Research on Speech Emotion Recognition Based on SpatioTemporal Attention CNN Model

Yancong Zhou, *Member, IAENG,* Xiaoli Xie

***Abstract*—With the development of science and technology, Speech Emotion Recognition (SER) has gradually become a hot spot for researchers. In recent years, the research on SER has focused on the display of local features in speech, while ignoring the contextual features of speech. In order to solve this problem, this paper proposed a SpatioTemporal Attention Convolutional Neural Networks (STA-CNN) model, in which the CFE layer of contextual feature extraction was added to the model to ensure that the contextual feature information was fully utilized. In order to avoid information loss, the feature fusion FF layer that fused features before the input layer was added to form a richer and more comprehensive data representation. In addition, in order to further obtain the emotional information features of speech, the global feature extraction module layer GFE was added to obtain the global features. Finally, the STA-CNN model was experimentally verified on EMO-DB, IEMOCAP, RAVDESS and SAVEE datasets, and the accuracy of the model's sentiment classification was 80.23%, 71.96%, 76.39% and 80.09%, respectively, which was 0.23%, 3.07%, 0.77% and 4.46% higher than that of the benchmark models. The experimental results have fully demonstrated the high accuracy and strong generalization ability of the STA-CNN model.**

***Index Terms*—Speech emotion recognition, emotion features, features fusion, feature extraction**

I. Introduction

As a natural communication medium, voice signals contain not only communicative content, but also a lot of information about the speaker, such as age, gender, ethnicity, health status, emotions, and thoughts. Effectively processing and understanding the emotional state of the speaker, enhancing the ability of system analysis and assisting in SER are important and arduous tasks in human-computer interaction[1], which have become a hot research field[2]. Related research not only has important theoretical significance, but also shows broad prospects in practical application, which is a necessary prerequisite for realizing more emotional, friendly, and vivid interaction between humans and machines, and plays an important role in the public fields such as industry, healthcare, transportation, and education[3].

SER which focuses on the technical analysis is a pivotal research direction in the field of artificial intelligence. The identification of emotional information in speech. The significance of SER research is multifaceted. It enhances the machines' ability to comprehend users' emotional states, thereby facilitating a more personalized and empathetic interactive experience. As the base of affective computing, SER has spurred extensive research in this domain. By utilizing voice emotion recognition, the personnel can more precisely understand the emotions of customers or patients, thereby it can offer more personalized services. SER technology is instrumental in monitoring and analyzing emotional fluctuations in individuals, holding promise for the early detection and diagnosis of conditions such as depression and anxiety. In educational settings, by recognizing students' speech and emotions, educators can gain insight into their learning status and emotional needs, then adjust the teaching methods and content to improve educational quality. SER aids in comprehending and responding to others' emotions during social interactions, which is particularly advantageous for individuals with social disabilities. Emotion expression varies across cultural and linguistic contexts; SER research aids in elucidating these nuances, thereby fostering cross-cultural communication. In essence, the exploration of SER not only carries academic significance but also presents a broad spectrum of application potential, and plays an important role in promoting social development, scientific and technological progress.

Through in-depth research and exploration of SER, researchers can further improve the understanding and processing ability of human emotions, thereby promote the development of related technologies and applications. In recent years, researchers have emerged in an endless stream in the field of SER, but most of them focus on the basic features of speech and do not involve the deep speech features, and often ignore the contextual features of speech, resulting in the low accuracy of the emotion recognition model of speech and the difficulty of accurately capturing the subtle emotional changes in speech. These models perform poorly when dealing with complex emotions and multilingual environments, so limit their effectiveness in real-world applications. In order to improve the accuracy of SER, researchers have begun to introduce Deep Learning (DL) models to solve this problem. DL models can learn and extract emotional features in speech more efficiently, which significantly improves the accuracy and robustness of recognition.

The contributions are summarized as follows:

(1) In order to solve the problem of information loss in speech, the feature fusion FF layer was proposed to fuse the speech signal features before the input layer of the model, so as to avoid information loss and reduce the propagation of

Manuscript received December 20, 2024; revised April 12, 2025.

This work was supported in part by enterprise project under Grant G24075.

Yancong Zhou is a professor of Tianjin University of Commerce, Tianjin 300134, China (corresponding author to provide phone:022-26667577 fax:022-26667577, e-mail: zycong78@126.com).

Xiaoli Xie is a postgraduate student of Tianjin University of Commerce, Tianjin 300134, China(e-mail: 2165297808@qq.com).

errors, so that the model could adapt to more complex tasks.

(2) In order to solve the problem of lack of contextual features in speech, the CFE layer of contextual feature extraction was added., which helped the STA-CNN model better understand the contextual information, so as to improve the recognition rate of the model.

(3) In order to solve the problem of insufficient feature representation in speech, the GFE (Global feature extraction) layer was added to extract the global features in speech, and automatically focused on the most important feature parts for SER, so as to obtain more feature representations from speech, analyzed speech features more comprehensively, and made full use of the features in speech.

II. Related Work

A. Speech Features

In the field of SER, feature processing and research of speech are the key to SER[4]. In order to learn advanced features from emotional discourse and form hierarchical representations of speech, many DL architectures have been introduced into SER. SÖNMEZ et al.[5] conducted a detailed analysis of SER research on speech signals from the past to the present, and compared the results of classical classifiers and DL methods. Given that manual feature extraction often ignores advanced features derived from the lower-level features, DL has been introduced into SER to learn advanced features from emotional discourse, such as Convolutional Neural Networks (CNN), which have become excellent models in the field of SER.

Because CNN can automatically extract various features needed from raw data, more and more researchers are using CNN for SER feature extraction. Flower et al.[6] improved SER performance by combining one-dimensional CNN, prevented overfitting and enhancing its accuracy, and achieved significant improvements on EMO-DB, EMOVO, SAVEE, and RAVDESS datasets. The performance of SER systems depends on the features extracted from speech signals, but most studies overlook the contextual information representation of speech signals, leading to low accuracy in model recognition. To address this issue, Ahmed et al.[7] proposed the 1D-CNN-LSTM-GRU (one-Dimensional Convolutional Neural Networks-Long Short-Term Memory-Gate Recurrent Unit) model, which focused on extracting local and long-term global context representations of speech signals. The highest weighted average accuracy was achieved on five benchmark datasets, TESS, EMO-DB, RAVDESS, SAVEE, and CREMA-D. In addition, Manohar et al. [8] implemented a new SER using a hybrid DL model. Wang et al.[9] proficiently integrated various feature extraction techniques and achieved high accuracy on multiple datasets. Hyeon et al.[10] proposed a new method that utilized both SSL models and domain agnostic spectral features through feature fusion techniques. Lin et al.[11] recently proposed a block based DeepEmoCluster framework, which incorporated the concept of deep clustering as a new semi-supervised learning framework. Compared with traditional reconstruction based methods, this framework improved recognition performance. Zhang et al.[12] proposed a maximum average pooling capsule network model. In recent studies, Soltani et al.[13] used deep echo state networks to enhance neural information processing, significantly improving weighted and unweighted accuracy in RAVDESS, Emo DB, SAVEE, and TESS databases.

B. Feature extraction techniques

In addition, in the field of SER, researchers are committed to improving the representation methods for capturing emotional information. Naderi et al.[14] proposed an Attention-based Feature Fusion method based on attention, feature fusion and transfer learning. To overcome the problem of insufficient feature fusion, Chen et al.[15] proposed a multi-scale SER parallel network based on connection attention mechanism, which can utilize the temporal and spatial features of speech signals. Yang et al. [16] proposed a novel multi-feature method for SER, which effectively solved the overfitting problem by reducing the feature dimension. Lian et al.[17] proposed a new transfer subspace learning framework with feature selection capability to address the problem of lack of interpretability for selected features, which made it difficult to determine which acoustic features have corpus invariance. In recent years, there have been numerous studies and feature processing methods in the field of SER. Saleem et al.[18] proposed a 3D-CNN model to capture temporal and spatial features of speech, while Li et al.[19] proposed a Bi-A2CEmo framework that simultaneously addressed the bidirectional acoustic to speech conversion problem of SER. Wang et al. [20] proposed a new domain adaptation method that solved the problem of domain differences in speech data from different domains. Wang et al.[21] proposed an enhanced SER method using TF Mix, while Liu et al.[22] proposed a one-Dimensional Convolutional Neural Network (1D CNN) to obtain comprehensive speech features using a multi-scale multi-channel features extraction structure with global and local information. Yu et al.[23] proposed a SER method based on multidimensional features extraction and multi-scale features fusion, which solved the problem of capturing rich emotional information at different scales. Harby et al.[24] used 2D convolutional neural networks for speech preprocessing, and provided a visual representation of how the frequency content of audio signals changing over time. Mishra et al.[25] proposed a combined method of Variational Mode Decomposition (VMD) and Hilbert Transform (HT) techniques to improve the performance of emotion classification.

Although human feature selection ability is very powerful, some DL models cannot fully learn human attention mechanisms, so the model results can not meet expectations, resulting in significant limitations for DL models in the field of SER. Therefore, attention mechanisms are used to screen features and help DL models solve the problem of limited feature selection in the field of SER.

III. Methodology

A. Speech Emotion Recognition Model

We have designed a SER model STA-CNN, which includes four layers: FF (feature fusion) layer, LFE (local feature extraction) layer, CFE (Context feature extraction) layer, and GFE (Global feature extraction) layer. The FF layer can integrate features from different dimensions such as Spectrograms, Zero-crossing rate, MFCC, and Histogram together to avoid information loss. By fusing different types of features, a richer and more comprehensive data representation can be formed, which helps to improve the accuracy of the model in tasks such as classification, detection, and segmentation. The model structure diagram is shown in Fig.1.

LFE layer extracts local features. In order to further extract representative features from speech, LFE layer is used to extract local features from speech, reduce computational complexity, and enable the network to better recognize the feature structure in speech, including pitch, intonation, and emotion.

For lacking utilization for the contextual features in previous studies, this paper adopts the CFE layer to extract contextual features in order to obtain more feature representations of speech. During the research process, the GFE layer is used to enrich the emotional features of speech and improve the accuracy of the model's emotion recognition.

In the next section, we will provide a detailed introduction to the functions of each layer in STA-CNN.

B. Model Introduction

(1) FF layer

To address the issue of information loss in speech, the features extracted from the speech signals are fused before the input layer. In order to adapt the model to more complex tasks, reduce error propagation, and avoid information loss, early feature fusion method is proposed.

These features fused in the speech before the input layer can more comprehensively describe the speech signals. The purpose of features fusing at the feature level is to reduce the propagation of errors caused by a single feature extraction process in subsequent processing, as errors in different features are often independent. Feature fusion can provide more complex and refine representations for these tasks. Refer to Fig.1 for the FF layer.

Before conducting SER, it is necessary to extract emotional feature information from the speech. As shown in Fig.1, this paper extracts Spectrograms, Zero-crossing rate, MFCC, and Chromatogram features from speech. The Spectrograms feature in speech is represented as x_S, the Zero-crossing rate feature is represented as x_Z, the MFCC feature is represented as x_M , and the Chromatogram feature is represented as x_C. After features fusion, x_F is generated, where x_F represents the fused features. Please refer to (1) for the fusion formula.

$$x_F = x_S + x_c + x_M + x_Z \tag{1}$$

(2) LFE layer

The LFE layer is a local feature extraction layer and is a key link in speech processing. It can extract local features information from speech signals, which can reflect the subtle changes of speech and provide important data support for subsequent speech recognition and speech analysis. It is composed of three convolutional layers and three pooling layers. The layers composition is shown in Fig.1. Among them, three convolutional layers are designed to extract different local features information. Conv1 extracts the pitch of the fused feature, Conv2 extracts the intonation of the fused feature, and Conv3 extracts the emotion of the fused feature. Refer to Fig.1 for the LFE layer.

The LFE layer extracts the local features from the special x_F which is through three convolutional layers (conv1, conv2, conv3). Each layer uses the ReLU activation function, and then performs the maximum pooling operation, as shown in Fig.1.

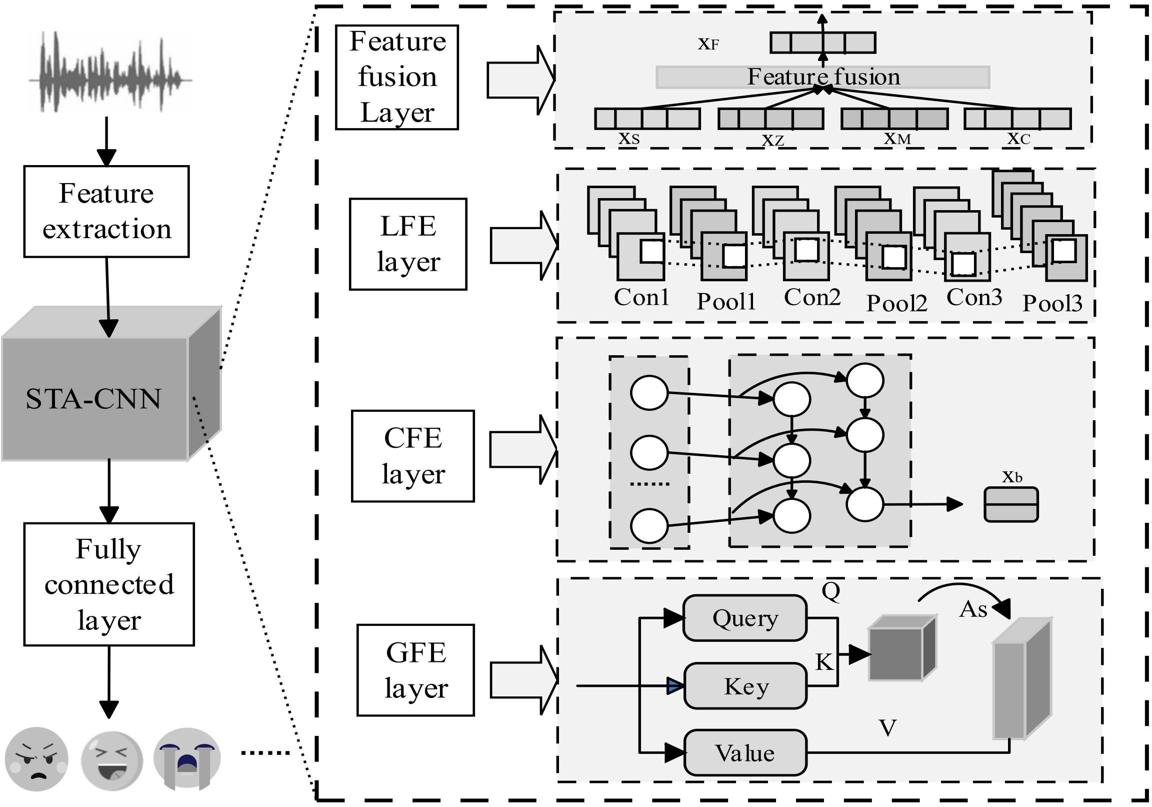

Fig.1. Structure diagram of speech emotion recognition model based on STA-CNN

Apply convolution kernels k_j of different scales and weight matrices W_j to input feature x_F, and perform nonlinear activation, refer to formula (2):

$$X_{c_j} = \mathrm{Re}\,LU\left(\sum_{i=1}^{C} W_j^{(i)} * X_F^{(i)}\right), j = \{1, 2, 3\} \tag{2}$$

Among them, C is the number of channels, $*$ represents the convolution operation, $W_j^{(i)}$ is the convolution kernel weight. Pooling operation is applied to each X_{c_j}.

(3) CFE layer

The CFE layer is a Bi-LSTM (Bidirectional Long Short-Term Memory) network, whose input x_c is generated by the LFE network layer and the output is recorded as x_b. x_b The calculation for x_b can refer to formula (3). Bi-LSTM network enable simultaneous processing of forward and backward dependencies of sequence data through the unique bidirectional architecture, in which the data stream is divided into two directions: a forward LSTM and a backward LSTM.

The forward LSTM processes data in a normal sequence order, and captures information from the past to the current moment， from the beginning to the end of the sequence. The backward LSTM, meanwhile, processes the data in reverse order, from the end to the beginning of the sequence, fetching information that never come to the current moment. These two LSTM layers are structurally identical, but their learning parameters are independent. At each time step, the forward and backward LSTMs produce an output respectively, then they are combined (usually by stitching) into a single output that contains the forward and backward context information for the current time step. This merging operation ensures that the model has access to the full context of the entire sequence at each point in time, enabling it more accurately to understand and predict the sequence data. Concatenate all pooled features and input them into a bidirectional LSTM layer.

$$X_b = B\left(\sum_{j=1}^{3} X_{\mathrm{F}j}\right) \tag{3}$$

(4) GFE layer

In order to further obtain the emotional representation of features, we use the global feature extraction layer GFE to extract the global features. the GFE layer adopts the self-attention mechanism module, in which the self-attention mechanism can calculate the attention score between any two elements in the sequence in parallel, so it can consider all the elements in the sequence at the same time. Because the self-attention directly calculates the relationship between any two positions in the sequence, it can effectively capture the long-distance dependence, without being limited by the length of the sequence in the process of calculating the weighted sum. The output features of each position are calculated based on the information of the entire sequence, which means that the representation of each position is integrated into the global features, and self-attention usually appears in the form of multi-head attention, which can significantly improve the expression ability of the model. Each head extracts features from different subspaces and learns different attention patterns to capture a richer global feature set.

Calculate the weighted similarity between each query and key by calculating the attention score, and normalize it. The output of the GFE layer can be referred to (4):

$$A_s^{(i,j)} = \frac{\exp\left(\frac{Q^{(i)} \cdot K^{(j)}}{\sqrt{d}}\right)}{\sum_{k=1}^{N} \exp\left(\frac{Q^{(i)} \cdot K^{(k)}}{\sqrt{d}}\right)}, \ i, j \in \{1, \ldots, N\} \tag{4}$$

Among them, $A_s^{(i,j)}$ represents weighted sum of attention score, $Q^{(i)}$ represents the query vector, $K^{(j)}$ represents the key vector, N represents the sequence length, $\exp$ is an exponential function used to convert the results of molecular components into positive numbers and enhance terms with high correlation. d represents the dimension of the vector, $\sqrt{d}$ represents the square root of the dimension and is used as a normalization factor to ensure that the dot product does not become too large as the dimension increases.

$W_{sum}^{(i)}$ represents the weight vector of the i-th word, $A_s^{(i,j)}$ represents the coefficient of influence of word s on words i and j, $V^{(j)}$ represents the feature vector of the j-th word, i and j respectively represent the subscripts of two different words. The meaning of this formula is to calculate the sum of the weights of all words in a sentence or text sequence, where the weight of each word is determined by its feature. $W_{sum}^{(i)}$ can be referred to (5):

$$W_{sum}^{(i)} = \sum_{j=1}^{N} A_s^{(i,j)} V^{(j)} \tag{5}$$

The output x_G of GFE can be referred to (6):

$$x_G = \sum_{i=1}^{N} W_{sum}^{(i)} \tag{6}$$

C. Output

The fully connected layer, also known as the dense layer, is a fundamental component of neural network. Its output can be represented by a mathematical formula that describes how a fully connected layer processes input data through a series of linear transformations and nonlinear activation processes to generate the final output result. Specifically, the output of the fully connected layer can be expanded to the following lengthy statement:

In neural network, the fully connected layer receives the outputs of all neurons from the previous layer as inputs, which are first weighted and summed with the weights of each neuron in the fully connected layer. This process can be seen as a linear combination of input data, where each weight represents the relative importance of each feature in the input data to the current neuron. After the weighted sum is completed, an additional bias term is added to the result, which allows the neural network to learn information beyond the input features. The output of this nonlinear activation function is the final output of the fully connected layer. Therefore, the output of a fully connected layer can be represented as the product of the input feature vector and the weight matrix pluses the bias vector, and then is processed by an activation function. This output not only includes the linear transformation of the input data, but also includes the nonlinear characteristics introduced by the activation function, enabling the fully connected layer to learn more complex data representations.

The output of the last layer can be referred to (7):

$$Output = soft\max(Wx + b) \qquad (7)$$

IV. Experiments

A. Datasets

In order to comprehensively evaluate the generalization ability and effectiveness of the proposed model, this paper carefully selects multiple datasets that are highly representative in the field of SER: EMODB[26], IEMOCAP[27], RAVDESS[28], SAVEE[29].

TABLE I.
SENTIMENT INFORMATION FOR FOUR DATASETS

Dataset	Language	Sample Quantity	Emotional Category
EMODB	German	535	7
IEMOCAP	English	5531	4
RAVDESS	English	1440	8
SAVEE	English	480	7

TABLE II.
DETAILED SENTIMENT INFORMATION FOR FOUR DATASETS

D	AN	BO	CL	DI	FE	HA	NE	SA	SU
E	127	81	-	46	69	71	79	62	-
I	1103	-	-	-	-	1636	1708	1084	-
R	192	-	192	192	192	192	96	192	192
S	60	-		60	60	60	120	60	60

D = dataset, E=EMODB, I=IEMOCAP, R=RAVDESS, S=SAVEE, AN=Angry, BO=Bored, CL=Clam, DI=Disgust, FE=Fearful, HA=Happy, NE=Neutral, SA=Sad, SU=Surprised.

The total number of samples in the EMODB speech dataset is 535, with a total of 7 emotion labels. The sample size of the IEMOCAP dataset is 5531, and 4 of the emotion labels are used. The sample size of the RAVDESS dataset totals 1440, with a total of 8 emotion labels. The sample size of the SAVEE Chinese language dataset totals 480, with a total of 7 sentiment labels. Each dataset gives the number of emotional information and the corresponding dataset division, and its division ratio is training set: validation set: test set = 6:2:2. TABLE I and TABLE II show the emotions contained in each dataset and the corresponding number of emotions. There are seven emotions in the EMODB dataset, which are neutral, angry, afraid, happy, sad, disgusted, and bored. The IEMOCAP dataset has four emotions, which are happy, angry, sad, and neutral, and the RAVDESS dataset has eight emotions, including calm, happy, sad, fearful, angry, surprised, disgusted, and neutral. The SAVEE dataset has seven emotions, which are anger, disgust, fear, happiness, neutrality, sadness, and surprise.

B. Experimental Setting

In this section, we introduced the parameter settings for the experiments. The settings and related experimental settings used in the experiments are shown in TABLE III. The Adam (Adaptive Moment Estimation) optimizer was selected for the experiments. Adam is a widely used optimization algorithm that can handle sparse gradient and noisy data, and has shown good convergence speed and effectiveness in practice. The Adam optimizer combines the advantages of momentum method and RMSProp algorithm, dynamically adjusting the learning rate by correcting the first-order and second-order moment estimates. The regularization coefficient used in the experiment is 0.001, and the learning rate is 0.0001.

TABLE III.
RELATED EXPERIMENTAL ENVIRONMENT SETTINGS

Experimental Environment	Setting or Version Number
Convolutional Kernel	3
Graphics	NVIDIA Geforce RTX 409
Video memory	128G
Processor	13th Gen Intel(R) Core(TM) i9-13900K
Optimizer	Adam

C. Performance metrics for different datasets

In order to evaluate the effectiveness of the experiments, three evaluation indicators (precision, recall, and F1-score) were selected to verify the experimental effect of the model on each emotion in the EMODB, IEMOCAP, RAVDESS and SAVEE dataset.

(1) Results and analysis on the EMODB dataset

TABLE IV shows the recognition rates of each emotion on the EMODB dataset. It can be concluded that the model achieves a recognition accuracy of 75% for nausea and sadness, as these two models have relatively strong emotional representation abilities when extracting speech emotion features, resulting in higher accuracy of the models.

TABLE IV.
SEVEN EMOTION RECOGNITION RESULTS ON EMODB DATASET

Emotion	Precision (%)	Recall (%)	F1 Score (%)
Angry	66.67	77.78	71.79
Bored	50.00	40.00	44.44
Disgusted	75.00	75.00	75.00
Fearful	53.85	50.00	51.85
Happy	42.11	44.44	43.24
Neutral	57.14	50.00	53.33
Sad	75.00	100.00	85.71

(2) IEMOCAP dataset results and analysis

TABLE V shows the experimental results of the STA-CNN model on the IEMOCAP dataset for four-classification emotion recognition.

TABLE V.
FOUR EMOTION RECOGNITION RESULTS ON IEMOCAP DATASET

Emotion	Precision (%)	Recall (%)	F1 Score (%)
Angry	73.65	47.19	57.52
Happy	56.44	50.29	53.19
Neutral	49.52	65.30	56.33
Sad	62.61	68.04	65.21

According to the analysis in TABLE V, only the Angry emotion exceeds 70%, and the recognition rate of Neutral emotion is low, only 56.44%, because the representation color of Neutral emotion is not as strong as that of Angry and Sad, and the representation ability of Happy is not strong, so the recognition accuracy of a single emotion is not high.

(3) Results and analysis on the RAVDESS dataset

In order to verify the effectiveness of the STA-CNN model in SER tasks, eight classification experiments were carried out on the RAVDESS dataset. In this paper, the speech part of the RAVDESS dataset is used for experiments, which are divided into eight types of emotion labels: neutral, calm, happy, sad, angry, scared, disgusted and surprised. The experimental results are shown in TABLE VI, from the table, it can be seen that Calm emotion recognition has the best performance, and this emotion exhibits extremely high accuracy in emotion recognition tasks, thanks to the unique context aware mechanism and efficient feature extraction ability of the STA-CNN model.

TABLE VI.
EIGHT EMOTION RECOGNITION RESULTS ON RAVDESS DATASET

Emotion	Precision (%)	Recall (%)
Angry	65.00	41.94
Calm	90.91	85.11
Disgusted	58.82	74.07
Fearful	53.85	65.62
Happy	73.81	91.18
Neutral	59.38	63.33
Sad	68.75	66.67
Surprised	80.00	66.67

(4) Results and analysis on SAVEE dataset

The STA-CNN SER model was used to conduct seven-classification experiments on the SAVEE dataset to evaluate the performance of the model in the multi-classification emotion recognition task. The experimental results are shown in TABLE VII, which respectively show the Precision, Recall and F1 performance indicators of the model on different emotion categories. The model has achieved good accuracy in the seven emotions of Angry, Disgusted, Fearful, Happy, Neutral, Sad and Surprised.

TABLE VII.
SEVEN EMOTION RECOGNITION RESULTS BASED ON SAVEE DATASET

Emotion	Precision (%)	Recall (%)	F1 Score (%)
Angry	25.00	50.00	33.33
Disgusted	71.43	83.33	76.92
Fearful	50.00	60.00	54.55
Happy	33.33	66.67	44.44
Neutral	100.00	84.21	91.43
Sad	75.00	85.71	80.00
Surprised	100.00	16.67	28.57

D. The recognition effect of the model on four datasets

The comparison of experimental data of different models on different datasets shows the performance of the STA-CNN model in dealing with different emotional states more comprehensively, and further verifies its effectiveness and reliability in practical applications. model.

As can be seen from TABLE VIII, the STA-CNN model has significant effects on four sentiment datasets, and the results of SER are improved compared with the previous ones. On the EMODB dataset the recognition accuracy is 0.23% higher than that of the benchmark model. According to the analysis in TABLE VIII, the accuracy of the proposed STA-CNN model is improved by 3.07% compared with the EMQDPS model on the IEMOCAP dataset. The accuracy of emotion recognition on the RAVDESS dataset is 76.39%, which is 0.77% higher than that of the previous Hybrid LSTM and 4.46% higher than SVM on the SAVEE dataset, which fully verifies the effectiveness of the model.The accuracy of the model was improved by extracting spectral features, MFCC parameters, chromatogram and zero-crossing rate features, and the effect of the model on EMODB, IEMOCAP, RAVDESS and SAVEE datasets was evaluated. The results on EMODB, IEMOCAP, RAVDESS and SAVEE datasets were 80.23%, 71.96%, 76.39% and 80.09%, respectively. Compared with the benchmark models, the effectiveness of the model was fully verified. Experimental results showed that the model achieved good classification results on EMODB and SAVEE datasets, because the amount of dataset is smaller. The model is more suitable for small-scale datasets, the training and debugging model with small datasets is usually faster, and the model training time is shorter. Different model architectures, hyperparameters and feature engineering methods can be tried in order to get more quickly model. The reason for the poor results on the IEMOCAP dataset is that the sample size of the dataset is large, the model cannot effectively extract all the features, so the accuracy of 71.96% is only achieved on the IEMOCAP dataset, while the sample size of the RAVDESS dataset is relatively balanced, so the accuracy is better than the benchmark models.

TABLE VIII.
EXPERIMENTAL EXPERIMENTAL RESULTS OF THE DATASET

Dataset	Model	Accuracy（%）
EMODB	a deep and shallow feature fusion convolutional network[30]	63.50
	SVM[31]	80.00
	STA-CNN	**80.23**
IEMOCAP	RNNs[32]	59.70
	IAAN[33]	66.30
	EMQDPS[34]	68.89
	STA-CNN	**71.96**
RAVDESS	SVM[35]	64.31
	BLSTM + Capsule routing[36]	69.40
	Hybrid LSTM Transformer[37]	75.62
	STA-CNN	**76.39**
SAVEE	Deep belief network (DBN)[38]	56.76
	Logistic model tree (LMT)[39]	70.40
	1D CNN[40]	65.83
	SVM[35]	75.63
	STA-CNN	**80.09**

In the final recognition results, we used a bar chart to visually display the performance comparison of different models or methods. Fig.2 shows the recognition results on the EMODB dataset, IEMOCAP dataset, RAVDESS dataset, and SAVEE dataset.

As shown in the figure, the height of the bar chart directly reflects the accuracy performance of each model. Among them, the bar chart of the proposed model is significantly higher than other comparison methods, indicating that our model has significant performance advantages in recognition tasks. Specifically, the height difference of the bar chart not only reflects the excellent performance of the model on training data, but also further verifies its strong generalization ability on test data. In addition, the high bar chart also indicates that our model can better adapt to complex task scenarios and has strong robustness and stability. This result fully proves the effectiveness and progressiveness of the proposed method.

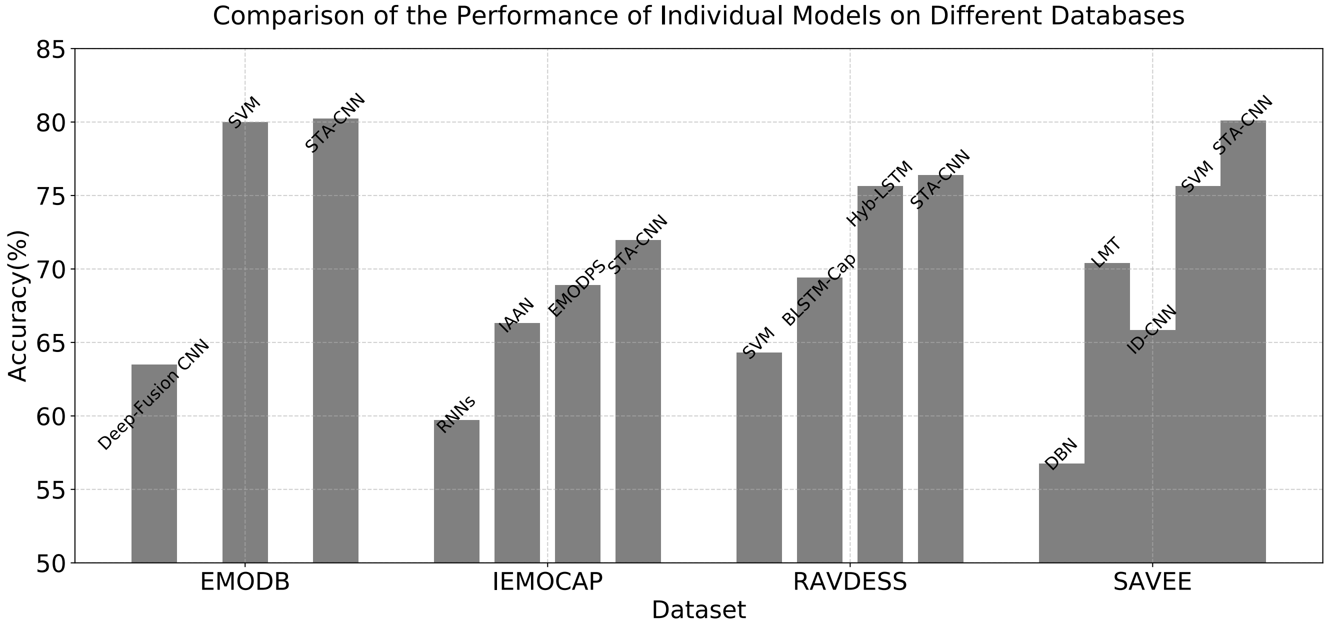

Fig.2. Column chart results of the model on four different datasets

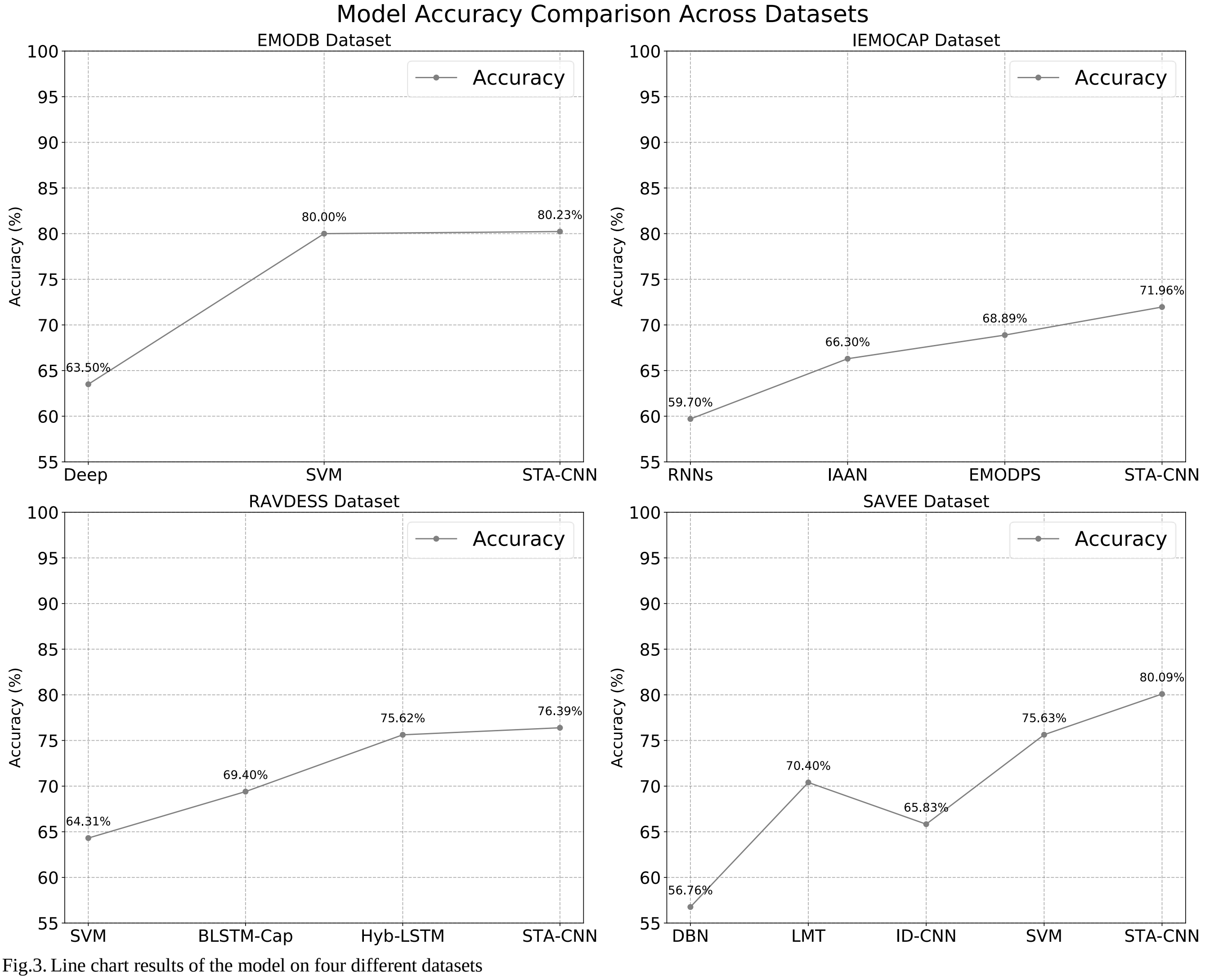

Fig.3. Line chart results of the model on four different datasets

The line chart can clearly illustrate the advantages of the proposed model and the comparative model. Fig.3 represent the line graphs of the EMODB dataset, IEMOCAP dataset, RAVDESS dataset, and SAVEE dataset. As shown in the figure, all line charts exhibit a clear upward trend, which fully demonstrates the significant improvement in accuracy of the proposed model. This result not only validates the excellent performance of the model on training data, but also demonstrates its strong generalization ability on test data, further proving that the model has high adaptability and robustness when facing different task scenarios. In addition, the continuous improvement of model performance also reflects its stability and reliability in handling complex tasks, highlighting its significant advantages among similar methods.

In summary, the STA-CNN model proposed in this article has shown good performance on four different datasets: EMODB, IEMOCAP, RAVDESS, and SAVEE, and the performance improved by 0.23%, 3.07%, 0.77% and 4.46% compared to the baseline datasets, respectively. This indicates that by introducing contextual feature information and capturing global feature information from the data, the model fully utilizes the feature information in speech, which can effectively capture the details and dynamic changes of speech without improving the model's expressive power.

V. Conclusion

In view of the problem of ignoring the contextual feature information, lossing information and further obtaining the emotional information features in speech, this paper proposed a model called STA-CNN to identify speech emotion, which extracted the local features of the speech signal through the LFE layer, the contextual information features of the speech signal through the CFE layer, the global features of the speech through the GFE layer, and finally outputs the classification results. The purpose of this paper is to systematically explore the effectiveness and optimization methods of deep learning based speech emotion recognition models. By adjusting the hyper parameters and comparing the baseline models, we find that the proposed model shows excellent performance on multiple performance indicators. The STA-CNN model is experimentally verified on EMO-DB, IEMOCAP, RAVDESS and SAVEE datasets, and the accuracy of the model's sentiment classification is 80.23%, 71.96%, 76.39% and 80.09%, respectively, which is 0.23%, 3.07%, 0.77% and 4.46% higher than that of the benchmark model. A series of experiments have been done to evaluate the impact of different hyper parameters on the performance of the model, and the important role of key parameters on the stability of the model is revealed.

Although this study has achieved certain results, there are also limitations such as limited data volume and high model complexity. Future research can be expanded in the following aspects: first, more diverse datasets can be collected to enhance the generalization ability of the model; Secondly, more efficient network structures and training strategies can be explored to reduce the complexity of the model. Thirdly, the model can be applied to a wider range of real-world scenarios to verify its application value. In addition, given the complexity of emotion recognition, an interdisciplinary approach may provide new perspectives for solving current problems. This study provides new insights and methods for the field of SER, and it is expected that future work can further promote the development of this field.

VI. Declarations

Contributions by each author: Yancong Zhou is responsible for the conception, guidance and important revision of the paper; Xiaoli Xie is responsible for work design and program verification.

We declare that we have no known competing economic interests or personal relationships that may appear to affect the work of this report.

The data involved in the study is public data. They do not involve moral and trade secrets. The public dataset is the speech emotion dataset portion of EMODB, IEMOCAP, SAVEE, and RAVDESS and can be passed separately.

The datasets can be download through the internet

EMODB Dataset: http://emodb.bilderbar.info/docu/#emodb

IEMOCAP Dataset: https://sail.usc.edu/iemocap/

RAVDES Dataset:

https://zenodo.org/record/1188976#.YuE4Z3ZBwak

SAVEE Dataset:

http://kahlan.eps.surrey.ac.uk/savee/Download.html

References

[1] C. Hema. and F.P.G Marquez, “Emotional speech recognition using CNN and deep learning techniques,” *Appl. Acoust.*, vol. 211, art. no.1386, August 2023, 211: Article 109492. https://doi.org/10.1016/j.apacoust.2023.109492

[2] W. Alsabhan, “Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention,” *Sensors*, vol. 3, art. no. 1386, Jan. 2023.https://doi.org/10.3390/s23031386.

[3] Z. Yang, S. Zhou, L. Zhang, and S. Serikawa, “Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network,” *Cognitive Robotics*, vol. 4, pp.30-41, 2024. https://doi.org/10.1016/j.cogr.2023.12.001

[4] A. Hashem, M. Arif, and M. Alghamdi, “Speech emotion recognition approaches: A systematic review,” *Speech Commun.*, vol. 154, art. no. 102974, October 2023.https://doi.org/10.1016/j.specom.2023.102974

[5] Y. Ü. SÖNMEZ, A. VAROL, “In-depth investigation of speech emotion recognition studies from past to present –The importance of emotion recognition from speech signal for AI–,” *Intelligent Systems with Applications*, vol. 22, art. no.200351, Jun.2024, https://doi.org/10.1016/j.iswa.2024.200351

[6] T. Mary, and T. Jaya, “A novel concatenated 1D-CNN model for speech emotion recognition,” *Biomedical Signal Processing and Control*, vol. 93, 2024, 93: Article 106201. https://doi.org/10.1016/j.bspc.2024.106201

[7] M. R. Ahmed., S. Islam, and A. K. Muzahidul, “An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition,” *Expert Syst. Appl.*, vol. 218, art. no. 119633, May.2023. https://doi.org/10.1016/j.eswa.2023.119633

[8] K. Manohar, D. E. Logashanmugam, “Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm,” *Knowl. - Based Systems*, vol. 246, art. no. 108659, Jun. Mar2022.

[9] M. Wang, H. Ma, Y. Wang, and X. Sun, “Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion.” *Appl. Acoust.*, vol.218, art. no. 109886, Mar.2024. https://doi.org/10.1016/j.apacoust.2024.109886

[10] J. Hyeon, Y. Oh, Y. Lee, H. Choi, “Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts,” *Data & Knowledge Engineering*, vol. 150, art. no. 102262, Mar. 2024. https://doi.org/10.1016/j.datak.2023.102262

[11] W Lin., C. Busso, “Deep temporal clustering features for speech emotion recognition,” *Speech Commun.*, vol. 157, art. no. 103027, Feb. 2024. https://doi.org/10.1016/j.specom.2023.103027

[12] H. Zhang, H. Huang, and H. Han, “MA-CapsNet-DA: Speech emotion recognition based on MA-CapsNet using data augmentation,” *Expert*

Syst. Appl., vol. 244, art. no. 122939, June, 2024. https://doi.org/10.1016/j.eswa.2023.122939

[13] R Soltani, and E. Benmohamed, "Newman-Watts-Strogatz topology in deep echo state networks for speech emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 133, art. no. 108293, July, 2024. https://doi.org/10.1016/j.engappai.2024.108293

[14] N. Naderi, and B. Nasersharif, "Cross Corpus Speech Emotion Recognition using transfer learning and attention-based fusion of Wav2Vec2 and prosody features," *Knowl. - Based Syst.*, vol. 277, art.no.110814, Oct.2023.

[15] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert. Syst. Appl.*, vol. 214, art.no. 118943, Mar.2023. https://doi.org/10.1016/j.eswa.2022.118943

[16] Z. Yang, Z. Li, S. Zhou, L. Zhang, and S. Serikawa, "Speech emotion recognition based on multi-feature speed rate and LSTM," *Neurocomputing*, vol. 601, art. no. 128177, October,2024. https://doi.org/10.1016/j.neucom.2024.128177

[17] H. Lian, C. Lu, Y. Zhao, S. Li, T. Qi, and Y. Zong, "Exploring corpus-invariant emotional acoustic feature for cross-corpus speech emotion recognition," *Expert Syst. Appl.*, vol.258, art. no. 125162, December ,2024. https://doi.org/10.1016/j.eswa.2024.125162

[18] N. Saleem, H. Elmannai, S. Bourouis, "Squeeze-and-excitation 3D convolutional attention recurrent network for end-to-end speech emotion recognition," *Appl. Soft Comput.*, vol.161, art. no. 111735, August, 2024. https://doi.org/10.1016/j.asoc.2024.111735

[19] H. Li, X. Zhang, S. Duan, and H. Liang, "Speech emotion recognition based on bi-directional acoustic–articulatory conversion," *Knowl. - Based Syst.*, Vol. 299, art. no. 112123, September, 2024. https://doi.org/10.1016/j.knosys.2024.112123

[20] H. Wang, P. Song, S. Jiang, R. Wang, S. Li, and T. Liu, "Domain adaptive dual-relaxation regression for speech emotion recognition," *Appl. Acoust.*, Vol. 224, art. no. 1101185, September, 2024.https://doi.org/10.1016/j.apacoust.2024.110118

[21] M. Wang, H. Ma, Y. Wang, and X. Sun, "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion," *Appl. Acoust.*, Vol. 218, art. no. 109886, March 2024. https://doi.org/10.1016/j.apacoust.2024.109886

[22] M. Liu, A. N. J. Raj, V. Rajangam, K. Ma, Z. Zhuang, S. Zhuang， Multiscale-multichannel feature extraction and classification through one-dimensional convolutional neural network for Speech emotion recognition[J], *Speech Commun.*, Vol. 156, art. no.103010, January 2024. https://doi.org/10.1016/j.specom.2023.103010

[23] L. Yu, F. Xu, Y. Qu, and K. Zhou, "Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion," *Appl. Acoust.*, Volume 216, art. no. 109752, January 2024. https://doi.org/10.1016/j.apacoust.2023.109752

[24] F. Harby, M. Alohali, A. Thaljaoui, A. S. Talaat, "Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition," *Computers, Materials and Continua*, Vol. 78, PP. 2689-2719, February 2024. https://doi.org/10.32604/cmc.2024.046623

[25] S. P. Mishra, P. W., S. Deb, "Speech emotion recognition using a combination of variational mode decomposition and Hilbert transform," *Appl. Acoust.*, Vol. 222, art. no. 110046, June 2024.https://doi.org/10.1016/j.apacoust.2024.110046

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier and B. Weiss, "A database of German emotional speech[C]. IEEE, INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Commun. and Technology, Lisbon, Portugal, September 2005.

[27] C. Busso, M. Bulut, and C. C. Lee, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol.42, iss. (4), PP.335-359, 2008.

[28] S.R Livingstone, and F.A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american English," *PloS one* 2018, 13(5): e0196391.

[29] S. Haq, P.J.B. Jackson, and J.D. Edge, "Audio-Visual feature selection and reduction for emotion classification," [C]. i*n: Proceedings of the International Conference on Auditory-Visual Speech*, PP. 185–190, 2008.

[30] L.H. Sun, et al. "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition," *Int J Speech Technol*, vol.21, iss. (4), pp. 931-940,2018

[31] B.C. Chiou, C.P. Chen, "Feature Space Dimension Reduction in speech emotion recognition using Support Vector Machine," in: Signal and Information Processing Association Annual Summit and Conference, Asia-Pacific, 29 October 2013, pp. 1–6.

[32] J. Tao, J. Huang, Y. Li, Z. Lian and M. Niu, "Semi-supervised ladder networks for speech emotion recognition," *International Journal of Automation and Computing*, vol. 16, (4), pp. 437-448, 2019. https://link.springer.com/article/10.1007/s1163019-1175-x

[33] S.L. Yeh, Y. Lin, and C. Lee. "An interaction-aware attention network for speech emotion recognition in spoken dialogs[C]. Paper presented at the ICASSP 2019–2019 *IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP),2019.

[34] F. Daneshfar, and S.J. Kabudian, "Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm," *Multimedia Tools and Applications*, vol. 79, pp. 1261-1289, 2020:. https://link.springer.com/article/10.1007/s11042-019-08222-8

[35] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179,: art. no. 108046, August 2021. https://doi.org/10.1016/j.apacoust.2021.108046

[36] M.A. Jalal, E. Loweimi, R.K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," *Proceedings of interspeech*, pp. 1701-1705, 2019. http://dx.doi.org/10.21437/Interspeech.2019-3068

[37] F. Andayani, L.B. Theng, M.T. Tsun, and C. Chua, "Hybrid lstm-transformer model for emotion recognition from speech audio files," *IEEE Access*, vol. 10, pp. 36018-36027, 2022.

[38] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *Computer Vision and Pattern Recognition*, 2018.

[39] G. Assunção, P. Menezes, F. Perdigão, "Speaker awareness for speech emotion recognition," *International Journal of Online and Biomedical Engineering*, vol. 16, (4), pp.15-22, 2020.

[40] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom. "Negative emotion recognition using deep learning for thai language," 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, *Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, pp. 71-74, IEEE 2020.