scientific reports

Check for updates

OPEN

Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC)

Yunhao Zhao[1,2]✉ & Xiaoqing Shu[1,2]

Speech emotion analysis is one of the most basic requirements for the evolution of Artificial Intelligence (AI) in the field of human–machine interaction. Accurate emotion recognition in speech can be effective in applications such as online support, lie detection systems and customer feedback analysis. However, the existing techniques for this field have not yet met sufficient development. This paper presents a new method to improve the performance of emotion analysis in speech. The proposed method includes the following steps: pre-processing, feature description, feature extraction, and classification. The initial description of speech features in the proposed method is done by using the combination of spectro-temporal modulation (STM) and entropy features. Also, a Convolutional Neural Network (CNN) is utilized to reduce the dimensions of these features and extract the features of each signal. Finally, the combination of gamma classifier (GC) and Error-Correcting Output Codes (ECOC) is applied to classify features and extract emotions in speech. The performance of the proposed method has been evaluated using two datasets, Berlin and ShEMO. The results show that the proposed method can recognize speech emotions in the Berlin and ShEMO datasets with an average accuracy of 93.33 and 85.73%, respectively, which is at least 6.67% better than compared methods.

Speech is one of the most basic means of human interaction, which can implicitly convey the feelings of the speaker to the listener. With the expansion of AI techniques, human–machine interaction systems have also been developed, and these systems must be able to process speech for improving interaction[1]. By analyzing the emotions hidden in speech, the feedback of intelligent systems can be improved and the information obtained from it can be used to make the output of these systems understandable by people[2,3]. Sentiment extraction can be useful in areas such as online support, lie detector systems, customer feedback analysis, and the like[4]. For this reason, achieving an accurate emotion extraction system can benefit from various practical aspects. Nevertheless, a significant part of the research in the field of sentiment analysis has focused on text processing, and speech analysis has received less attention[5]. The reason for this can be the higher complexity of speech analysis systems and the significant impact of noise (or background sounds) on system performance.

Achieving an accurate system for recognizing speech emotions requires solving several challenges. The complexity of emotional states is one of the first challenges. Because an emotional state can include a combination of several basic emotional states, and for this reason, recognizing basic emotions in speech is of considerable importance[6]. On the other hand, the extraction of speech features should be performed efficiently so that the relationship between verbal patterns and different emotional states can be interpreted based on it. The relatively large number of emotional states is another existing challenge that causes most learning techniques to be unable to correctly classify the speech's emotional features[7]. In this paper, a hybrid solution is presented to solve the main challenges in the problem of speech emotions analysis. The difference between the current research and previous similar works can be investigated from feature extraction and classification. This method uses modulation analysis and deep learning techniques to extract speech features and it presents a new classification model based

[1]Department of Chinese Language & Literature, The Catholic University of Korea, 43 Jibong-Ro, Gyeonggi-Do, Bucheon-Si 14662, Republic of Korea. [2]These authors contributed equally: Yunhao Zhao and Xiaoqing Shu. ✉email: zhaoyunhao1994@126.com

on the combination of ECOC and GC to overcome the challenge of, a large number of target classes. With these explanations, the contribution of the current paper can be summarized in the following cases:

- The proposed method in this article uses the combination of STM and entropy features of the speech signal to describe the emotional features, and the dimensions of these features are reduced by using a CNN. The features extracted through this CNN contain the most relevant features with emotional states.
- In this paper, a new classification model based on the combination of ECOC and GC is presented, which can be effective in solving classification problems with a large number of target classes. This classification model includes several GCs that are trained based on the ECOC matrix.

Based on the current knowledge of the authors, these two methods have not been used in previous research and can be considered as new approaches to fill the existing research gap. The rest of the paper is organized as follows: In section "Research background", the research background is studied. In section "Research method" the description of the proposed method is provided, and in section "Results", the findings of the research are discussed. Finally, sections "Discussion" and "Conclusion" contain the discussions and summary of the research findings, respectively.

Research background

In[8], speech emotion recognition was done with Artificial Neural Network (ANN) and Support Vector Machine (SVM). Since the effect of feature dimension reduction has been carefully evaluated, the effect of dimension reduction on these two models was compared. These features from the CASIA Chines Emotional Corpus dataset were extracted. In[9], the Deep Neural network (DNN) classification method was applied to a custom dataset to recognize speech emotions. Only the Mel Frequency Cepstral Coefficients (MFCC) feature was considered for testing in this research.

In[10] by using the Multiple Linear Regression (MLR) classification method, and with the combination of MFCC and MS features, seven emotions anger, hatred, happiness, fear, surprise, sadness, and normal state were reported with a recognition accuracy of 82.41%.

In[11], the MFCC feature (which is widely used to analyze any speech signal) was employed to recognize emotions in speech. This set of features, which compared to other features, performed well for speech emotion recognition systems, includes 39 coefficients. In this research, Long Short-Term Memory (LSTM) was applied to recognize emotions. A multifaceted self-care method was proposed in[12], which implements five convolution layers and one attention layer using MFCC for emotion recognition. This method applies functional descriptors to the output of the MFCC signal to increase recognition accuracy by combining them with features based on parts. This model is trained for the IEMOCAP dataset and considers only four emotions: happy, sad, angry, and neutral.

In[13], DNN and Hidden Markov Model (HMM) classification methods using two MFCC and Epoch-based features were compared for four emotional states: happy, sad, angry, and normal in the IEMOCAP dataset. In[14], emotion recognition was performed on four datasets, including the ShEMO dataset. The features used in this research are MFCC, formant, and prosodic features (such as the lowest and highest pitch and other pitch features) as well as a combination of these three features. However, in this paper, only the recognition of two emotional states, neutral and anger, was tested.

In[15], the aggressiveness of a person's speech and state through mobile phone voice was discussed. It examines the feelings of anger and fears that cause stress from the speech available in various datasets such as BERLIN EMO-DB and recognizes the stress from the person's speech and expression. In this method, 13 MFCC-based features, 7 functional features, and 4 low-level features are used to describe the features of each speech signal. Then, SVM was used to classify the features. In[16], emotion recognition in the speech was performed on six datasets, including ShEMO, and it was shown that the classification method can be made sensitive and dependent on a small number of phonetic labels that are clustered by the K-Means method and other feature components can be ignored. These specific phonetic components are taken from the MFCC feature and recognize speech emotions with DNN and SVM classification methods. The advantages of this reduction in dimensions are in reducing calculations, execution time, and cost, but; It does not achieve high accuracy.

In addition to speech, sentiment analysis can be done using other types of data. For example, in[17], the issue of emotion evaluation with interaction levels in blended learning is discussed. In[18], EEG brain signals are used to detect emotions. The methods presented in[19] and[20] have analyzed sentiments in texts. Also, the method presented in[21] discusses the topic of emotion recognition in long dialogues.

Research method

In this section, after describing the specification of datasets used in the research, the details of the proposed method for emotion analysis in speech signal of these datasets is presented.

Data acquisition

In this research, the speech signals available in ShEMO[22] and Berlin Emotional[23] datasets were used as input for experiments. In the following, the specifications of the mentioned datasets are described first, and then the obtained results are analyzed.

The data set used from the Berlin dataset is 535 speech signals in six emotional categories, and a subset including 150 speech signals from this set was used in the proposed method. The selected speech signals are without background and the speakers include both male and female genders. The emotional states in this dataset are 1- Anger (31 samples), 2- Hatred (25 samples), 3- Fear (29 samples), 4- Joy (28 samples), and 5- Sadness (37

samples). The average length of speech signals in this dataset is equal to 3.051 s. Also, all samples of this dataset were recorded through two speakers (one male and one female). On the other hand, the ShEMO dataset has 3000 speech signals, all of which were collected from radio shows. This dataset's total sample length is 3 h and 25 min. This dataset was gathered through 78 Persian speakers (male and female) and includes six basic emotions, which are: 1- Anger (1059 samples), 2- Fear (38 samples), 3—happiness (201 samples), 4- neutral state (1028 samples), 5- sadness (449 samples) and 6- surprise (225 samples). Due to the shortness of a significant number of samples in this dataset, all samples with a length of less than one second were ignored. Following this process, 874 samples were removed from the dataset, and as a result, the experiments conducted in this research were performed on 2126 samples from the ShEMO dataset. In the experiments, the cross-validation technique was used with 10 iterations, and in each iteration, 90 samples were used for training the learning models and the remaining 10% were used for testing.

The proposed method for emotion analysis in speech

A model for recognizing speech emotion can generally include pre-processing, feature extraction, and classification steps. Each of these processes is of great importance and the resulting combination should be able to efficiently extract emotional states with the highest possible accuracy. In the proposed method, the recognition of speech emotion is done using the following calculation steps (Fig. 1):

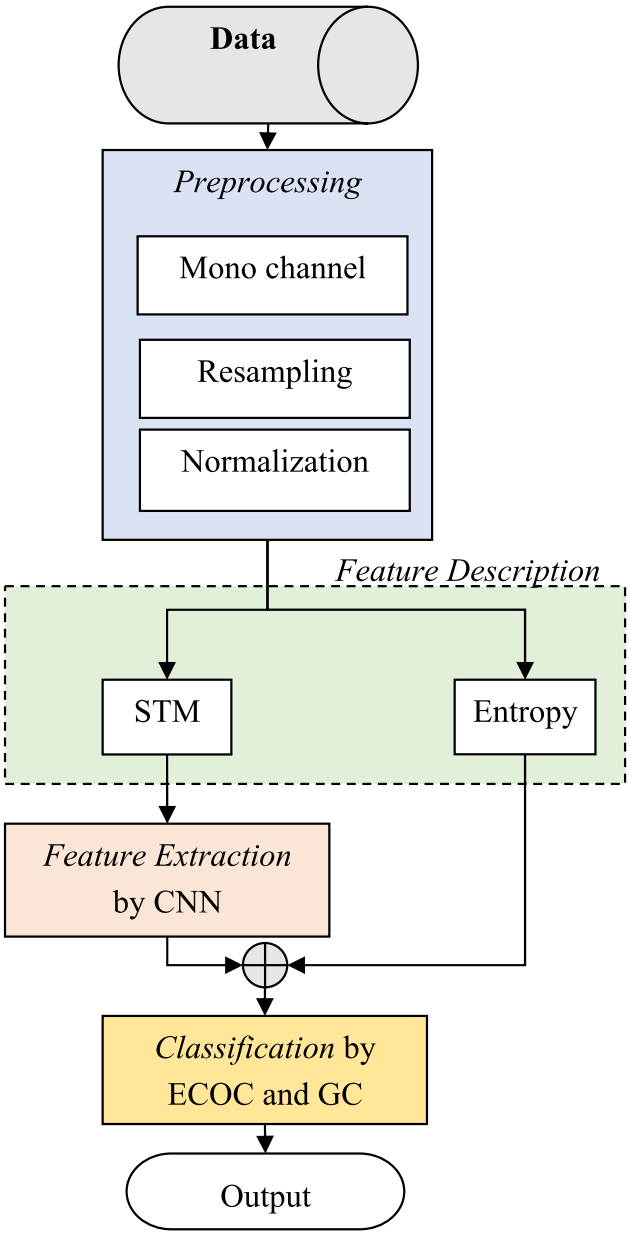

Figure 1. Flowchart of the proposed method.

1. Preprocessing speech signal
2. Feature description based on spectro-temporal modulation and entropy features
3. Extracting features using CNN
4. Classification based on GC and ECOC

According to Fig. 1, the proposed method starts with preprocessing speech signals including converting to mono-channel, converting frequency by sampling, and signal normalization. In the second step, feature description is performed using two sets of features: entropy and spectro-temporal modulation. These two techniques process the normalized signal independently and extract features describing the signal. The features extracted through spectro-temporal modulation are reduced by using a CNN in the third step of the proposed method and are extracted as a vector containing significant emotional features in speech. This vector is merged with entropy features to recognize the emotion in the speech in the fourth step of the proposed method. For this purpose, a detection model based on the combination of ECOC and GC is used. This classification model includes a set of GCs that are trained based on the ECOC code matrix. In the rest of this section, the details of each step of the proposed method are described.

Preprocessing speech signal
The proposed method starts with preprocessing the input signals. This step includes three main steps:

- Converting to mono-channel
- Converting signal frequency
- Signal normalization

The purpose of the pre-processing step is to remove redundant data from the audio signal, convert all signals into a standard intermediate form and prepare them for use in the next steps. For this purpose, the processes of audio signal normalization, audio signal sampling, and transformation of two-channel signals into mono-channel vector form are used. In this way, at the beginning of the pre-processing step, the two-channel nature of the input audio signal is checked. If the signal is two (or more)-channel, we convert it to a mono-channel signal. Since audio signals are recorded in different conditions and by different devices; Therefore, the frequency of some input signals may be different from other signals. For this reason, in the second step of the pre-processing phase, all the vectors of the input signal are converted to the same frequency of 16 kHz. At the end of the pre-processing step, the signal vector is converted into a vector with zero mean and unit variance, so that the special conditions of a signal (high or low volume) are eliminated as much as possible.

Feature description based on spectro-temporal modulation and entropy features
In the second step, each speech signal is described with entropy features and spectro-temporal modulation. Each of these two techniques processes the pre-processed signal independently and extracts the descriptive features of the signal, which are described below.

Extraction of entropy features. The first feature set used to describe speech signal features is entropy features. In the proposed method, two types of features, approximate entropy, and sample entropy, are used to describe the general characteristics of each speech signal. Approximate entropy can describe speech signals based on the features of certainty, chaos, or randomness. A high approximate entropy value indicates disorder and a high level of chaos in the signal, while a low approximate entropy value indicates repetitive changes in the signal. On the other hand, sample entropy is a modified version of approximate entropy that can be used to evaluate the complexity of physiological and speech time series signals. Compared to approximate entropy, this criterion has two advantages: independence from the length of the data and relative implementation without problems. In[24] calculation of these entropy values is explained; Therefore, it is omitted to deal with the details of the calculation of these features.

Feature description based on spectro-temporal modulation. The second set of features utilized to describe audio signal features is spectro-temporal modulation. This technique also performs the feature description process using auditory modeling and includes two basic processing steps:

- Modeling the human auditory system
- Generating temporal modulation based on the Auditory Spectrogram (AS)

In the first step, the human auditory system is modeled, during which the speech signal is converted into a neural pattern, i.e., AS.

AS is a time–frequency distribution along the tonotopic axis or logarithmic frequency, which is obtained by applying three stages of transformation on the input signal. In the second stage, the temporal modulation content is obtained through AS and by applying wavelet transformation to each line of the AS.

Modeling the human auditory system: The process of modeling the human auditory system consists of three main steps, which are based on the initial stages of sound processing by humans. In the first step, a constant Q transformation is applied to the input signal. This transformation is done using the filter bank, in all of the filters, the ratio of the central frequency and resolution is always a constant value. In the proposed method, 96

overlapped filters are used, whose central frequencies are linearly and uniformly distributed. To properly distribute these filters, the logarithmic frequency axis is divided into the following four octave ranges:

- Octave 1: 100 to 200 Hz
- Octave 2: 200 to 400 Hz
- Octave 3: 400 to 800 Hz
- Octave 4: 800 to 1600 Hz

All the 96 mentioned filters are distributed on the logarithmic frequency axis in such a way that they can cover these four octaves. If we consider f as the logarithmic frequency of this filter bank, then the impulse response of each filter can be expressed as $h_{cochlea}(t,f)$. Considering the impulse response caused by each filter and s(t) as the input speech signal, the output of the cochlear filter can be calculated as follows[25]:

$$y_{\text{cochlea}}(\text{t},\ \text{f}) = \text{s(t)} *_{\text{t}} \text{h}_{\text{cochlea}}(\text{t},\ \text{f}) \tag{1}$$

where *t represents a twist in the time domain. In this way, by calculating the output of the cochlear filter, the first stage of modeling the auditory system is completed. In the second step, the output obtained from the previous step (i.e., $y_{\text{cochlea}}(\text{t},\text{f})$) is converted into an auditory neural pattern by a hair cell. Using this process, the cochlear output can be modeled as an intracellular pattern. This transformation can be implemented through the following steps: First, the output obtained from each filter is derived with respect to high-pass (t,f)). This action works as a high-pass filter. Then, by applying a non-linear compress function such as $gh_c(.)$ on the output obtained from the previous step, ion channels can be modeled. The compress function $gh_c(.)$ is defined as follows[25]:

$$gh_c(f) = \frac{1}{1+e^{-\gamma * f}} - 0.5 \tag{2}$$

Finally, by using a low-pass filter $\mu h_{c(t)}(0)$, the output of hair cells in the auditory system can be modeled. With this filter, frequencies higher than 4.5 kHz can be passed through the filter. The three stages described in the second step of modeling the auditory system can be described as the following relationship[25]:

$$y_{an}(t,f) = gh_c\left(\frac{\partial y_{cochlea}}{\partial t}(t,f)\right) * t\mu h_c(t) \tag{3}$$

where $y_{an}(t,f)$ represents the auditory neural pattern obtained through speech signal processing. Next, by applying the Lateral Inhibitory Network (LIN), the discontinuities of the response along the logarithmic frequency for the existing auditory neural pattern should be determined. This LIN can be simulated by the first-order differential in terms of logarithmic frequency and then by using a half-wave rectifier as follows[25]:

$$yLIN(t,\ f) = \max\left(\frac{\partial y_{an}}{\partial f}(t,\ f),\ 0\right) \tag{4}$$

The last step in the process of modeling the human auditory system is to integrate the result of the above relationship ($yLIN(t,f)$) in a short range, as follows[25]:

$$\mu_{midbrain}(t;\tau) = e^{-\frac{t}{\tau}} u(t) \tag{5}$$

where u(t) represents the unit step function and τ defines a short time constant in the range of 2 to 8 ms. With these explanations, the AS, $y(t,f)$ can be calculated as follows[25]:

$$y(t,\ f) = yLIN(t,\ f) * t\mu_{midbrain}(t;\tau) \tag{6}$$

The described process for modeling the human auditory system is shown as a diagram in Fig. 2. The output matrix resulting from the steps described above is an AS, an example of which is shown in the lower part of Fig. 2.

Temporal modulation generation based on the AS: At the higher levels of the human central auditory system, especially in the main cortex of the auditory system, the analysis is performed on the AS by estimating the signal content. In order to model the human auditory system's perception of temporal modulation, in the proposed method, the modulation dimensions analysis is used, in order to provide a more detailed view of the spectro-temporal features of speech signals. The previous works show that by using the logarithmic frequency vector along with the Q discriminator, the best mechanism can be achieved for modeling the human auditory system's perception from temporal modulation[26]. In this way, by applying continuous wavelet transformation to each of the lines in the standard spectrum, the effect of Q can be modeled in an efficient way[27]. In the proposed method, instead of using the standard spectrogram, the AS is used as the input of the modulation dimension analysis step.

The modulation dimension analysis process consists of two main steps. First, given r coefficients, a wavelet filter is applied to each time row in the AS ($y(t,f)$)[25]:

$$X^{SP}(r,\ t,\ f) = \frac{1}{r} y(t,\ f) * t\Psi\left(-\frac{t}{r}\right) \tag{7}$$

By applying (7), the output obtained from each of the cochlear channels can be filtered. In order to reduce complexity and increase computational efficiency, wavelet filters can be simulated by a filter bank, including a set of Gabor filters. Each of these filters can be adjusted for different values of spectro-temporal

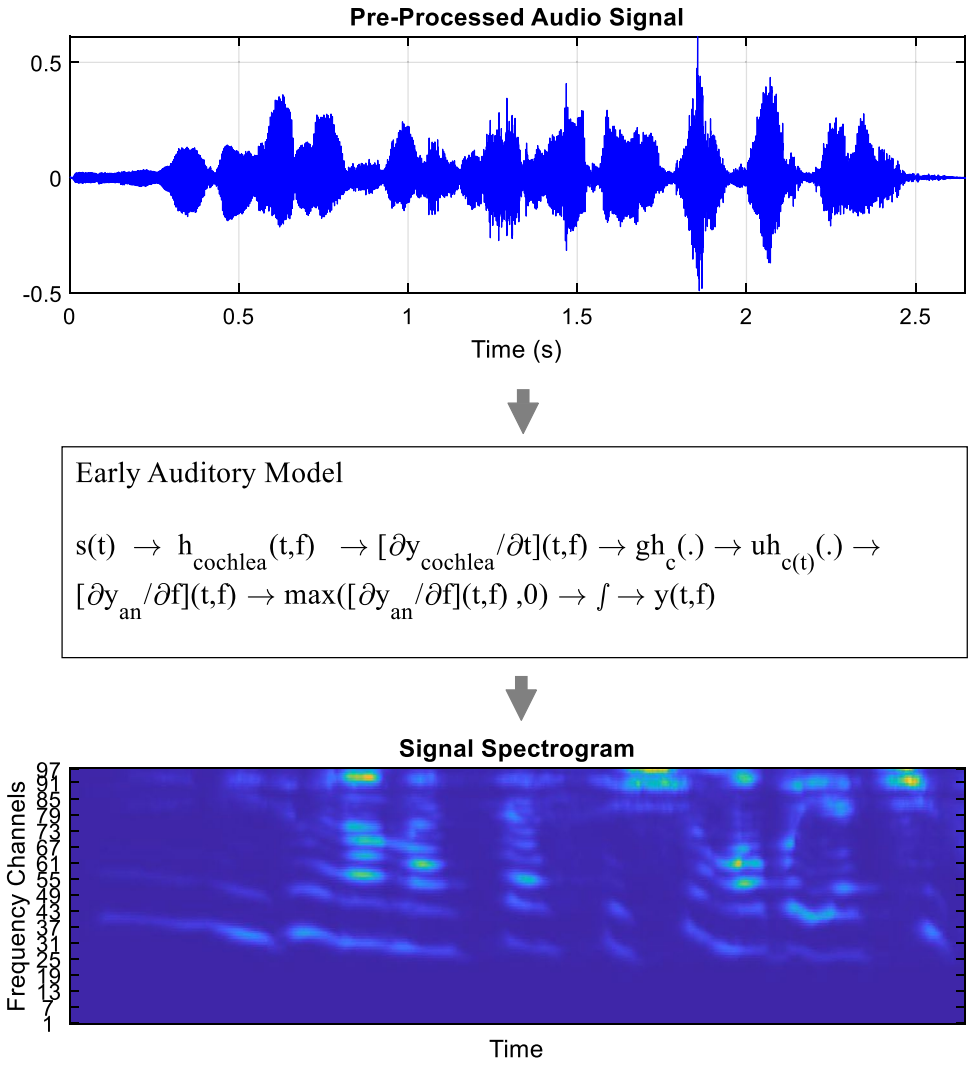

Figure 2. Auditory system modeling diagram.

parameters (low to high speeds and based on Hertz). Modulation rates determined for the set of Gabor filters are $r = \{2, 4, 8, 16, 32, 64, 128, 256\}$.

It should be noted that the output obtained in this step uses speed-time–frequency criteria to analyze the input signal. In this way, the obtained AS can be depicted in the form of a three-dimensional spectrogram in terms of speed, time, and frequency. The mentioned filters are applied on each row of this AS. This process is shown in Fig. 3.

In the last step of the temporal modulation generating process based on the AS, by integrating three-dimensional spectrogram with respect to time, the spectro-temporal modulation is achieved.

This process can be implemented as an integration of each member of the set $X^{SP}(r, t, f)$. By doing this, a two-dimensional model is obtained in terms of rate and frequency, and the obtained two-dimensional model is called auditory temporal modulation[25]:

$$X^{JF}(r, f) = \int \left| X^{SP}(r, t, f) \right|^2 dt \tag{8}$$

The process of generating temporal modulation based on the AS is provided as a diagram in Fig. 4. In short, during this process, Gabor filters are applied on each row of the AS. Then, the integral is taken from the members of the resulting set with respect to time to extract the temporal modulation.

In Fig. 4, the AS is shown in the upper part and the temporal modulation diagram extracted from this AS is shown in the lower part of the figure.

Feature extraction with CNN

In the proposed method, feature extraction is done with a CNN model. It should be noted that this step is only applied to the spectro-temporal modulation features and the entropy features extracted from the speech signal are transferred to the next step without change. The CNN structure used in the proposed method for extracting speech signal features is shown in Fig. 5. In the design of this CNN model, the simplest possible structure for

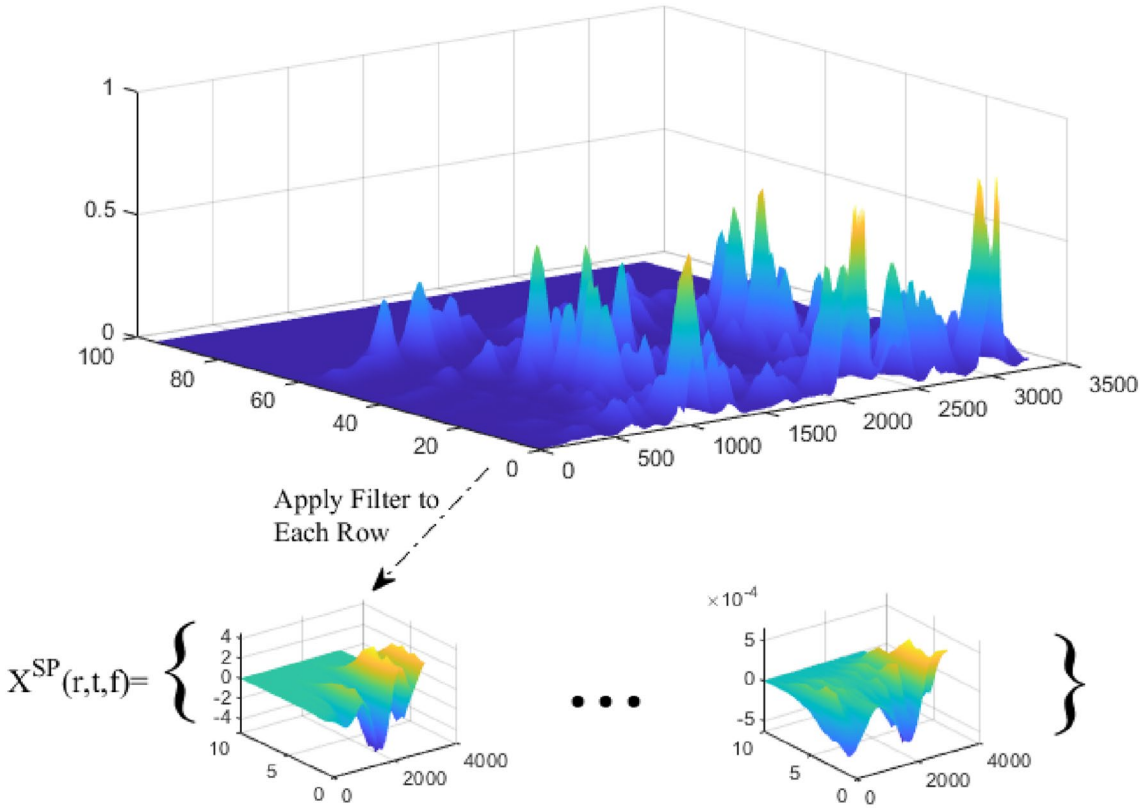

Figure 3. Applying Gabor filters to the row of the spectrogram.

feature extraction is used. After examining different architectures of CNN models for feature extraction, it was determined that the highest performance can be achieved by using a model with two convolution layers, two MaxPool layers, and two fully connected layers.

According to Fig. 5, the proposed CNN model does not have the necessary layers for feature classification, and the output of the last fully connected layer of this model is considered as the extracted features for each sample. The modulation matrices obtained from the previous step have dimensions of 97×8, which are applied to the input layer as the input of this CNN model. Two convolution layers consisting of 32 filters with dimensions of 5 and 3 respectively are responsible for extracting input patterns. Finally, the extracted patterns are converted into vectors by using two consecutive fully connected layers and the dimensions of the feature matrix are reduced to 50. The feature vector extracted through this CNN model is combined with entropy features for each sample so that they are finally used as the input of the proposed classification model.

Classification based on GC and ECOC

In the last step of the proposed method, the combination of ECOC and GC is used to recognize emotions in speech. The ECOC model is presented as a comprehensive method for solving complex problems, and it is a complete structure for solving multi-class problems through the combination of several binary problems, which can be used to solve problems such as face recognition and emotion recognition, etc. The framework of ECOCs consists of two steps: encryption and decryption. In the coding stage, a code word is assigned to each class in the problem (emotional state). Each code word contains a string of bits, each bit of which indicates the belonging or non-belonging of the class corresponding to that bit for the given binary classification. By combining these codes, a matrix called the code matrix is created, whose rows correspond to the codes created for each of the target classes. In the ECOC model, a binary classifier is assigned to each column of the code matrix, which is trained based on the binary codes corresponding to its column. There are different ways to create the code matrix. In the proposed method, the dense random algorithm[28] is used to form the code matrix, and $10 \times log_2 C$ numbers of binary gamma classification are trained based on this strategy.

After training the GCs based on the columns of the code matrix, the Hamming distance criterion is used to recognize emotional states in new samples. For this purpose, first, the features of each training sample are processed by all GCs so that each classifier creates a binary code as an output. By combining these outputs, a binary string is created, then by matching it with the row of the code matrix, the emotional state in the input sample can be determined. For this purpose, the Hamming distance between the mentioned binary string and each line of the code matrix is calculated, and thus the sample belongs to the class that has the smallest distance. In the following, the mechanism of each GC in the proposed model is explained.

GC is a supervised method whose name is derived from the similarity operator which it uses: the generalized gamma operator. This operator receives two binary vectors x and y and a positive integer like θ as input and

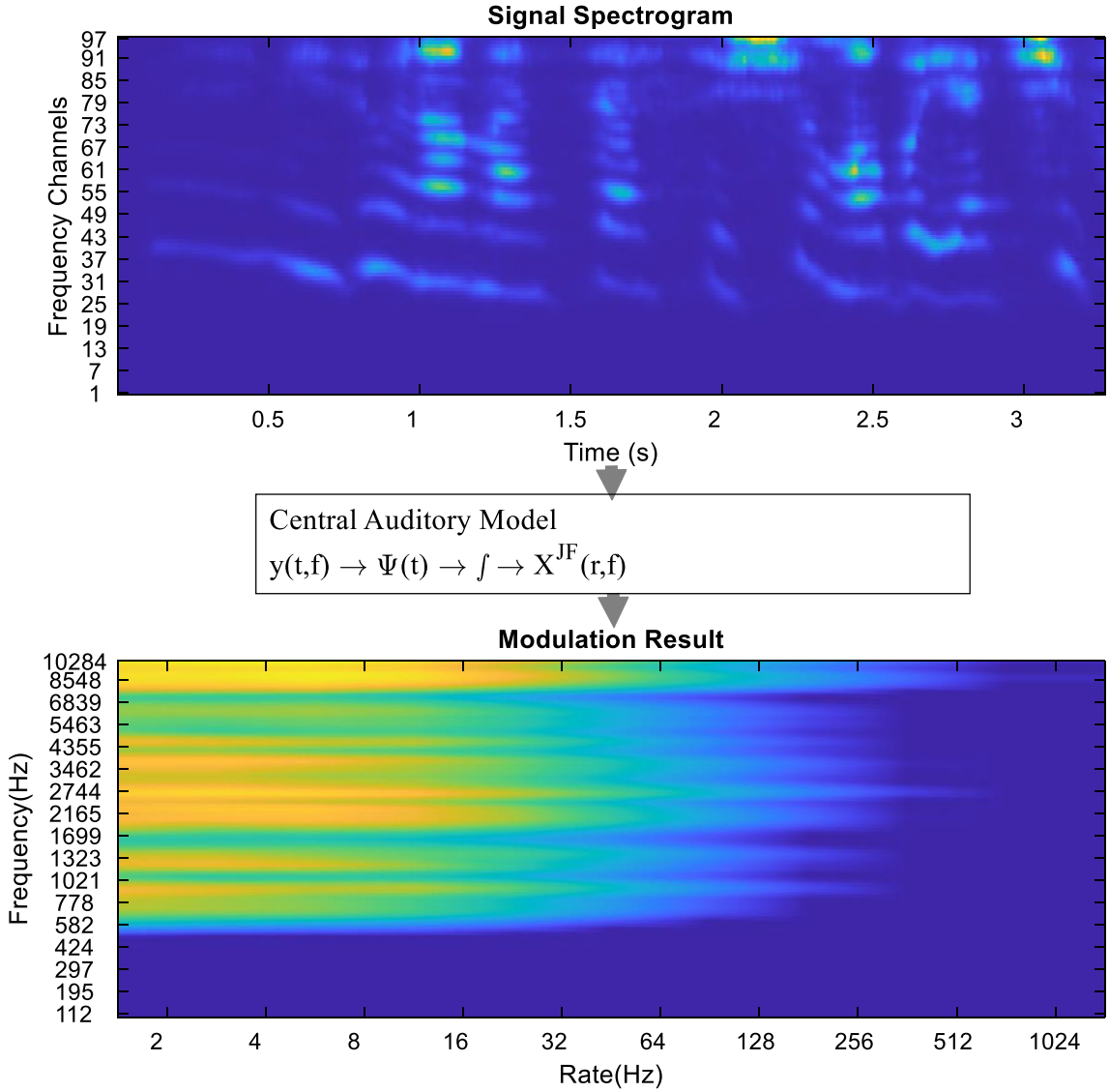

Figure 4. Diagram of the temporal modulation generation steps based on the AS.

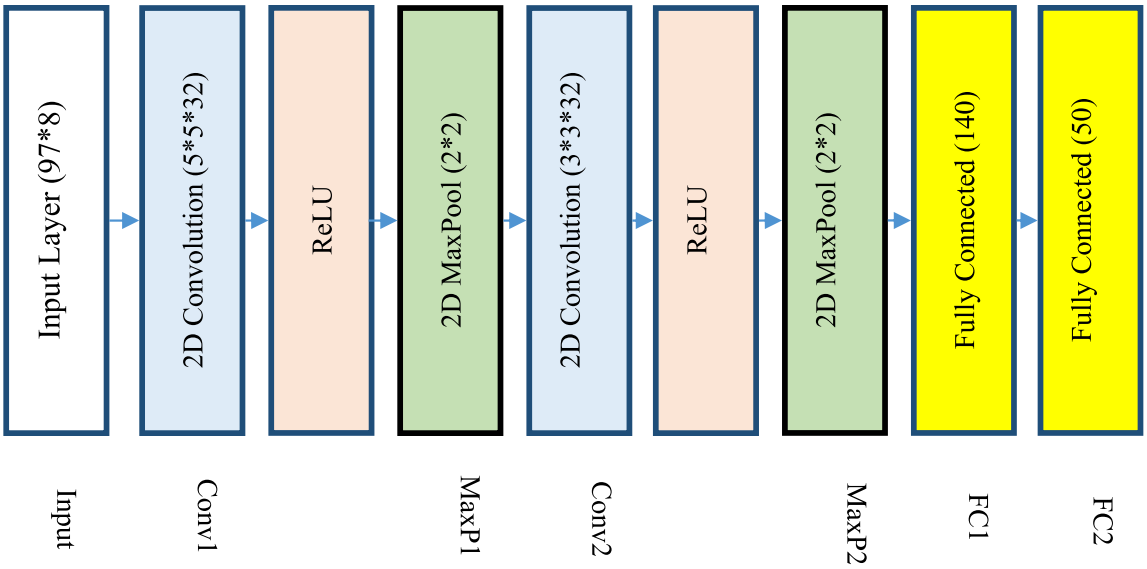

Figure 5. Proposed CNN structure for feature extraction.

returns 1 if both vectors are the same and 0 otherwise. The gamma operator uses other operators (such as α, β, etc.), which are defined below.

Definition 1 α and β operators: Given the set $A = \{0,1\}$ and $B = \{0,1,3\}$, the alpha (α) and beta (β) operators are as shown in Table 1.

Definition 2 α operator: Let $x, y \in A^n$ be the input column vectors. The output $\alpha(x, y)$ is the n-dimensional vector whose components are calculated as follows:

$$\alpha(x,\ y)_i = \alpha(x_i,\ y_i) \tag{9}$$

Definition 3 u_β operator: considering the binary pattern $x \in A^n$ as input, this operator generates the following non-negative integer as output and is calculated as follows:

$$u_\beta(x) = \sum_{i=1}^{n} \beta(x_i,\ x_i) \tag{10}$$

Definition 4 Pruning operator: Let $x \in A^n$ and $y \in A^m$ and $n < m$ be two binary vectors. Then y pruned by x is displayed as $y \| x$ and is a binary and n-dimensional vector whose components are calculated as follows:

$$(y||_x)_i = y_{i+m-n}, (i = 1,\ 2,\ \ldots,\ n) \tag{11}$$

The gamma operator requires a binary vector as input. To deal with real vectors or integers, a method to represent these vectors in binary form is needed. In this work, the modified Johnson-Mobius code in[29] is used. Because the full details of this algorithm have been discussed in [29], its process is not explained in this paper.

Definition 5 Gamma operator: Gamma similarity operator takes two binary patterns such as $x \in A^n$ and $y \in A^m$ and $n \leq m$ and a non-negative integer such as θ as input and generates a binary output for each of the following two states.

Case 1 If $n = m$, the output is calculated according to the following equation[30]:

$$\gamma(x,\ y,\ \theta) = \begin{cases} 1,\ if\ m - u_\beta[\alpha(x,\ y) mod\ 2] \leq \theta \\ 0,\ otherwise \end{cases} \tag{12}$$

In (12), "mod" indicates the remainder of the division.

Case 2 If $n < m$, the output is calculated using $y||_x$ instead of y according to the following equation[30]:

$$\gamma(x,\ y,\ \theta) = \begin{cases} 1,\ if m - u_\beta[\alpha(x, y||_x) mod 2] \leq \theta \\ 0,\ otherwise \end{cases} \tag{13}$$

Considering the above operators, the GC operator is explained in the following. Assuming that the basic patterns are $\{(x^u, y^u) | u = 1, 2, \ldots, p\}$ with cardinality p and the test pattern is described as $\tilde{x} \in R^n$. In this case, the GC classifies the experimental sample $\tilde{x}$ through the following calculation steps:

Step 1 Convert the training base set to binary form using the modified Johnson-Mobius code so that e_m value is calculated for each component in the n-dimensional vectors of the base set as follows[30]:

$$e_m(j) = \max_{i=1}^{p} \left(x_j^i\right),\quad \forall j \in \{1,\ 2,\ \ldots,\ n\} \tag{14}$$

$\alpha : A \times A \rightarrow B$			$\beta : B \times A \rightarrow A$		
x	y	$\alpha(x,y)$	x	y	$\beta(x,y)$
0	0	1	0	0	0
0	1	0	0	1	0
1	0	2	1	0	0
1	1	1	1	1	1
			2	0	1
			2	1	1

Table 1. α and β operators in GC.

Step 2 The stopping parameter is as $\rho = \max_{j=1}^{n}[e_m(j)]$.

Step 3 The test pattern is also modified using the Johnson-Mobius code and coded with the same parameters used to code the original set. If each obtained e_j is greater than its corresponding $e_m(j)$, it is coded with more bits.

Step 4 The indicators of all base patterns are converted into two indicators: one for their class and another for their position in the class.

Step 5 The initial value of the parameter θ is set to zero.

Step 6 If $\theta = 0$‹, then by calculating $\gamma(x_j^u, \tilde{x}_j, 0)$ it is checked whether $\tilde{x}$ is a fundamental pattern and then the initial weighted addition c_u^0 for each basic pattern is calculated as follows[30]:

$$c_u^0 = \sum_{i=1}^{n} \gamma\left(x_j^u,\ \tilde{x}_j,\ 0\right), \quad \text{for u} = 1,\ 2,\ \ldots,\ \text{p} \tag{15}$$

If there is a unique maximum value equal to n; Then the class corresponding to this maximum value is assigned to the test pattern[30]:

$$\tilde{y} = y^{\sigma} \quad such\ that \quad \max_{i=1}^{p} c_i^0 = c_\sigma^0 = n \tag{16}$$

Step 7 The value of $\gamma\left(x_j^{i\omega}, \tilde{x}_j, \theta\right)$ is calculated for each component of the basic patterns.

Step 8 The value of the weighted sum c_i for each class is calculated[30] with (17).

$$c_i = \frac{\sum_{\omega=1}^{k_i} \sum_{j=1}^{n} \gamma\left(x_j^{i\omega},\ \tilde{x}_j,\ \theta\right)}{k_i} \tag{17}$$

In the above relationship, k_i represents the cardinality of the base set of class i.

Step 9 If there is more than one maximum value among different c_i, the value of θ increases by one unit then steps 7 and 8 are repeated until there is only one maximum value, or the termination condition $\theta > \rho$ is met.

Step 10 If there is a unique maximum value among different c_i; Then the class corresponding to the experimental pattern $\tilde{x}$ is calculated as follows[30]:

$$\tilde{y} = y^{j} \quad such\ that \quad \max c_i = c_j \tag{18}$$

Step 11 Otherwise, the pattern $\tilde{x}$ is assigned to the class with the first maximum value.

Results

The implementation of the proposed method was done to evaluate its performance using MATLAB 2016a software. During these tests, the effectiveness of the proposed method was investigated in terms of accuracy and classification quality, and the results were compared with previous similar works. Also, the speech signals available in ShEMO[22] and Berlin Emotional[23] datasets were used as input for experiments. The specification of these datasets were described in Section "Introduction"–"Research method". The evaluation process was done separately for each of these datasets using tenfold cross-validation. Figure 6 shows the results of the accuracy of the proposed method in comparison with other methods for extracting sentiments from the Berlin and ShEMO

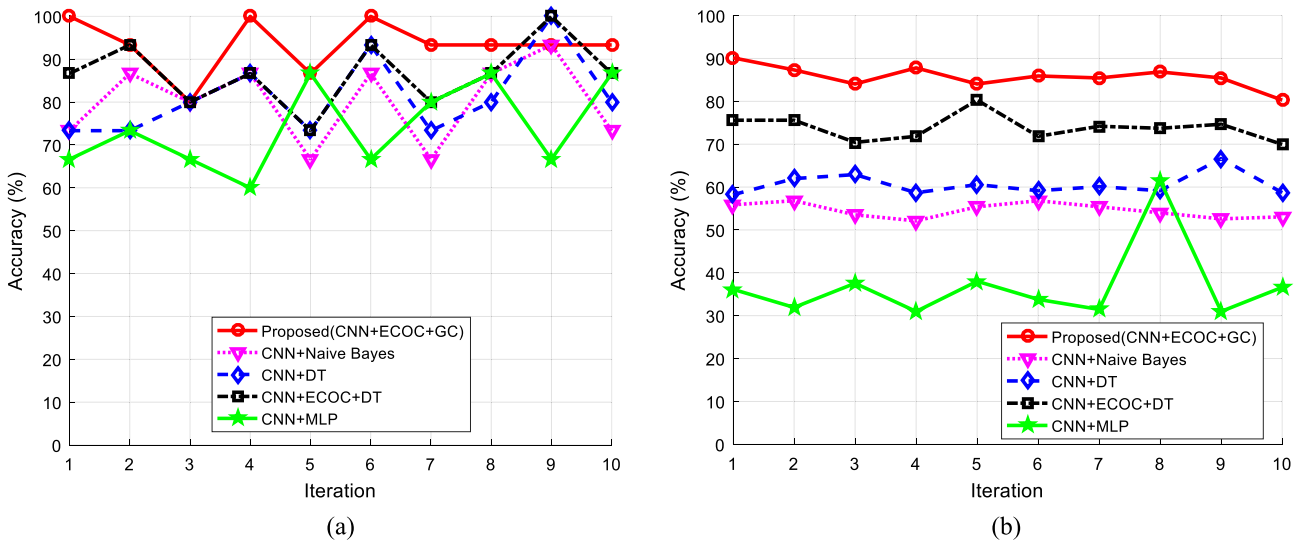

Figure 6. The accuracy of different methods in extracting emotions from two datasets (**a**) Berlin and (**b**) ShEMO during 10 iterations of cross-validation.

datasets during 10 iterations of cross-validation. In these graphs, the performance of the proposed method is compared with the case where other learning models use the extracted features of the CNN model.

Based on Fig. 6, the proposed method can perform the process of recognizing emotions in speech more accurately in most iterations, and it can be said that it is generally more accurate than the compared methods. Since the only difference between the compared methods is the type of classifier applied to them, and all these methods' input features are the same; Therefore, the improvement of accuracy obtained in the proposed method can be attributed to the appropriate performance of its classification model. In the proposed method, a combination of ECOC and GC is used in order to recognize emotional states. This combination can correctly recognize emotional states in the Berlin and ShEMO datasets with an average accuracy of 93.33 and 85.73%, respectively. The higher accuracy of the proposed method in the Berlin dataset comes from two factors. First, the data set used from the Berlin dataset includes 5 emotional states. While the number of emotional states in ShEMO is equal to 6 and this increases the complexity of the problem. Secondly, the number of speakers in the ShEMO dataset is much more than the Berlin dataset, which can add to the complexity of the problem. Because each speaker may convey an emotional state through a specific pattern in speech. In Fig. 7, the average accuracy values of the proposed method and other methods are compared for the Berlin and ShEMO datasets.

The Fig. 7 confirms that the proposed method for both Berlin and ShEMO datasets leads to increased recognition accuracy. As mentioned, this improvement can be attributed to the performance of the proposed classification. On the other hand, the combination of ECOC and decision tree has the closest results to the proposed method, which is a direct result of using the ECOC model to overcome the high complexity of multiclass problems.

Inspecting Figs. 6 and 7 shows that the proposed method has two advantages in the recognition process. First, the proposed method can recognize emotions more accurately than other methods. Second, the accuracy variation range of the proposed method is more limited than the compared methods. High accuracy and at the same time, the limitation of its variation range can be considered as a strong point of a recognition system. Because this feature shows the reliability of the outputs generated by that recognition system. These conditions are shown in Fig. 8 for two datasets, Berlin and ShEMO.

In Fig. 8, each box represents the variation range of the accuracy during 10 times of cross-validation. The middle circle indicates the median value of the accuracy during the iterations, and each part of the box shows one of the quartiles of the accuracy changes. Outliers are also drawn as points outside the box. Based on these plots, the proposed method has more compact boxes that are at higher levels than other methods, and this confirms the effectiveness of the proposed method in more accurately recognizing emotions from speech.

The Fig. 9 shows the confusion matrix of the proposed method for recognizing emotional states in the Berlin and ShEMO datasets. The numbering of the classes in these confusion matrices is based on the order of the classes explained at the beginning of this section. For example, the number 2 in the Berlin dataset indicates the state of hatred; While this tag in the ShEMO dataset describes the emotional state of fear. The rows of each confusion matrix represent the outputs of the learning model (recognized emotional states); While the real emotional states of the samples are displayed in the columns of the matrix. In this way, the elements of the main diameter of each matrix represent the number of samples of each emotional state that have been correctly recognized, and the other elements represent the number of errors in classification. Also, the state confusion matrix of emotion recognition by the combination of ECOC and decision tree (the method with the closest performance to the proposed method) is presented in Fig. 10.

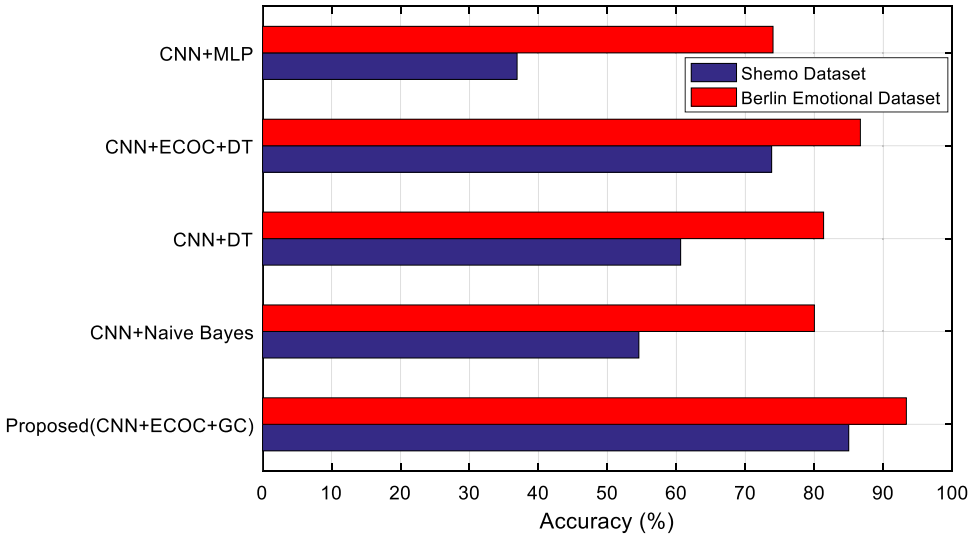

Figure 7. Comparison of the accuracy of different methods for extracting sentiments from the Berlin and ShEMO datasets.

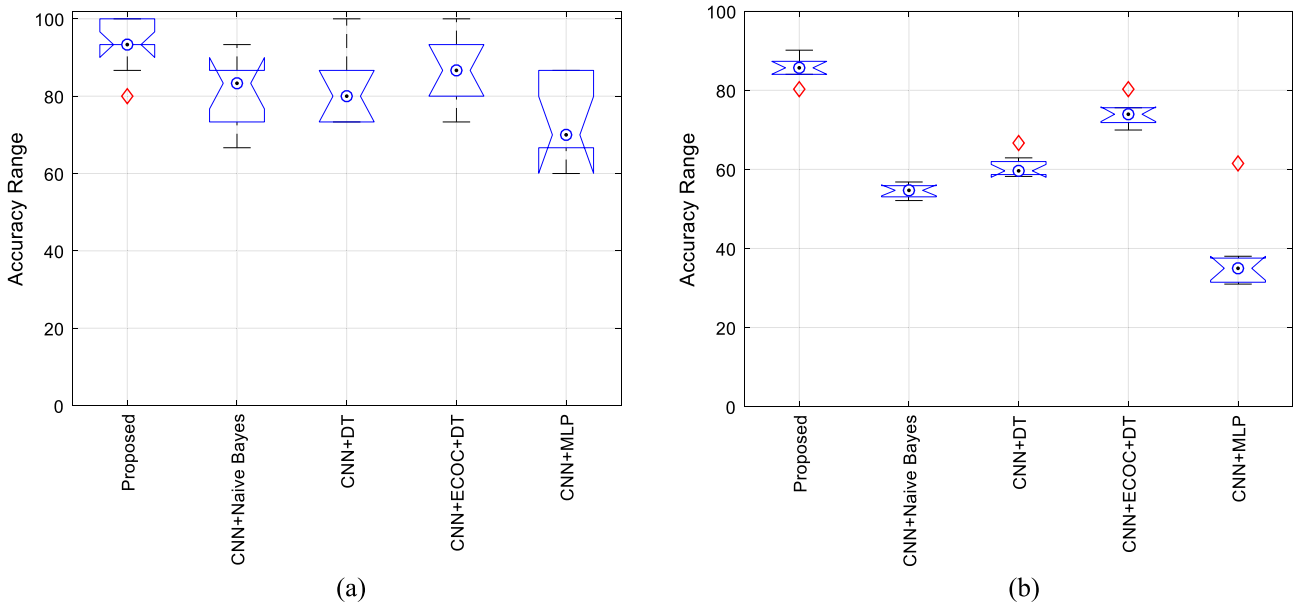

Figure 8. Box plot of different methods accuracy for two datasets (**a**) Berlin and (**b**) ShEMO during 10 iterations of cross-validation.

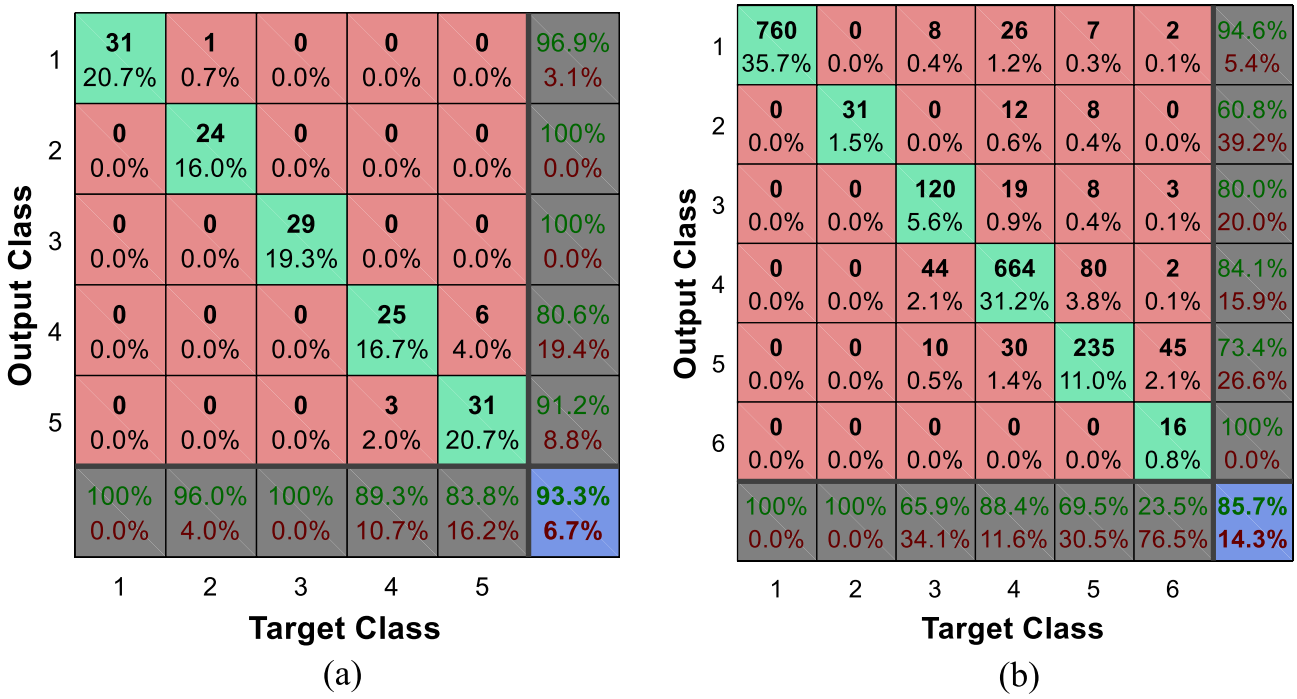

Figure 9. Confusion matrix of the proposed method for recognizing types of emotional states in two datasets (**a**) Berlin and (**b**) ShEMO, after 10 folds of cross-validation.

Comparison of Figs. 9 and 10 shows that the proposed method can recognize any emotional state with higher accuracy than other methods. By inspecting each emotional state, it can be seen that in the Berlin dataset, the proposed method can recognize samples of each class with higher accuracy. On the other hand, in the ShEMO dataset, the proposed method does not perform well in recognizing samples related to the emotional state of surprise, but it has an acceptable performance in recognizing other emotional states. The relatively small number of surprise class samples can be the main reason for the ineffectiveness of the proposed method in recognizing samples of this class. In addition, the high similarity of the features of this emotional state to the state of sadness in the ShEMO dataset can also be one of the reasons for this; Because about 90% of this class samples are wrongly classified as an emotional state of sadness.

Figure 10. Confusion matrix of the combination of ECOC and decision tree for emotion recognition in the dataset (**a**) Berlin, (**b**) ShEMO, after 10 folds of cross-validation.

In Fig. 11, the Received Operating Characteristics (ROC) curve of different methods for recognizing emotional states in the two test bases is presented. In this curve, the values of the true positive rate (TPR) are displayed for changes in the false positive rate (FPR). The goal of any emotion recognition method is to achieve higher TPR values and at the same time, reduce FPR. Therefore, an emotion recognition system with a higher level of the ROC curve has a better performance. Considering that the number of target classes (emotional states) in the research problem is more than two; Therefore, to draw this curve, the values of TPR and FPR were calculated for

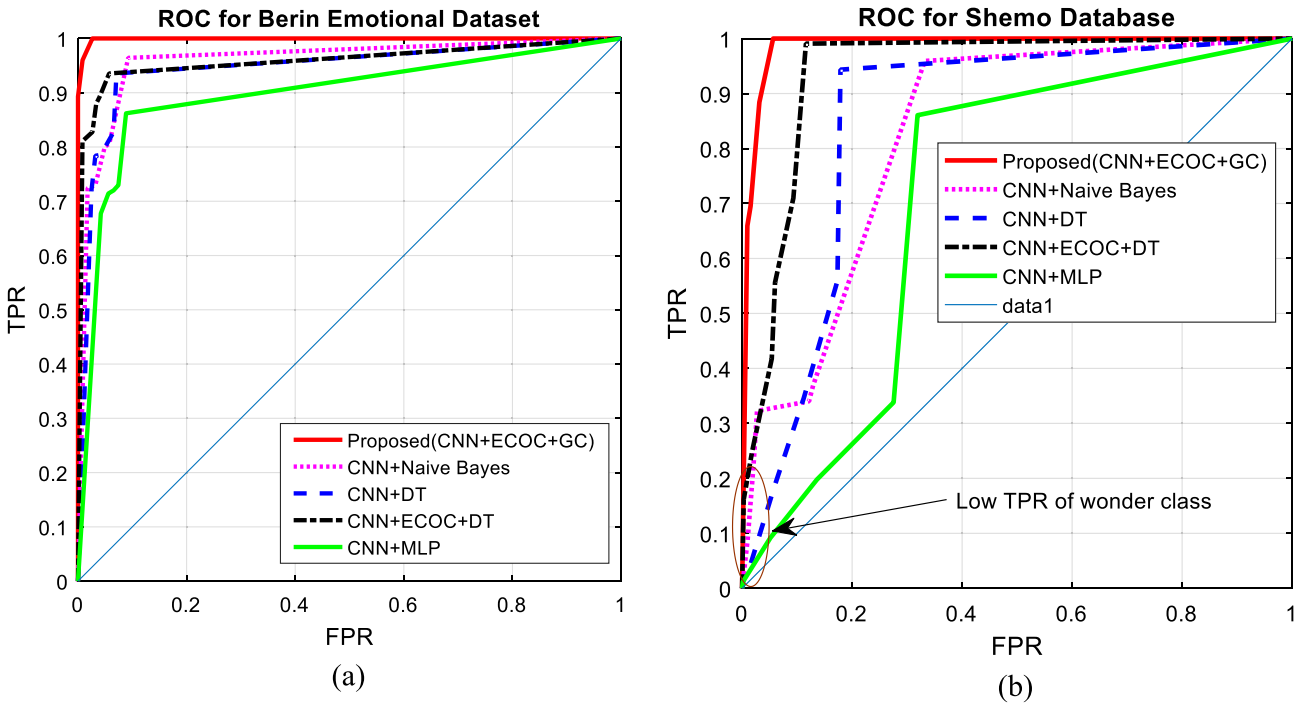

Figure 11. ROC curve for speech emotion recognition in (**a**) Berlin and (**b**) ShEMO datasets after 10 folds of cross-validation.

different classes, and each time, the current emotional state was considered as a positive class and other classes as a negative class.

The graphs depicted in Fig. 11, show that the proposed method can achieve higher TPR and lower FPR values for both tested datasets. These results confirm that the proposed method is superior in recognizing emotional states separately for each class. The ROC curve in Fig. 11b for the ShEMO dataset shows that the proposed method has a lower TPR at the initial points of the curve than the combination of ECOC and decision tree. This low TPR results from the inappropriate performance of the proposed method in recognizing the emotional state of surprise, which was discussed in Fig. 9. However, the proposed method has worked well in accurately recognizing other emotional states, which results in achieving higher levels of the ROC curve.

To better check the effectiveness of the proposed method, the precision, recall, and F-Measure criteria can be used. The precision criterion, can demonstrate the accuracy of system in recognizing samples of each target class, separately. Also, the recall criterion shows that which ratio of samples belonging to each target class has been recognized correctly. Finally, the F-Measure is used to describe the classification efficiency of the system by harmonic mean of precision and recall. These criteria are formulated by Eqs. (19) to (21):

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{21}$$

Since these criteria consider the classification task as a binary problem, they are calculated for each target class, separately. Therefore, for calculating these criteria for each target class, the mentioned class is considered as positive while the other classes are assumed to be negative. In the above equations, TP refers to the number of correctly recognized samples belonging to positive class. Also, FP demonstrates the number of samples which actually belong to negative class but are classified as positive. Finally, FN refers to the number of positive samples which have been incorrectly labeled as positive.

In Fig. 12, the efficiency of the proposed method is compared with other methods in terms of precision, recall, and F-Measure criteria. Also, the numerical results related to these criteria are given in Table 2.

Comparing precision, recall, and F-Measure values in Fig. 12 and Table 2 shows that the proposed method can be more successful in separating the emotional states of each class. These results confirm the claim made in this article regarding the greater effectiveness of the combination of ECOC and GC for more accurate recognition of emotions in speech. On the other hand, the comparison of the criteria values in Table 2 also confirms that this model can classify emotional states more accurately than previous methods.

The research conducted in[14] used the ShEMO dataset to recognize emotional states, but this method was limited to recognizing only two emotional states: anger and neutral. Therefore, in order to compare the effectiveness of the proposed method with the results of this research, other target classes are ignored and emotion recognition is done only based on samples belonging to the two classes "anger" and "neutral".

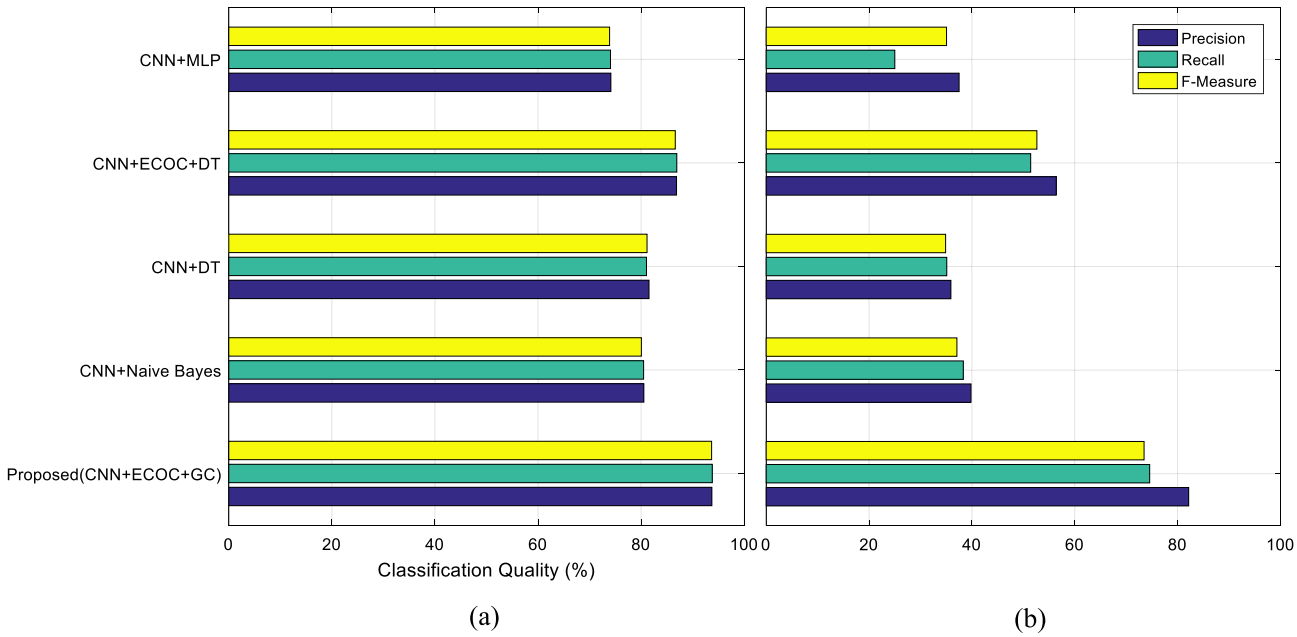

Figure 12. Classification rates of speech emotion recognition in (**a**) Berlin and (**b**) ShEMO datasets.

Dataset	Method	Accuracy	F-measure	Recall	Precision
Berlin	Proposed	93.3333	93.6883	93.8139	93.7393
	NB[31]	80.0000	80.0573	80.4921	80.5399
	DT[32]	81.3333	81.1562	81.0514	81.5426
	ECOC + DT[33]	86.6667	86.6354	86.9348	86.8789
	MLP[34]	74.0000	73.8902	74.0701	74.1400
	Kerkeni et al.[10]	82.4100	82.1761	83.0122	81.3567
ShEMO	Proposed	85.7277	73.4749	74.5676	82.1529
	NB[31]	54.5540	37.0834	38.3228	39.8235
	DT[32]	60.6103	34.8818	35.1087	35.9131
	ECOC + DT[33]	73.8028	52.6375	51.4266	56.4003
	MLP[34]	36.9014	35.0632	25.0046	37.5138
	Liou et al.[16]	70.5333	69.6606	71.0404	68.3333

Table 2. Comparing the efficiency of the proposed method with other classification models.

The method in[14] can perform two emotional states "anger" and "neutral" without separating the gender samples with 90.97% accuracy. Meanwhile, by using the proposed method, these two emotional states can be separated with an accuracy of 93.23%.

These results show that by using the proposed method, the recognition accuracy can be increased by 2.26% compared to the method of[14]. Figure 13 shows the confusion matrix resulting from the recognition of the two mentioned emotional states by the proposed method. Also, in Fig. 14, the accuracy of the proposed method is compared with other speech feature extraction methods.

These results clearly show that by using the combination of spectro-temporal modulation and entropy features in the proposed method, the accuracy of recognizing emotional states can be increased compared to previous methods.

Discussion

Using the experiments conducted in this research, it was attempted to identify the advantages and limitations of the proposed strategy from different aspects. In this regard, two different datasets that include various emotions were used to conduct experiments in order to prove the generality of the proposed strategy. The evaluation and comparison of the proposed method in terms of identification accuracy (Figs. 6 and 7) and its variation range (Fig. 8) showed that the combination of ECOC and GC can perform better than the compared classifications (such as the combination of ECOC with NB or DT), and achieved higher accuracy values in both tested data sets. In addition, the range of accuracy changes reported by the proposed method during CV folds has been narrower than other methods, which confirms the higher reliability of the output of this model. Nevertheless, the examination of confusion matrices (Figs. 9 and 10) for two datasets, ShEMO and Berlin, showed that the proposed method is more successful in recognizing each emotion than other classifiers. Examining the ROC curve (Fig. 11) and the classification rates (Fig. 12) proved this claim. The higher precision values show that this model is more accurate in assigning labels of each target category to samples, and the superiority of the recall criterion confirms the higher efficiency of the proposed method in recognizing samples of each emotional state.

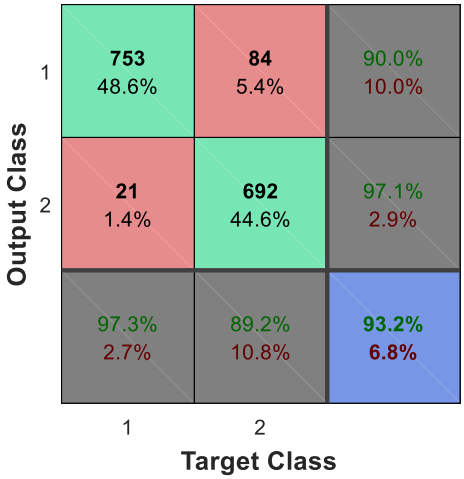

Figure 13. Confusion matrix of the proposed method for recognizing two emotional states of anger and neutral.

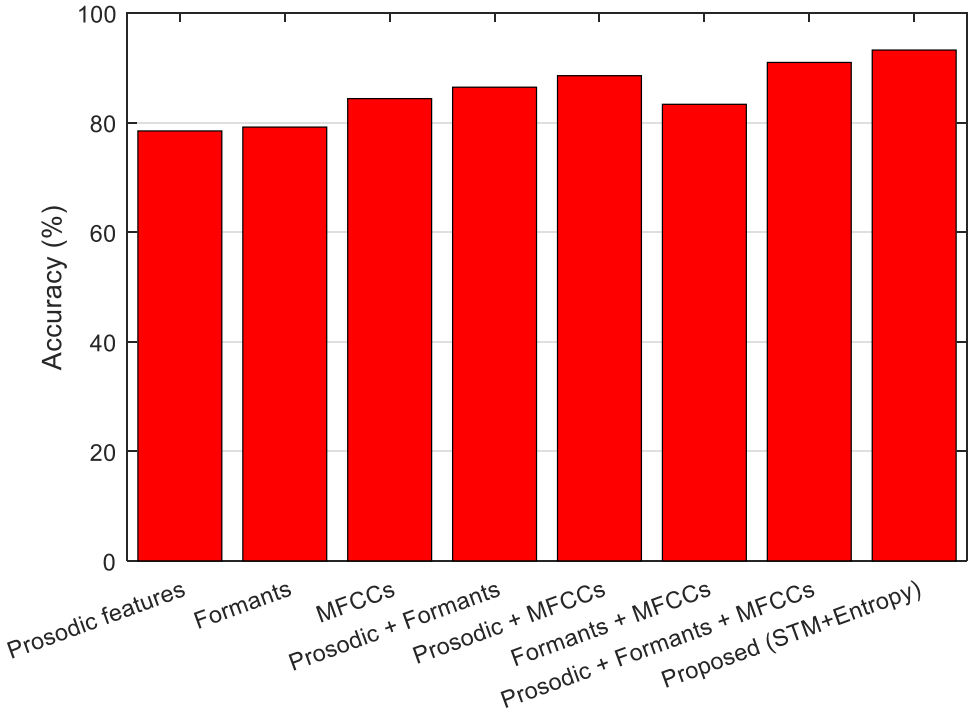

Figure 14. Accuracy comparison of the proposed method with other feature extraction methods for recognizing the two emotional states of anger and neutral.

Overall, these results showed that the proposed method performs better than methods such as Kerkeni et al.[10] and Liou et al.[16] for both Berlin and ShEMO datasets. This superiority can be attributed to two main factors: First, the combination of ECOC and GC has led to the formation of a more powerful classification model for identifying emotional states, which can be used to resist the increased complexity caused by the increase in the number of emotional states. Secondly, the use of STM has made it possible to represent emotional states in speech signals in a more efficient way, and the extraction of STM features by CNN and the combination of entropy features with it can be effective in increasing the accuracy of the model by at least 2.26%. The results presented in Figs. 13 and 14 prove this claim. Based on these results, the combination of STM and entropy features leads to higher accuracy compared to conventional feature description models such as MFCC, Formants or Prosodic[14].

One of the limitations of the proposed method is the need for higher computing power of its learning model, which results from the use of several GCs in the ECOC model. This has caused the processing time of the proposed method to train the learning model to be more than the conventional methods. Although this increase in processing time only occurs in the training phase; this problem can be solved by using parallel processing techniques. In future works, the use of other feature extraction techniques to describe the emotional features of speech can be investigated. Also, combining the ECOC model with other existing classifiers to achieve a more accurate emotion recognition system can be a topic for further research.

Conclusion

In this paper, a new method is presented to recognize emotion in speech using machine learning techniques. The proposed method uses a set of entropy features and spectro-temporal modulation to describe speech features, and the feature extraction is done by a CNN. Also, a new model based on the combination of GC and ECOC is utilized to classify features and recognize emotional states. These two techniques make it possible to recognize the emotional states of speech with higher accuracy and efficiency in comparison with previous works. The performance of the proposed method was tested through two datasets, Berlin and ShEMO, and the results were compared with previous similar works. The obtained results showed that the proposed method is an accurate solution in recognizing the emotional states of speech and could recognize speech emotions in the Berlin and ShEMO datasets with an average accuracy of 93.33 and 85.73%, respectively, which had an improvement of at least 2.1% compared to the compared methods. On the other hand, comparing the performance of feature extraction techniques in recognizing emotional states showed that by combining spectro-temporal modulation and entropy in the proposed method, the accuracy of recognizing emotional states can be 2.26% higher than compared methods.

Data availability

All data generated or analyzed during this study are included in this published article.

Received: 19 July 2023; Accepted: 9 November 2023
Published online: 21 November 2023

References

1. Kadiri, S. R. & Alku, P. Excitation features of speech for speaker-specific emotion detection. *IEEE Access* **8**, 60382–60391 (2020).
2. Ramesh, S., Gomathi, S., Sasikala, S. & Saravanan, T. R. Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes. *Int. J. Speech Technol.* **2**, 1–8 (2021).
3. Lalitha, S., Tripathi, S. & Gupta, D. Enhanced speech emotion detection using deep neural networks. *Int. J. Speech Technol.* **22**, 497–510 (2019).
4. Atmaja, B. T., Sasou, A. & Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* **140**, 11–28 (2022).
5. Saxena, A., Khanna, A. & Gupta, D. Emotion recognition and detection methods: A comprehensive survey. *J. Artif. Intell. Syst.* **2**(1), 53–79 (2020).
6. Akçay, M. B. & Oğuz, K. Speech emotion recognition: Emotional models, datasets, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **116**, 56–76 (2020).
7. Abbaschian, B. J., Sierra-Sosa, D. & Elmaghraby, A. Deep learning techniques for speech emotion recognition, from datasets to models. *Sensors* **21**(4), 1249 (2021).
8. Ke, X., Zhu, Y., Wen, L. & Zhang, W. Speech emotion recognition based on SVM and ANN. *Int. J. Mach. Learn. Comput.* **8**(3), 198–202 (2018).
9. Alghifari, M. F., Gunawan, T. S. & Kartiwi, M. Speech emotion recognition using deep feedforward neural network. *Indones. J. Electr. Eng. Comput. Sci.* **10**(2), 554–561 (2018).
10. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. & Mahjoub, M. A. Speech emotion recognition: Methods and cases study. *ICAART* **2**, 20 (2018).
11. - Kumbhar, H. S., & Bhandari, S. U. (2019). Speech emotion recognition using MFCC features and LSTM network. In *2019 5th International Conf. On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1–3). IEEE.
12. - Xu, M., Zhang, F., & Khan, S. U. (2020). Improve accuracy of speech emotion recognition with attention head fusion. In *2020 10th Annual Computing and Communication Workshop and Conf. (CCWC)* (pp. 1058–1064). IEEE.
13. Fahad, M. S., Deepak, A., Pradhan, G. & Yadav, J. DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. *Circ. Syst. Signal Process.* **40**, 466–489 (2021).
14. Horkous, H. & Guerti, M. Recognition of anger and neutral emotions in speech with different languages. *Int. J. Comput. Dig. Syst.* **10**, 563–574 (2021).
15. - Samarasekara, I., Udayangani, C., Jayaweera, G., Jayawardhana, D., & Abeygunawardhana, P. K. (2020). Non invasive continuous detection of mental stress via readily available mobile-based help parameters. In *2020 IEEE Region 10 Conf. (TENCON)* (pp. 579–584). IEEE.
16. Liu, Z. T., Rehman, A., Wu, M., Cao, W. H. & Hao, M. Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. *Inf. Sci.* **563**, 309–325 (2021).
17. Huang, C., Han, Z., Li, M., Wang, X. & Zhao, W. Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis. *Aust. J. Educ. Technol.* **37**(2), 81–95. https://doi.org/10.14742/ajet.6749 (2021).
18. Zhang, X. *et al.* Self-training maximum classifier discrepancy for EEG emotion recognition. *CAAI Trans. Intell. Technol.* https://doi.org/10.1049/cit2.12174 (2023).
19. Liu, X. *et al.* Emotion classification for short texts: an improved multi-label method. *Hum. Social Sci. Commun.* **10**(1), 1–9 (2023).
20. Liu, Z. *et al.* Emotion-semantic-aware dual contrastive learning for epistemic emotion identification of learner-generated reviews in MOOCs. *IEEE Trans. Neural Netw. Learn. Syst.* https://doi.org/10.1109/TNNLS.2023.3294636 (2023).
21. Nie, W., Bao, Y., Zhao, Y. & Liu, A. Long dialogue emotion detection based on commonsense knowledge graph guidance. *IEEE Trans. Multim.* https://doi.org/10.1109/TMM.2023.3267295 (2023).
22. Mohamad Nezami, O., Jamshid Lou, P. & Karami, M. ShEMO: A large-scale validated dataset for Persian speech emotion detection. *Lang. Resour. Eval.* **53**, 1–16 (2019).
23. - EMO-DB: Berlin Emotional Dataset (Access time: 2022), *Institute of Communication Science, Technical University, Berlin*, Available online at: https://www.kaggle.com/datasets/piyushagni5/berlin-dataset-of-emotional-speech-emodb.
24. Delgado-Bonal, A. & Marshak, A. Approximate entropy and sample entropy: A comprehensive tutorial. *Entropy* **21**(6), 541 (2019).
25. Panagakis, Y., Kotropoulos, C. & Arce, G. R. Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. *IEEE Trans. Audio Speech Lang. Process.* **18**(3), 576–588 (2009).
26. Edraki, A., Chan, W. Y., Jensen, J. & Fogerty, D. Speech intelligibility prediction using spectro-temporal modulation analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 210–225 (2020).
27. Edraki, A., Chan, W. Y., Jensen, J. & Fogerty, D. Spectro-temporal modulation glimpsing for speech intelligibility prediction. *Hear. Res.* **52**, 108620 (2022).
28. Zhou, J., Yang, Y., Zhang, M. & Xing, H. Constructing ECOC based on confusion matrix for multiclass learning problems. *Sci. China Inf. Sci.* **59**(1), 1–14 (2016).
29. - Yáñez, C., Felipe-Riveron, E., López-Yáñez, I., & Flores-Carapia, R. A novel approach to automatic color matching. In *Iberoamerican Congress on Pattern Recognition* (pp. 529–538). (Springer, Berlin, 2008).
30. Uriarte-Arcia, A. V., López-Yáñez, I., Yáñez-Márquez, C., Gama, J. & Camacho-Nieto, O. Data stream classification based on the gamma classifier. *Math. Prob. Eng.* **2015**, 939175. https://doi.org/10.1155/2015/939175 (2015).
31. - Khan, A., & Roy, U. K. (2017, March). Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier. In *2017 International Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 1017–1021). IEEE.
32. Liu, Z. T. *et al.* Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neuro Comput.* **273**, 271–280 (2018).
33. Hamsa, S., Shahin, I., Iraqi, Y. & Werghi, N. Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier. *IEEE Access* **8**, 96994–97006 (2020).
34. Alnuaim, A. A. *et al.* Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. *J. Healthc. Eng.* https://doi.org/10.1155/2022/6005446 (2022).

Author contributions

All authors wrote the main manuscript text. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.