(W_c z_t + U_c h_{t-1} + b_c\right), \tag{11}$$

$$h_t = o_t \cdot \tau_g(c_t), \tag{12}$$

where σ_g and τ_g denote the sigmoid and tanh activation functions, the $(\cdot)$ operator is the element-wise product. The i_t, f_t, o_t, c_t, z_t and h_t represent the input gate, forget gate, output gate, cell state with a self-recurrent, input vector, and hidden state at the time step t, respectively. All weight matrices are set to W, U and corresponding bias vectors are set to b.

In this implementation, we choose the cell units of 256 for each LSTM direction. All features obtained in the framework of DCNN are fed into the BiLSTM to learn the sequential global features.

3) Attention mechanism

After performing the BiLSTM to learn sequential features, we add the attention mechanism to exploit the context of each time step. The attention-based model has been effectively used in almost sequence-to-sequence and the SER tasks [17, 18, 20–23, 38, 39]. In the SER task, not all emotional content from speech signal contributes equally to represent the emotional states. Hence, in this research, the attention-based method is constructed to concentrate more on the specific part of the spectrogram-based features that involve mostly the emotional classification task.

In this experiment, the attention structure of BiLSTM at the time step t is defined as below:

$$Att = \sum_{t=1}^{T} \beta_t h_t, \tag{13}$$

where the Att is the output of the attention layer, the β_t is the attention weight computed by the softmax function as follows:

$$\beta_t = \frac{\exp\left(W \cdot h_t\right)}{\sum\limits_{j=1}^{T} \exp\left(W \cdot h_t\right)}, \tag{14}$$

where the $(\cdot)$ operator is the element-wise product, the W is a trainable parameter, the h_t is the hidden state at time step t from the BiLSTM that is $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. Next, we integrate the contrastive-center loss with the softmax loss to classify the higher-level representation from the attention output.

2.3. Loss function for classification

For most of the classification tasks, people commonly used softmax cross-entropy with logit for multi-class classification. In this article, we not only want to separate the emotional states but also discriminate against them. To investigate that we integrated the contrastive-center loss with the softmax loss as the loss function to update weights during training process. The contrastive-center loss [33] performs better than the center loss for deep neural networks and classification problems.

Due to the weakness of the center loss, the contrastive-center loss has been proposed to discriminate the intra-class compactness and inter-class separability as follows:

$$\mathscr{L}_C = \frac{1}{2}\sum_{s=1}^{S} \frac{||x_i - C_{y_i}||_2^2}{\left(\sum\limits_{\substack{e=1,\\ e\neq y_i}}^{E} ||x_i - C_e||_2^2\right) + \lambda}, \tag{15}$$

where S denotes the number of training samples in a mini-batch , $||x_i - C_{y_i}||_2^2$ denotes the distances between the training samples and their corresponding class centers, $||x_i - C_m||_2^2$ denotes the distances between the training samples and their non-corresponding class centers, E denotes the number of classes, and the constant λ is set to 1 to ensure that the denominator not equal zero. Finally, we add the softmax loss with the contrastive-center loss to obtain final loss function to update weights during training progress as below:

$$\mathscr{L}_{total} = \mathscr{L}_C + \mathscr{L}_{softmax}. \tag{16}$$

3. Experimental results

3.1. Datasets

To evaluate the robustness and effectiveness of our proposed method with the 3-D log Mel-spectrogram, we used the Berlin Database for Emotional Speech (Emo-DB) and the Emotion Recognition Challenge 2019 dataset (ERC2019).

1) Emo-DB

The Berlin Database of Emotional Speech [40] is recorded with 44.1 kHz and downed sampling rate to 16 kHz simultaneously. It contains 535 recorded files by 5 males and 5 females. They spoke the sentences with several different emotional states, such as anger, sadness, happiness, neutral, disgust, fear, and boredom. We presented in detail the distribution of the emotional states in the Emo-DB dataset in the pie chart as Fig. 4.

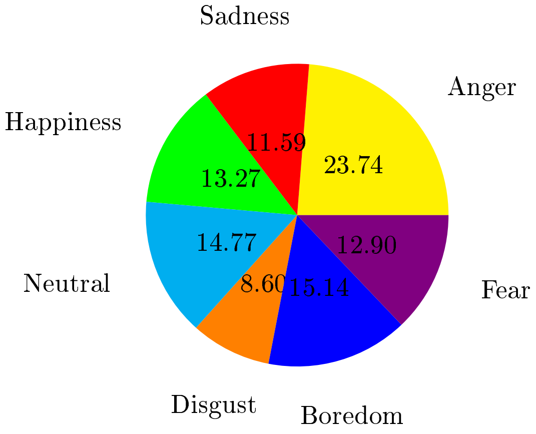

Fig. 4: The distribution of emotional states in the Emo-DB dataset.

2) ERC2019 dataset

ERC2019 dataset is the dataset in Emotion Recognition Challenge 2019 held by Robotics-IoT club, Vietnam National University Ho Chi Minh City - University of Science, which is a subset of Crema-D in [41]. It contains 5,230 recorded files that were spoken from a selection of 12 sentences in six different emotional states:

anger, sadness, happiness, neutral, disgust, and fear. All actors included 48 males and 43 females between the ages of 20 and 74 coming from a variety of races and ethnicities, such as Asian, African, American, Hispanic, Caucasian, and Unspecified. We presented in detail the distribution of the emotional states in the ERC2019 dataset in the pie chart as Fig. 5.

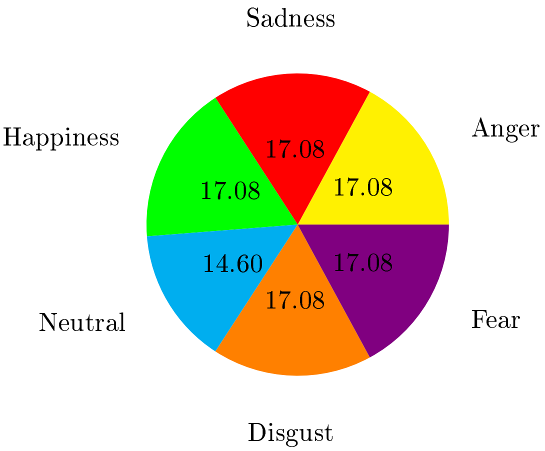

Fig. 5: The distribution of emotional states in the ERC2019 dataset.

3.2. Experimental setup

The system was used for running the experiment is built in Intel CORE i5 8th Gen with NVIDIA Graphics Card 1080Ti. We used TensorFlow deep learning framework [42] to implement the whole model.

For the feature extraction, 'python speech features' framework [43] is used to compute and extract the 3-D log Mel-spectrogram as follows:

- Compute the static coefficient of the raw audio/speech signal: The window length, the overlap between windows, FFT size, and pre-emphasis coefficient are set by default at 0.025 sec, 0.01 sec, 512, and 0.97, respectively. The sampling rate, the lowest band edge of Mel filters, and the highest band edge of Mel filters are set to 16 kHz, 300 Hz, and 8,000 Hz, respectively.
- Compute the differential coefficient is computed by taking the time derivative of the static coefficient.
- Compute the acceleration coefficient is computed by taking the time derivative of the differential coefficient.

For parameter optimization, we set the batch-size 32 to compatible with the limited memory. Then, we chose Adam optimizer with a learning rate of e^{-4}. Besides, we also integrated the contrastive-center loss with standard softmax loss for the proposed model to improve the classification performance. Furthermore, to get the best results, we also employed k-fold cross-validation with $k = 5$ to get the mean and standard deviation accuracy.

3.3. Results

1) Experiment on the Emo-DB

Tab. 2: The comparison of accuracy on the Emo-DB.

Model	Loss function	Accuracy
ACRNN [17]	Center [18]	0.83 ± 0.03
	Proposed	0.86 ± 0.01
ADCRNN	Center [18]	0.86 ± 0.05
	Proposed	0.88 ± 0.03

In Tab. 2, our proposed loss function achieved better accuracy than the center loss [18] in both ACRNN [17] and our ADCRNN architectures. Our proposed loss function reaches 0.86 ± 0.01 and 0.88 ± 0.03, respectively. The results in Tab. 2 also prove that our ADCRNN architecture with both proposed and center losses were higher accuracy than ADRNN with the center loss [18] by 0.88 ± 0.03, 0.86 ± 0.05, and 0.85 ± 0.02, respectively.

The confusion matrix shows the model predicted results and the ground truth labels for each emotional state in Fig. 6. The A, B, D, F, H, S, and N labels represented anger, boredom, disgust, fear, happiness, sadness, and neutral emotional states, respectively. As the confusion matrix is shown in Fig. 6, the proposed method was better than the

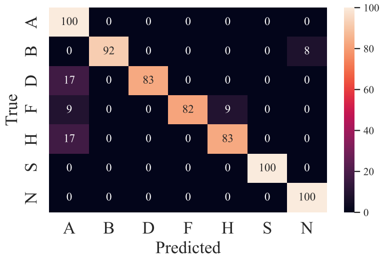

Fig. 6: The confusion matrix of proposed method on the Emo-DB.

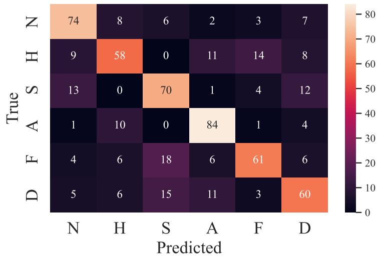

Fig. 7: The confusion matrix of proposed method on the ERC2019.

previous methods in [17, 18]. In terms of comparison with [18], our proposed method achieved 3%, 27%, and 11% better performance in the fear, happy, and bored emotional states, respectively. Besides, our proposed method was more accurate 12%, 36%, 1%, 17%, and 7% than [17] in the angry, happy, fear, bored, and neutral emotional states, respectively.

2) Experiment on the ERC2019

Tab. 3: The comparison of accuracy on the ERC2019.

Model	Loss function	Accuracy
ACRNN [17]	Center [18]	0.63 ± 0.01
	Proposed	0.63 ± 0.00
ADCRNN	Center [18]	0.65 ± 0.01
	Proposed	0.67 ± 0.01

In Tab. 3, our proposed loss function achieved better accuracy than the center loss [18] with our ADCRNN architectures. Our proposed loss function reaches 0.86 ± 0.01 with our ADCRNN architecture while the center loss reaches 0.65 ± 0.01. Besides, the accuracy is likely the same in both the center and proposed loss functions with ACRNN architecture. The results in Tab. 3 also prove that our ADCRNN architecture with the proposed loss function is better than the others on the ERC2019.

The confusion matrix shows the model predicted results and the ground truth labels for each emotional state in Fig. 7. The N, H, S, A, F, and D labels represented neutral, happiness, sadness, anger, fear, and disgust emotional states, respectively.

4. Conclusions

In this article, we proposed an architecture ADCRNN with the contrastive-center loss for SER systems. The model not only learned spatial local features by DCNN, but also learned long-term global features and exploited the context for each time step by bidirectional RNN with attention mechanism from the 3-D log Mel-spectrogram (static, differential, and acceleration coefficients) of raw speech/audio signal. The DCNN with a residual block that consisted of three dilated convolution layers with one Leaky ReLU activation function in each layer and the skip connection with one dilated convolution layer. Then, we employed the contrastive-center loss together with softmax loss to improve performance classification.

The proposed method was tested on the benchmark Emo-DB and also validated on the ERC2019 which was used in the Emotion Recognition challenge. The experimental results show that the proposed model with the contrastive-center loss not only extracted the spatial features of the 3-D log Mel-spectrogram and learn the long-term global features but also discriminated the emotional states instead of separating them. Our proposed method achieved better accuracy than state-of-the-art methods by 88% and 67% on the Emo-DB and ERC2019, respectively.

Acknowledgement

The authors would like to thank the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.01-2019.328.

Abbreviations

ACRNN	3-D attention-based convolutional recurrent neural networks
ADCRNN	Attention-based dilated convolution and bidirectional recurrent neural networks
ADRNN	Dilated CNN with residual block and Bi-LSTM based on the attention mechanism
ANN	Artificial neural networks
Bi-LSTM	Bidirectional long short-term memory
BN	Batch normalization
CNN	Convolutional neural network
CRGNN	Convolutional recurrent global neural network
DCNN	Dilated convolution neural network
Emo-DB	Berlin database of emotional speech
ERC2019	Emotion recognition challenge 2019
FFT	Fast Fourier Transform
GMM	Gaussian mixture model
HCI	Human-Computer Interaction
HMM	Hidden Markov model
KNN	K-nearest neighbors
LeReLU	Leaky rectified linear unit
LSTM	Long short-term memory
LPC	Linear predictive coding
MFCC	Mel-frequency cepstral coefficients
RNN	Recurrent Neural Networks
SER	Speech Emotion Recognition
STFT	Short-time Fourier Transform
SVM	Support Vector Machine

References

[1] Yoon, W.J., Cho, Y.H., & Park, K.S. (2007). A study of speech emotion recognition and its application to mobile services. In *International Conference on Ubiquitous Intelligence and Computing*, Springer, 758–766.

[2] Huahu, X., Jue, G., & Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, IEEE, 537–541.

[3] Cen, L., Wu, F., Yu, Z.L., & Hu, F. (2016). A real-time speech emotion recognition system and its application in online learning. In *Emotions, technology, design, and learning*, Elsevier, 27–46.

[4] Nwe, T.L., Foo, S.W., & De Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. *Speech communication*, *41*(4), 603–623.

[5] El Ayadi, M.M., Kamel, M.S., & Karray, F. (2007). Speech emotion recognition using Gaussian mixture vector autoregressive models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, IEEE, IV–957.

[6] Albornoz, E.M., Milone, D.H., & Rufiner, H.L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, *25*(3), 556–570.

[7] Yeh, J.H., Pao, T.L., Lin, C.Y., Tsai, Y.W., & Chen, Y.T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior*, *27*(5), 1545–1552.

[8] Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, *49*(10-11), 787–800.

[9] Yang, B. & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal processing*, *90*(5), 1415–1423.

[10] Lee, C.C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*(9-10), 1162–1171.

[11] Chen, L., Mao, X., Xue, Y., & Cheng, L.L. (2012). Speech emotion recognition: Features and classification models. *Digital signal processing*, *22*(6), 1154–1160.

[12] Zhang, S., Zhang, S., Huang, T., & Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, *20*(6), 1576–1590.

[13] Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, *16*(8), 2203–2213.

[14] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 5200–5204.

[15] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323.

[16] Chan, W. & Lane, I. (2015). Deep convolutional neural networks for acoustic modeling in low resource languages. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2056–2060.

[17] Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440–1444.

[18] Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access*, *7*, 125868–125881.

[19] Zayene, B., Jlassi, C., & Arous, N. (2020). 3D Convolutional Recurrent Global Neural Network for Speech Emotion Recognition. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–5.

[20] Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2227–2231.

[21] Neumann, M. & Vu, N.T. (2017). Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. In *Proc. Interspeech 2017*, 1263–1267.

[22] Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019). Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access*, *7*, 97515–97525.

[23] Zhang, Z., Wu, B., & Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 6705–6709.

[24] Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, IEEE, II–1.

[25] Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., & Sahli, H. (2013). Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In *2013 Humaine association conference on*

affective computing and intelligent interaction, IEEE, 312–317.

[26] Tashev, I.J., Wang, Z.Q., & Godin, K. (2017). Speech emotion recognition based on gaussian mixture models and deep neural networks. In *2017 Information Theory and Applications Workshop (ITA)*, IEEE, 1–4.

[27] Navyasri, M., RajeswarRao, R., DaveeduRaju, A., & Ramakrishnamurthy, M. (2017). Robust features for emotion recognition from speech by using Gaussian mixture model classification. In *International Conference on Information and Communication Technology for Intelligent Systems*, Springer, 437–444.

[28] Shahin, I., Nassif, A.B., & Hamsa, S. (2019). Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access*, *7*, 26777–26787.

[29] Lanjewar, R.B., Mathurkar, S., & Patel, N. (2015). Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques. *Procedia computer science*, *49*, 50–57.

[30] Tripathi, S., Kumar, A., Ramesh, A., Singh, C., & Yenigalla, P. (2019). Focal Loss based Residual Convolutional Neural Network for Speech Emotion Recognition. *arXiv preprint arXiv:190605682.*

[31] Zhu, Z., Dai, W., Hu, Y., & Li, J. (2020). Speech Emotion Recognition Model Based on Bi-GRU and Focal Loss. *Pattern Recognition Letters.*

[32] Dai, D., Wu, Z., Li, R., Wu, X., Jia, J., & Meng, H. (2019). Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 7405–7409.

[33] Qi, C. & Su, F. (2017). Contrastive-center loss for deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, 2851–2855.

[34] Huang, X., Acero, A., Hon, H.W., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice hall PTR.

[35] Yu, F. & Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. In Bengio, Y. & LeCun, Y. (editors), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.*

[36] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

[37] Schuster, M. & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

[38] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., & Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (editors), *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc., 5998–6008.

[40] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., & Weiss, B. (2005). A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology.*

[41] Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390.

[42] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. *et al.* (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603044467.*

[43] Lyons, J., Wang, D.Y.B., Gianluca, Shteingart, H., Mavrinac, E., Gaurkar, Y., Watcharawisetkul, W., Birch, S., Zhihe, L., Hölzl, J., Lesinskis, J., Almér, H., Lord, C., & Stark, A. (2020), jameslyons/python_speech_features: release v0.6.1.

About Authors

Nhat Truong PHAM received a B.Eng. degree in Electronics and Telecommunication Engineering, Ton Duc Thang University (TDTU), Vietnam, in 2019. He is currently pursuing an M.Eng. degree in Automation and Control Engineering. He is a research assistant in the Division of Computational Mechatronics, Institute for Computational Science, TDTU, Vietnam. His research interests include Artificial Intelligence, Deep Learning, Computer Vision, Audio/Speech Processing, Robotics, and Intelligent Computation.

Duc Ngoc Minh DANG received the B.Eng and M.Eng degrees in Telecommunications Engineering from Ho Chi Minh City University of Technology, Vietnam, in 2005 and 2007, respectively, and a Ph.D. degree in Computer Engineering from Kyung Hee University, Korea, in 2014. His research interests include the MAC protocols in Wireless Ad hoc Networks and Vehicular Ad hoc Networks. From 2005 to 2008, he was a Senior Telecom Engineer with TMA Solutions company, Vietnam. He joined Ton Duc Thang University, Vietnam, where he worked as Head of Electronics and Telecommunications Engineering Department and Laboratory Head from 2008 to 2011. From 2011 to 2014, he worked as a Ph.D. assistant and a post-doc researcher at Kyung Hee University, Korea. Currently, he has worked as Vice Dean, School of Graduate Studies, Ton Duc Thang University, Vietnam.

Sy Dzung NGUYEN received an M.E. degree in Manufacturing Engineering from Ho Chi Minh City University of Technology (HCMUT) – Vietnam National University (VNU) in 2001, a Ph.D. degree in Applied Mechanics in 2011 from HCMUT - VNU. He was a postdoctoral fellow at Inha University, South Korea, 2011-2013, at Incheon National University, South Korea, 2015-2016. He is currently a Head of Division of Computational Mechatronics, Institute for Computational Science, Ton Duc Thang University, Vietnam. His research interests include Artificial Intelligence and its applications to NonLinear Adaptive Control, System Identification, Prediction, and Structure Damage Managing. Dr. Nguyen has been the main author of various ISI papers in these fields.