IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 June 2023, accepted 3 July 2023, date of publication 11 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294274

RESEARCH ARTICLE

A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering

LI-MIN ZHANG[1,2], GIAP WENG NG[2], YU-BENG LEAU[2], (Senior Member, IEEE), AND HAO YAN[1]

[1]Key Laboratory for Artificial Intelligence and Cognitive Neuroscience of Language, Xi'an International Studies University, Xi'an 610116, China
[2]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Sabah 88400, Malaysia

Corresponding author: Hao Yan (haoyan@xisu.edu.cn)

This work was supported in part by the Social Science Foundation of Shaanxi Province of China under Grant 2022K014, in part by the National Social Science Foundation of China under Grant 20BYY097, and in part by the Natural Science Basic Research Program of Shaanxi Province of China under Grant 2023-C-QN-0725.

ABSTRACT Speech Emotion Recognition (SER) is a common aspect of human-computer interaction and has significant applications in fields such as healthcare, education, and elder care. Although researchers have made progress in speech emotion feature extraction and model identification, they have struggled to create an SER system with satisfactory recognition accuracy. To address this issue, we proposed a novel algorithm called F-Emotion to select speech emotion features and established a parallel deep learning model to recognize different types of emotions. We first extracted the emotion features from speech and calculated the F-Emotion value for each feature. These values were then used to determine the combination of speech emotion features that was optimal for speech emotion recognition. Next, a parallel deep learning model was established with the speech emotion feature combination as input to train and test for each type of emotion. Finally, decision fusion was applied to the parallel output results to obtain an overall recognition result. These analyses were conducted on two datasets, RAVDESS and EMO-DB, with the accuracy of speech emotion recognition reaching 82.3% and 88.8%, respectively. The results demonstrate that the F-Emotion algorithm can effectively analyze the correspondence between speech emotion features and emotion types. The MFCC feature best describes emotions of neutrality, happiness, fear, and surprise, and Mel best describes emotions of anger and sadness. The parallel deep learning model mechanism can improve the accuracy of speech emotion recognition.

INDEX TERMS Speech emotion recognition, F-emotion algorithm, feature clustering, parallel model, deep learning.

I. INTRODUCTION

Speech is an essential mode of human communication [1], as it allows us to express not only our thoughts but also our emotions. Acoustic cues present in speech can reveal various emotional states, allowing a deeper understanding of the message being conveyed. Speech emotion recognition involves the identification and analysis of these cues within recorded speech signals, and their mapping to the corresponding emotional states. This process is crucial for facilitating effective communication and intelligent human-machine interaction. Speech emotion recognition has numerous applications in fields such as medicine [2], education [3], criminal investigation [4], and elder care [5].

Selecting features that can represent different emotions is a major challenge in speech emotion recognition, as the precision of feature selection greatly impacts recognition [6]. Luengo et al. [7] employed the J1 Criterion formula to compute speech emotion feature parameters, parameter combinations and evaluated them using feature fusion technology. The findings showed that spectral features outperformed the other features in emotion recognition. Chen et al. [8] combined non-personalized speech emotion features based on derivatives with traditional personalized speech emotion

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

features to enhance recognition accuracy. Özseven [9] proposed a statistical feature selection approach capable of filtering out effective features and improving classification success rates. Abdulmohsin et al. [10] designed a method for extracting feature means and standard deviations, and the method exhibited strong recognition ability. Farooq et al. [11] utilized deep convolutional neural networks (DCNNs) to extract speech emotion features and integrated feature selection techniques to identify suitable and discriminative features for the classifier, achieving an accuracy of 81.30% on the RAVDESS dataset. Er [12] combined acoustic and deep features to generate mixed features, and employed the Relief F algorithm to select more effective speech emotion features from the mixed feature vector, obtaining an accuracy of 79.41% on the RAVDESS dataset. Sönmez et al. [13] developed a new lightweight effective SER method called 1BTPDN, which has low computational complexity. Previous research has indicated that there are variations in the types of acoustic features exhibited by different emotions [14], necessitating a separate analysis of the speech emotion features or feature combinations that represent different emotion types.

To examine the relationship between speech emotion features and different emotion types, this study used statistical methods to analyze the degree of aggregation and relative dispersion of each emotion type for each speech feature parameter. F-Ratio [7] is a statistic commonly used in variance analysis to test for significant differences in mean values between two or more samples. In speaker recognition, F-Ratio has been successfully applied [15], [16] to evaluate the discriminative ability of individual speaker recognition parameters and their dependence on other parameters to achieve favorable outcomes. Poh et al. [15] improved speaker recognition systems by normalizing F-Ratio and standardizing the data before selecting the decision threshold. Chen et al. [16] used the phoneme-average F-Ratio method to examine the contributions of different frequency regions to Chinese speakers' phoneme recognition, and applied it to speaker recognition. Compared with mel-frequency cepstrum coefficient (MFCC) features, this feature reduced the recognition error rate by 32.94%.

Building on the principles outlined above, we propose an **F-Emotion algorithm** for the cluster analysis of speech emotion features to examine the association between speech emotion features and emotion types. The F-Emotion algorithm can precisely describe the degree of clustering of a particular emotion under a speech feature parameter and its relative dispersion compared with other emotions. The higher the degree of clustering, the more prominent the feature, whereas the higher the dispersion, the easier it is for the classifier to recognize it. To assess the impact of speech emotion features or feature combinations on different types of emotions, a parallel deep learning model is required to test the effect of emotion recognition on distinct speech emotion features or speech emotion feature combinations.

However, current speech emotion recognition models use a single network architecture to recognize multiple emotions. Singh et al. [17] proposed a hierarchical deep learning approach for the overall recognition of eight emotions using the RAVDESS dataset. Xiao et al. [18] achieved an 81% recognition rate in the Danish emotion corpus by testing a multilevel classification recognition model based on a dimensional emotion model. Zehra et al. [19] used an ensemble classifier with a majority voting mechanism to recognize multilingual speech emotions across corpora. Atila et al. [20] proposed an end-to-end attention-guided 3D CNN-LSTM model that predicts emotion in speech. Sun et al. [21] proposed a deep neural network (DNN) decision tree support vector machine-based method that trains different DNN networks for different emotion groups, thereby reducing confusion between emotions. However, most existing models ignore differences in feature combinations on the input side and rely on single-channel deep learning models. To improve the accuracy of speech emotion recognition, it is crucial to establish a parallel deep learning model that recognizes each emotion individually and demonstrates the impact of different feature combinations on different emotions.

In this study, we analyzed the impact of speech emotion feature parameters on emotion classification, and established a parallel deep learning model mechanism for speech emotion recognition. This study makes several contributions to the literature.

First, a novel feature selection algorithm called F-Emotion is proposed. For each type of selected speech emotion feature, the F-Emotion value corresponding to each type of emotion was calculated. Based on the calculation results of F-Emotion algorithm, the effect of each speech emotion feature on the recognition accuracy of different emotion types was analyzed.

Second, a parallel deep learning model was established for speech emotion recognition. It uses the optimal features or feature combinations analyzed by F-Emotion as inputs and produces separate recognition results for each emotion category.

Finally, a voting mechanism is established for decision fusion, whereby the emotion classification of each parallel channel was assigned a different weight. The overall result of speech emotion recognition was achieved by combining the outputs of all parallel channels.

The article was organized as follows. In Section II, the overall framework of speech emotion recognition is introduced, and the derivation process of the F-Emotion algorithm was described in detail. In Section III, the calculation results of speech emotion feature parameters and the recognition accuracy of the parallel deep learning model were presented. In Section IV, the proposed Multi-DNN model was compared with other deep learning models. Section V summarizes the research conducted in this study.

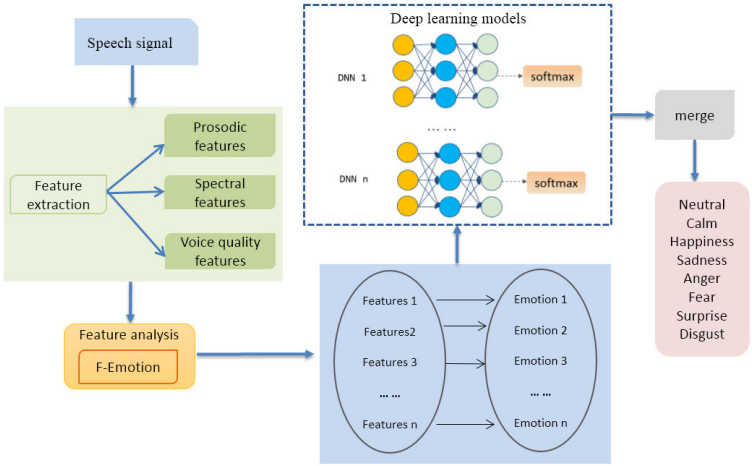

FIGURE 1. Framework of speech emotion recognition.

TABLE 1. Speech emotion feature parameters.

Type	Features	Statistical value
Prosody	Pitch, Energy, Zero-crossing rate, Short-time energy, Sound pressure level	Max, Min, Mean, Standard Deviation, Range
Spectral	MFCC, Chroma, Mel, RMS, Spectral contrast, Spectral flatness,Spectrum centroid, Spectral amplitude	
Voice Quality	Formants_f1, Formants_f2, Formants_f3, Formants_b1, Formants_b2, Formants_b3, Jitter- abs, Jitter- rel, Shimmer-abs, Shimmer-rel	

II. PROPOSED METHODOLOGY

Figure 1 shows the overall framework of speech emotion recognition. First, speech emotion features are extracted from the original speech, which includes prosodic features, spectral features, and voice quality features. The F-Emotion value of each speech emotion feature is then calculated for each emotion category. Based on these values, the effect of each speech emotion feature type on the recognition of different emotion types was analyzed, and then the optimal feature combination of each speech emotion type was identified. Next, a parallel deep learning model was established with a separate channel built for each emotion type. Using the optimal feature parameter combination as input, each channel outputs the recognition results for the corresponding emotion type. Finally, a weighted decision fusion mechanism was used to obtain the overall recognition result. DNN was the module chosen for each channel in this study.

A. EXTRACTION OF SPEECH EMOTION FEATURE PARAMETERS

Acoustical-based emotion features, including prosodic features, voice quality features, and spectral features, are commonly used to represent emotion information in speech. Prosodic features, also called supra-segmental features [22], have a good generalization performance for speech emotion recognition across different languages. Voice quality features are used to evaluate the clarity, purity, and intelligibility of speech [23], whereas spectral-based features describe the correlation between changes in the shape of the vocal tract and vocal movements [24]. In this study, we selected 23 acoustic features and analyzed their maximum, minimum, mean, standard deviation, and range values to obtain 115 statistical features. The speech emotion feature parameters are listed in Table 1 (see the Appendix for details).

B. F-EMOTION ALGORITHM

The F-ratio is a statistical measure that compares the variance between groups to the variance within groups. In the domain of speech emotion recognition, we devised an improved algorithm called F-Emotion, which was tailored to recognize emotions in speech. The F-Emotion algorithm was specially designed to be applicable to this task. In the formula for F-E motion,

μ_i denotes the mean of the *ith* emotion for a particular feature parameter $i = 1 \ldots m$ where m denotes the total number of emotion types. $\bar{\mu}$ denotes the overall mean obtained by averaging all μ_i values for a particular feature parameter.

F_i denotes the difference between the *ith* emotion type and the other emotion types in a specific speech emotion feature. The difference between the *ith* emotion type and other emotion types in a particular speech emotion feature is directly proportional to the variance of μ_i and μ_j, and it is given in Equation (1).

$$F_i \propto \sum_{j=1}^{m} (\mu_i - \mu_j)^2 \tag{1}$$

where $j = 1 \ldots m$, where $j \neq i$, and $i, j \in [1, m]$.

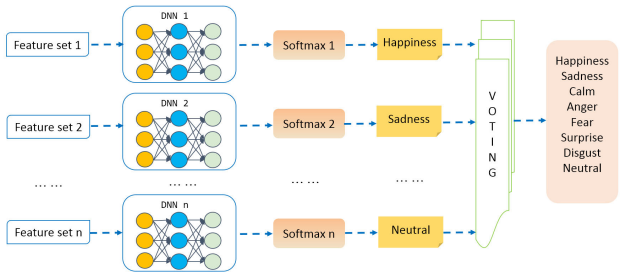

FIGURE 2. Parallel deep learning model.

It is also directly proportional to the variance of μ_i and $\bar{\mu}$, and it is given in Equation (2) and (3).

$$F_i \propto \left(\mu_i - \bar{\mu}\right)^2, \ i \in [1, m] \tag{2}$$

So

$$F_i \propto \left(\mu_i - \bar{\mu}\right)^2 \sum_{j=1}^{m} \left(\mu_i - \mu_j\right)^2 \tag{3}$$

where $j = 1 \ldots m$, where $j \neq i, i, \ j \in [1, m]$

For each emotional feature of speech, it is necessary to calculate the coherence of the speech samples. The coherence of n speech samples in the *ith* emotion type and is given in Equation (4).

$$\sum_{k=1}^{n} (x_{ik} - \mu_i)^2 \tag{4}$$

where $k = 1 \ldots n$.

Here, x_{ik} denotes the feature parameters of the *kth* speech sample for the *ith* emotion type. n represents the number of speech samples in the *ith* emotion type. Therefore, the coherence of all speech samples is given in Equation (5)

$$\sum_{j=1}^{m} \sum_{k=1}^{n} (x_{kj} - \mu_j)^2 \tag{5}$$

where $k = 1 \ldots n$,where $j = 1 \ldots m$.

F_i is inversely proportional to the coherence of the samples, and it is given in Equation (6)

$$F_i \propto \frac{1}{\sum_{j=1}^{m} \sum_{k=1}^{n} \left(x_{kj} - \mu_j\right)^2} \tag{6}$$

where $k = 1 \ldots n$, where $j = 1 \ldots m$.

Therefore, under the *ith* speech emotion feature, the F-Emotion formula is given in Equation (7).

$$F_i = \frac{\left(\mu_i - \bar{\mu}\right)^2 \sum_{j=1}^{m} \left(\mu_i - \mu_j\right)^2}{\frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{k=1}^{n} \left(x_{kj} - \mu_j\right)^2} \tag{7}$$

where $k = 1 \ldots n$ and $j = 1 \ldots m$.

The F-Emotion algorithm can be described using pseudocode as follows:

F-Emotion Algorithm

Input: A sample feature set X={X1, X2, X3,..,Xn}
the sample sum N
the feature classes F
the emotion classes C
Output: F-Emotion

```
Calculate the feature type count Fn |F|;
Calculate the emotion count Cn |C |;
Calculate the sample count of every emotion En |E |;
for y = 1 to Cn do
    Calculate the μ̄
End
for y = 1 to Cn do
  for z = 1 to Fn do
    Calculate the μ_i
  End
End
for z = 1 to Fn do
  for y = 1 to Cn do
    Calculate the A = (μ_i − μ̄)^2 Σ_{j=1}^{m} (μ_i − μ_j)^2
  for y = 1 to Cn do
    for arr = 1 to En do
      Calculate the sum = Σ_{k=1}^{n} (x_ik − μ_i)^2
    End
  End
  B = sum / ( Cn * En )
  F-Emotion = A/B
End
```

TABLE 2. Summary of used datasets.

Dataset	Language	Number of utterances	Number of Emotions	Type of emotions
RAVDESS	English	1440	8	Neutral, Happiness, Sadness, Anger, Fear, Disgust, Surprise, Calm,
EMO-DB	German	535	7	Neutral, Happiness, Sadness, Anger, Fear, Disgust, Boredom

C. PARALLEL DEEP LEARNING MODEL

Most researchers have developed speech emotion recognition models using a single network structure to identify multiple emotions simultaneously. However, these approaches didn't consider the varying levels of contribution that different speech feature results in different emotion recognition outcomes [25], [26].

We proposed a new model that divides the problem into several subchannels. Each subchannel is designed to identify a particular emotion, resulting in n subchannels for a classification task with n emotions. The DNN blocks used in all subchannels have identical structures. See Figure 2.

TABLE 3. The F-Emotion calculation results of speech emotion features on the RAVDESS dataset.

RAVDESS															
Neutral		Happiness		Sadness		Anger		Fear		Disgust		Surprise		Calm	
Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion
Mfcc-std	10.13	Mfcc-std	4.78	Mel-mean	4.79	Mel-mean	367.02	Mfcc-mean	8.78	Cent-mean	3.78	Mfcc-std	6.68	Energy-max	22.77
Mfcc-ptp	9.13	Mfcc-ptp	3.9	Mel-std	4.73	Mel-std	339.12	Mel-min	7.32	Mel-mean	3.12	Mfcc-ptp	4.9	Spl-max	22.3
Mfcc-mean	8.78	Mfcc-min	2.35	Mel-max	4.27	Mel-max	312.24	Mfcc-std	5	Mel-std	2.63	Mfcc-min	4.85	Mfcc-mean	14.14
Energy-max	7.94	Mfcc-max	1.73	Mel-ptp	3.72	Mel-ptp	282.79	Mfcc-max	4.37	Mel-max	2.29	Mfcc-max	3.61	Mfcc-min	11.96
Spl-max	7.73	Mfcc-mean	1.28	Spl-max	2.12	Magnitude-mean	47.25	Energy-max	3.11	Mel-ptp	2.21	Contrast-std	3.06	Mfcc-std	11.8
Mfcc-min	7.03	Spl-max	1.17	Energy-max	2.05	Rms-std	37.67	Spl-max	2.99	Zerocr-std	1.89	Cent-mean	2.73	Mfcc-ptp	11.57
Mel-std	6.56	Energy-max	1.14	Magnitude-mean	1.24	Rms-ptp	36.82	Mfcc-ptp	2.87	Mfcc-mean	1.32	Mel-mean	2.43	Mfcc-max	8.51
Mel-mean	6.55	Contrast-std	1.06	Mfcc-mean	1.21	Rms-max	36.81	Mel-mean	2.54	Mfcc-ptp	1.29	Mel-std	2.1	Mel-mean	6.88
Mel-max	6.02	Fmt_f3-mean	0.7	Mfcc-ptp	1.19	Magnitude-std	36.49	Mel-ptp	2.31	Mfcc-std	1.05	Mel-ptp	1.94	Mel-std	6.85
Mfcc-max	5.34	Mel-mean	0.66	Mfcc-min	1.15	Rms-mean	34.86	Mel-std	2.3	Mfcc-max	0.75	Mel-max	1.94	Mel-max	6.2

TABLE 4. The F-Emotion calculation results of speech emotion features on the EMO-DB dataset.

EMO-DB													
Neutral		Happiness		Sadness		Anger		Fear		Disgust		Boredom	
Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion	Feature	F-Emotion
Mfcc-ptp	29.39	Mfcc-std	66.35	Mel-min	299.95	Mel-mean	89.02	Mfcc-ptp	19.22	Mfcc-mean	19.33	Mfcc-std	39.73
Mfcc-max	27.72	Mfcc-ptp	61.65	Mfcc-std	145.42	Mfcc-min	65.59	Mel-mean	10.45	Mel-std	13.49	Mfcc-ptp	30.40
Mfcc-std	24.68	Mfcc-max	41.46	Mel-ptp	117.33	Mel-std	63.20	Mfcc-std	9.87	Mel-max	13.33	Mfcc-max	27.49
Mfcc-min	18.79	Mfcc-min	36.18	Mfcc-ptp	115.57	Mel-ptp	56.09	Mfcc-min	9.47	Zerocr-ptp	13.01	Mfcc-mean	18.05
Mfcc-mean	17.20	Mfcc-mean	24.03	Fmt_f2-mean	113.98	Mel-max	53.62	Mfcc-max	9.11	Zerocr-max	12.25	Mel-mean	17.46
Mel-mean	14.95	Mel-mean	15.44	Fmt_f1-mean	108.86	Mfcc-ptp	48.61	Mel-min	7.19	Mel-ptp	11.09	Mel-std	13.91
Cent-ptp	10.89	Mel-min	14.67	Mfcc-min	107.77	Mfcc-std	48.16	Mel-std	6.62	Mel-mean	10.53	Mfcc-min	12.14
Mel-std	10.05	Mel-ptp	12.00	Mfcc-mean	97.68	Mfcc-mean	39.50	Contrast-max	6.12	Mfcc-max	9.74	Mel-ptp	11.33
Mel-ptp	9.52	Mel-std	11.67	Mfcc-max	93.40	Mfcc-max	24.28	Mel-ptp	4.97	Mfcc-ptp	7.72	Mel-max	11.12
Cent-std	8.87	Mel-max	10.15	Fmt_b1-mean	74.43	Fmt_f1-std	21.96	Contrast-std	4.64	Mfcc-std	7.20	Mel-min	6.64

For each type of selected speech emotion feature, the F-Emotion value corresponding to each type of emotion was calculated. The feature combination for each emotion was established according to the allocation algorithm. The allocation algorithm distributes feature combinations based on the probability distribution to cover a threshold value of the F-Emotion sum, denoted by th. To determine the optimal feature combination, the speech emotion features corresponding to speech were ranked in descending order by their F-Emotion values, and the top M features were selected. The resulting probability distribution value, denoted as δ, should satisfy coverage threshold th. Assuming that there are N speech emotion features, this formula is used to establish the optimal feature combination for speech emotion recognition.

$$\delta = \frac{\sum_{i=1}^{M} F_i}{\sum_{j=1}^{N} F_j} > th \tag{8}$$

The aforementioned method was used for each subchannel to develop the best feature combination for each emotion type. This feature combination served as the input parameter for training and testing the DNN model in each subchannel, which produced a probability distribution of the emotion recognition results. Subsequently, a decision fusion mechanism was integrated into the backend of the parallel deep learning model to combine the emotion recognition results from all subchannels. In this decision fusion process, a weighted voting mechanism was utilized, with each model's recognition results being given a weight of 10.

III. EXPERIMENT RESULTS

A. DATASET AND SETTING

1) DATASETS

The data came from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [27] and Berlin Emotional Speech Database (EMO-DB) [28]. In the training and testing process, a five-fold cross-validation method was adopted, with 80% of the dataset used as training data and 20% as test data.

The RAVDESS dataset is an English language dataset consisting of two modalities: audio and audio-video. In this study, only 1440 audio files were used. The audio samples were categorized into eight emotions and recorded by 24 professional actors (12 male and 12 female). In addition, each expression was recorded at two levels of emotional intensity. The approximate time of the utterances in RAVDESS was three-five seconds with a 48 kHz sampling rate.

The EMO-DB dataset is a German language dataset. It was recorded by ten experienced actors and included 535 utterances with seven emotions. The dataset contained five male and five female actors who read predetermined sentences to express different emotions. The approximate time of the utterances in the EMO-DB was three-five seconds with a 16 kHz sampling rate. Further details are provided in Table 2.

2) SETTING

The DNN model comprises an input layer, hidden layers, and an output layer with the incorporation of batch normalization and activation layers. The initial input layer of the DNN model uses a 312-dimensional numerical array as input data and is composed of 256 filters that undergo batch normalization. Dropout was applied at a rate of 0.15, and the output was activated using a rectified linear unit (ReLU). The subsequent layer consists of 64 filters that receive the output from the preceding input layer. Dropout was applied at a rate of 0.3, and the output was also activated by ReLU. The DNN model utilizes the Adam optimizer with a learning rate of 0.001 and a decay rate of 1e-3. The threshold value for Eq. (8) was set to 93%.

The experiment was carried out on Windows 7, where the computer hardware was configured as an Intel i7 CPU at 2.80 GHz, with 16 GB of memory. Python version 3.8 was used as the programming language. The running process of the program relied mainly on the CPU for calculation. GPU is not used. The runtimes are approximately 10 min and 7 min on the RAVDESS and EMO-DB datasets, respectively. The CPU resource usage is approximately 60%.

B. F-EMOTION CALCULATION RESULTS

The speech emotion features for each emotion category were calculated using F-Emotion on the RAVDESS and EMO-DB datasets. The resulting scores were sorted in descending order. The results were presented in Tables 3 and 4, respectively.

Based on the computation results, the MFCC feature had the most significant impact on emotion categories, such as neutral, happiness, fear, surprise, and boredom. The Mel frequency spectrum (Mel) feature had the most significant impact on emotion categories, such as sadness and anger. Speech energy and sound pressure level (SPL) had the greatest influence on the calm emotion category, while the spectrum centroid had the greatest influence on the disgust emotion category.

The Mel frequency spectrum is based on the characteristics of human auditory perception and corresponds nonlinearly to the frequency. MFCC is calculated by applying a discrete cosine transform (DCT) to the Mel frequency spectrum. In this study, 50 coefficients were used for the MFCC, while 128 band parameters were used for the Mel frequency spectrum. Compared with the Mel frequency spectrum, MFCC features have lower inter-correlations and higher discriminative power, which is advantageous for linear models. Therefore, the MFCC features have relatively high F-Emotion values for most emotion categories. The spectrum centroid describes the brightness of sound, with a darker and deeper sound quality in the lower frequency range and a more cheerful sound quality in the higher frequency range. Calm emotion speech usually has low amplitude and frequency, low speech energy, and low sound pressure level, which distinguishes it from other speech emotion features and results in

TABLE 5. Speech emotion recognition results from 8-channel on the RAVDESS dataset.

	Neutral	Happiness	Sadness	Anger	Fear	Disgust	Surprise	Calm	Accuracy
Channel 1- Neutral	**75%**	71%	66%	76%	85%	76%	87%	90%	
Channel 2- Happiness	65%	**74%**	74%	79%	82%	71%	79%	87%	
Channel 3- Sadness	70%	63%	**87%**	76%	72%	71%	82%	92%	
Channel 4- Anger	70%	68%	76%	**84%**	87%	79%	74%	87%	
Channel 5- Fear	75%	71%	66%	76%	**85%**	76%	87%	90%	
Channel 6- Disgust	75%	68%	66%	84%	79%	**79%**	84%	85%	
Channel 7- Surprise	75%	63%	63%	76%	74%	79%	**87%**	97%	
Channel 8- Calm	75%	61%	66%	68%	87%	76%	84%	**97%**	
Decision-level Fusion	75%	71%	79%	87%	85%	74%	89%	95%	**82.3%**

TABLE 6. Prediction performance of the proposed model in terms of Precision, Recall, F1_score, Support, Weighted score, and Un-weighted score on the RAVDESS dataset.

Classes	Precision	Recall	F1_score	Support
Neutral	0.64	0.80	0.71	20
Happiness	0.81	0.68	0.74	38
Sadness	0.79	0.71	0.75	38
Anger	0.76	0.84	0.80	38
Fear	0.85	0.87	0.86	39
Disgust	0.82	0.71	0.76	38
Surprise	0.81	0.89	0.85	38
Calm	0.93	0.95	0.94	39
macro avg	0.81	0.82	0.81	288
micro avg	0.82	0.82	0.82	288
Weight Acc	0.82	0.82	0.82	288
Un-Weight Acc	0.80	0.80	0.80	288

high F-Emotion values. Furthermore, an algorithm for feature combination allocation was used to establish the optimal feature combination for different emotion categories. This algorithm assigns feature combinations that cover a threshold value of *th* for the total F-Emotion score, based on the principle of probability distribution.

C. SPEECH EMOTION RECOGNITION RESULTS

Table 5 presents the recognition results for each emotion category using eight channels on the RAVDESS dataset. The recognition rate for fear was 85%, which was lower than the rate for the calm channel at 87%, while the recognition results for the other seven emotion types corresponding to their respective channels were all optimal. After applying decision fusion, the final overall recognition rate was 82.3%.

Table 6 presents the predictive performance of the system, including metrics such as Precision, Recall, F1_score, and Support. The model achieved macro accuracy of 82%, micro, weighted, and unweighted accuracies of 82%, 82% and 80%, respectively. Figure 3 shows the confusion matrix of the recognition results, displaying the actual and predicted labels for each emotion. The model demonstrated recognition rates of over 85% for calm, angry and surprised emotions. However, it had a lower precision for happy than for the other emotion types.

The recognition results for seven channels in the EMO-DB dataset are listed in Table 7. The recognition rate for the neutral channel was 94%, which was lower than the 100% recognition rate for the disgust channel. The recognition results for the other six channels for the corresponding

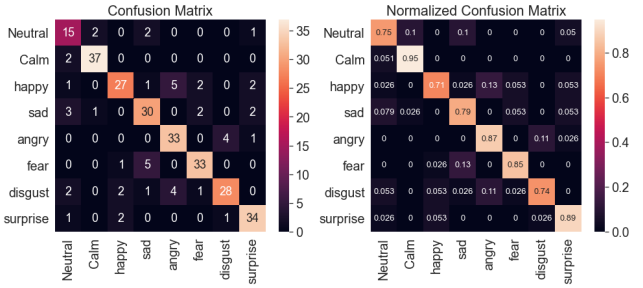

FIGURE 3. Confusion matrix on the RAVDESS dataset.

emotion types were optimal. After decision fusion, the final overall recognition result was 88.8%.

Table 8 presents the predictive performance of the system, including metrics such as Precision, Recall, F1_score, and Support. The model achieved macro accuracy of 90%, micro accuracy of 89%, a weighted accuracy of 89% and 90%, respectively. Figure 4 shows the confusion matrix of the recognition results, showing the actual and predicted labels for each emotion. The recognition rates of this model for neutral emotion, happiness, sadness, anger, fear, disgust, and boredom were 94%, 64%, 100%, 88%, 93%, 100% and 88% respectively.

D. ABLATION STUDY

We demonstrated the necessity of the F-Emotion algorithm and parallel deep learning model through ablation experiments. For verifying the validity of our proposed module, we removed the F-Emotion algorithm and parallel deep learning model. The experimental results are presented in Tables 9 and 10, respectively.

Table 9 compares the performance of our model with the experimental results when adding either the F-Emotion algorithm or the parallel deep learning model to the RAVDESS dataset. The recognition accuracy of a single network without F-Emotion was 70%. The parallel deep learning model alone improved the accuracy by 7.0%, while the F-Emotion algorithm alone improved the accuracy by 8.0%. Finally, our proposed methodology improves the accuracy by 12.0%.

Table 10 compares the performance of our model with the experimental results when adding either the F-Emotion algorithm or parallel deep learning model to the EMO-DB dataset. The recognition accuracy of a single network without

TABLE 7. Speech emotion recognition results from 8-channel on the EMO-DB dataset.

	Neutral	Happiness	Sadness	Anger	Fear	Disgust	Boredom	Accuracy
Channel 1- Neutral	**94%**	64%	92%	96%	86%	89%	94%	
Channel 2- Happiness	75%	**86%**	100%	92%	86%	100%	100%	
Channel 3- Sadness	94%	64%	**100%**	88%	86%	100%	88%	
Channel 4- Anger	94%	64%	100%	**96%**	86%	67%	82%	
Channel 5- Fear	94%	57%	92%	92%	**93%**	89%	71%	
Channel 6- Disgust	100%	71%	100%	96%	86%	**100%**	82%	
Channel 7- Boredom	94%	86%	100%	84%	79%	78%	**100%**	
Decision-level Fusion	94%	64%	100%	88%	93%	100%	88%	**88.8%**

TABLE 8. Prediction performance of the proposed model in terms of Precision, Recall, F1_score, Support, Weighted score, and Un-weighted score on the EMO-DB dataset.

Class	Precision	Recall	F1-score	Support
Neutral	0.94	0.94	0.94	16
Happiness	0.82	0.64	0.72	14
Sadness	1.00	1.00	1.00	12
Anger	0.79	0.88	0.83	25
Fear	1.00	0.93	0.96	14
Disgust	0.82	1.00	0.90	9
Boredom	0.94	0.88	0.91	17
macro avg	0.90	0.90	0.89	107
micro avg	0.89	0.89	0.89	107
Weight Acc	0.89	0.89	0.89	107
Un-Weight Acc	0.90	0.90	0.89	107

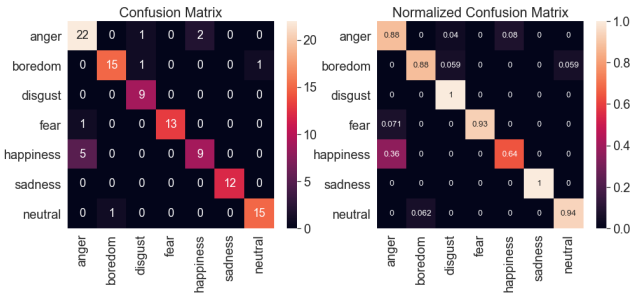

FIGURE 4. Confusion matrix on the EMO-DB dataset.

TABLE 9. Experimental results of ablation study on the RAVDESS dataset.

F-Emotion	Parallel Model	Weighted Precision	Weighted Recall	Weighted F1-score	Acc
×	×	0.70	0.70	0.69	0.70
×	✓	0.78	0.77	0.77	0.77
✓	×	0.79	0.78	0.78	0.78
✓	✓	0.82	0.82	0.82	0.82

TABLE 10. Experimental results of ablation study on the EMO-DB dataset.

F-Emotion	Parallel Model	Weighted Precision	Weighted Recall	Weighted F1-score	Acc
×	×	0.85	0.83	0.82	0.83
×	✓	0.87	0.87	0.87	0.87
✓	×	0.85	0.84	0.84	0.84
✓	✓	0.89	0.89	0.89	0.89

F-Emotion was 83%. The parallel deep learning model alone improved the accuracy by 4.0%, while the F-Emotion algorithm alone improved the accuracy by 1.0%. Finally, our proposed methodology improves the accuracy by 6.0%.

Experimental results show that the F-Emotion algorithm and parallel deep learning model can improve the accuracy of speech emotion recognition.

IV. DISCUSSION

This study introduces a new method for analyzing speech emotion features and a novel mechanism for emotion recognition. We developed the F-Emotion algorithm to calculate the weight of each speech emotion feature to select the optimal speech emotion feature combination. A parallel deep learning model was established, where the optimal speech emotion feature combination was used as the input, and the recognition results for each emotion category were output separately. Finally, a voting mechanism is implemented for decision fusion, which yields overall recognition results. The evaluation was conducted on two datasets: RAVDESS and EMO-DB. The results showed that the proposed method achieved a recognition performance that surpassed the results reported in previous studies.

TABLE 11. Comparison of recognition results in previous work and in the proposed method on the RAVDESS dataset.

Method	Dataset	Model	Overall Recognition Rate
Prabhav Singh et al. (2021)	RAVDESS	DNN	81.2%
Ftoon Abu Shaqra et al. (2019)	RAVDESS	Multi-layer perceptron	74.0%
Mustaqeem et al. (2019)	RAVDESS	DSCNN	79.5%
Nivedita Patel et al. (2021)	RAVDESS	CNN	80.0%
Hemin Ibrahim et al. (2022)	RAVDESS	parallel ESN	75.3%
Proposed	**RAVDESS**	**Multi-DNN**	**82.3%**

Table 11 presents a comparison of the speech emotion recognition results in our study and those in previous studies that also used the RAVDESS dataset [17], [29], [30], [31], [32] and the present study. Singh et al. [16] proposed a hierarchical deep learning-based approach that employs hierarchical DNN models and achieved an accuracy of 81.2%. Shaqra et al. [29] proposed a hierarchical classification model using the eGeMAPS feature set, achieving 74.0% accuracy. Kwon et al. [30] proposed an AI-assisted deep-stride convolutional neural network (DSCNN) architecture that learned salient and discriminative features from the

spectrogram of speech signals. Tests of the network showed an improved recognition performance. Patel et al. [31] used a traditional auto-encoder to reduce audio dimensionality, and achieved an 80.0% recognition rate using a Support Vector Machine (SVM), decision tree, and CNN models. Ibrahim et al. [32] recently proposed a novel bidirectional reservoir computing model by adopting two parallel reservoirs when the same direction output from the different reservoirs is fused together, achieving 75.3% accuracy. In comparison, the proposed method achieved an accuracy of 82.3% on the RAVDESS dataset, surpassing the best performing approach by Prabhav [17] by 1.1%.

TABLE 12. Comparison of recognition results in previous work and in the proposed method on the EMO-DB dataset.

Method	Dataset	Model	Overall Recognition Rate
Sajjad, M. et al.(2020)	EMO-DB	BiLSTM	85.6%
Pengxu Jiang et al. (2019)	EMO-DB	PCRN	84.5%
Dias Issa et al.(2020)	EMO-DB	CNN	86.1%
Premjeet Singh et al.(2023)	EMO-DB	DNN-SVM	79.8%
Dongdong Li et al.(2021)	EMO-DB	BLSTM-DSA	85.9%
Proposed	**EMO-DB**	**Multi-DNN**	**88.8%**

Table 12 shows the speech emotion recognition results of our study on the EMO-DB dataset and the results of previous studies [33], [34], [35], [36], [37]. Sajjad et al. [33] proposed an algorithmic transformation strategy that extracted distinctive and salient features from spectrograms, and used a deep bidirectional long short-term memory (BiLSTM) network to learn and recognize long-term sequences in audio data. Their method achieved a recognition accuracy of 85.6% on the EMO-DB dataset. Jiang et al. [34] introduced a parallel convolutional recursive neural network (PCRN) that utilized spectral features for speech emotion recognition, and achieved an accuracy of 84.5% on the EMO-DB dataset. Issa et al. [35] directly extracted features from raw audio files and used an incremental approach to modify the initial CNN model, achieving a recognition accuracy of 86.1% on the EMO-DB dataset. Singh et al. [36] explored the use of constant-Q transform based modulation spectral features (CQT-MSF) and used a DNN-SVM framework, achieving a recognition accuracy of 79.8% on the EMO-DB dataset. Li et al. [37] proposed a new deep network architecture, bidirectional long short-term memory with directional self-attention (BLSTM-DSA). The proposed algorithm automatically annotates the weights of frames using a self-attention mechanism to improve the efficiency of SER, achieving a recognition accuracy of 85.9%.

Our proposed method achieved the highest recognition accuracy of 88.8% on the EMO-DB dataset, which was 2.7% higher than that of the best-performing method by Issa et al. [35].

V. CONCLUSION

In this study, we proposed an F-Emotion feature selection algorithm to calculate the F-Emotion value of the extracted speech emotion features for each emotion category. Based on the F-Emotion value, we analyzed the weight of each speech emotion feature, or feature combination for each emotion classification. We then determined the optimal combination of speech emotion features for each emotion type. Based on this approach, we established a parallel deep learning model mechanism that inputs the optimal speech emotion feature combination for each channel and assigns different weights to the output emotion classification. A voting mechanism was then used to obtain the overall emotion recognition results. We evaluated our approach on two datasets, RAVDESS and EMO-DB, and achieved accuracy rates of 82.3% and 88.8%, respectively. These results demonstrate that using the F-Emotion feature selection algorithm and parallel deep learning model mechanism can improve the accuracy of speech emotion recognition.

APPENDIX

Chroma_cens	The normalization of chromatographic energy, which converts the speech signal into the corresponding spectrogram and performs normalization processing.
Energy	The loudness of the sound, also known as volume.
Formants	Frequencies produced by physical vibrations of objects that do not change in pitch.
Mel	Mel spectrogram. The speech signal is converted into the corresponding spectrogram, the data on which are utilized as the feature of the signal.
MFCC	Cosine transform is performed after the Mel spectrogram is obtained, and some of the coefficients are taken.
Pitch	The vibration frequency of the vocal cords.
Shimmer abs	The absolute value of shimmer. Shimmer describes the change of sound wave amplitude between adjacent periods, mainly reflecting the degree of hoarseness.

Short-time energy	The sum of the squares of the amplitude values of the frame speech signal.
Sound pressure level	The pressure level of a sound. Take the common logarithm of the ratio of the sound pressure to be measured p to the reference sound pressure p(ref) and multiply it by 20. The unit is decibels.
Spectral contrast	The centroid of the spectrum.
Zero-crossing rate	The number of times the speech signal passes through the zero point (from positive to negative or from negative to positive) in each frame.

ABBREVIATIONS

BiLSTM	bidirectional long short-term memory
BLSTM-DSA	Bi-directional Long-Short Term Memory with Directional Self-Attention
CQT-MSF	Constant-Q transform based modulation spectral features
DCNNs	deep convolutional neural networks
DCT	discrete cosine transform
DNN	deep neural network
DSCNN	deep strides convolutional neural network
ESN	Echo state network
Max	maximum
Mean	average value
MFCC	mel-frequency cepstrum coefficient
Min	minimum
PCRN	parallel convolutional recursive neural network
Ptp	Range: The difference between the maximum and minimum values
SER	speech emotion recognition
SPL	sound pressure level
STD	standard deviation
SVM	support vector machine
ZCR	zero-crossing rate

ACKNOWLEDGMENT

This Ph.D. research was conducted at Universiti Malaysia Sabah (UMS), Malaysia. The authors gratefully thank Universiti Malaysia Sabah for there support to publish this article.

REFERENCES

[1] D. Crystal, "Non-segmental phonology in language acquisition: A review of the issues," *Lingua*, vol. 32, nos. 1–2, pp. 1–45, Sep. 1973.

[2] R. Matin and D. Valles, "A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions," in *Proc. Intermountain Eng., Technol. Comput. (IETC)*, Oct. 2020, pp. 1–6.

[3] J. Liu, X. Wu, and X. Wu, "Prototype of educational affective arousal evaluation system based on facial and speech emotion recognition," *Int. J. Inf. Educ. Technol.*, vol. 9, no. 9, pp. 645–651, 2019.

[4] H. Nasri, W. Ouarda, and A. M. Alimi, "ReLiDSS: Novel lie detection system from speech signal," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2016, pp. 1–8.

[5] P. Laukka and P. N. Juslin, "Similar patterns of age-related differences in emotion recognition from speech and music," *Motivat. Emotion*, vol. 31, no. 3, pp. 182–191, Sep. 2007.

[6] 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, Mar. 2011.

[7] I. Luengo, E. Navas, and I. Hern, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[8] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human–robot interaction," *Inf. Sci.*, vol. 509, pp. 150–163, Jan. 2020.

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

[10] H. A. Abdulmohsin, "A new proposed statistical feature extraction method in speech emotion recognition," *Comput. Electr. Eng.*, vol. 93, pp. 107–172, Jul. 2021.

[11] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, p. 6008, Oct. 2020.

[12] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.

[13] Y. Ü. Sönmez and A. Varol, "A speech emotion recognition model based on multi-level local binary and local ternary patterns," *IEEE Access*, vol. 8, pp. 190784–190796, 2020.

[14] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, Feb. 2015.

[15] N. Poh and S. J. I. Bengio, "F-ratio client dependent normalisation for biometric authentication tasks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. I/721–I/724.

[16] Z. Chen, H. Wang, S. Hyon, J. Wei, and J. Dang, "Efficient feature extraction of speaker identification using phoneme mean F-ratio for Chinese," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, Dec. 2012, pp. 345–348.

[17] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107316.

[18] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Multi-stage classification of emotional speech motivated by a dimensional emotion model," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 119–145, Jan. 2010.

[19] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex Intell. Syst.*, vol. 2021, pp. 1–10, Aug. 2021.

[20] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," *Appl. Acoust.*, vol. 182, Nov. 2021, Art. no. 108260.

[21] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, Dec. 2019.

[22] A. Mahdhaoui, M. Chetouani, and C. Zong, "Motherese detection based on segmental and supra-segmental features," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[23] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2007, pp. IV-17–IV-20.

[24] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[25] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Commun.*, vol. 120, pp. 11–19, Jun. 2020.

[26] A. L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C. T. Lengerich, D. Jurafsky, and A. Y. Ng, "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017.

[27] 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*, vol. 13, no. 5, May 2018, Art. no. e0196391.
[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, vol. 5, Sep. 2005, pp. 1517–1520.
[29] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Proc. Comput. Sci.*, vol. 151, pp. 37–44, Jan. 2019.
[30] S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.
[31] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *J. Ambient Intell. Humanized Comput.*, vol. 2022, pp. 1–19, Feb. 2022.
[32] H. Ibrahim, C. K. Loo, and F. Alnajjar, "Bidirectional parallel echo state network for speech emotion recognition," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 17581–17599, Oct. 2022.
[33] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[34] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.
[35] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.
[36] P. Singh, M. Sahidullah, and G. Saha, "Modulation spectral features for speech emotion recognition using deep neural networks," *Speech Commun.*, vol. 146, pp. 53–69, Jan. 2023.
[37] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114683.

LI-MIN ZHANG received the M.Sc. degree from the School of Computer Science and Technology, Xidian University, in 2011. She is currently pursuing the Ph.D. degree in computer science with the Faculty of Computing and Informatics, Universiti Malaysia Sabah.

In 2019, she was a Visiting Scholar with the University of Leicester, U.K. Since 2021, she has been an Associate Professor with the Key Laboratory for Artificial Intelligence and Cognitive Neuroscience of Language, Xi'an International Studies University. Her research interests include speech emotion recognition, deep learning, and big data analytics.

GIAP WENG NG received the B.Sc. degree (Hons.) in cognitive sciences from UNIMAS, and the Ph.D. degree in computer science from Salford University, Manchester, U.K.

From 2007 to 2018, he was with the Faculty of Cognitive Sciences and Human Development. He was the Director of the Centre of Excellence for Semantic Technology and Augmented Reality (CoESTAR), UNIMAS. He is currently a Professor with the Faculty of Computing and Informatics and also the Director of the Centre for E-Learning, Universiti Malaysia Sabah (UMS). His areas of research interests include augmented reality (AR), virtual reality (VR), interactive technologies and artificial intelligence (AI), blockchain, and the Internet of Things (IoT).

YU-BENG LEAU (Senior Member, IEEE) received the B.Sc. degree in multimedia technology from Universiti Malaysia Sabah, the Master of Computer Science degree in information security from Universiti Teknologi Malaysia, and the Ph.D. degree in internet infrastructures security from Universiti Sains Malaysia.

He is currently a Senior Lecturer with the Faculty of Computing and Informatics, Universiti Malaysia Sabah. His areas of interests include cybersecurity, network management and situational awareness, IPV6 security, the Internet of Things, and future network architectures.

HAO YAN was born in Jiangsu, China, in 1980. He received the Bachelor of Arts degree in English from Xidian University, in 2003, the Master of Arts degree in foreign linguistics and applied linguistics from Northwestern Polytechnical University, in 2006, and the Ph.D. degree in cognitive psychology from Shaanxi Normal University, in 2013.

From 2015 to 2017, he was a Postdoctoral Fellow with the Health Science Center, School of Biomedical Engineering, Shenzhen University. Currently, he is a Professor and the Founding Director of the Key Laboratory for Artificial Intelligence and Cognitive Neuroscience of Language, Xi'an International Studies University. His research interests include cognitive neuroscience of language, natural language processing, and speech rehabilitation.