Journal of Computing and Security

Research Article, 2024, Volume 11, Number 1 (pp. 39–48)

http://www.jcomsec.org

# Speech Emotion Recognition based on Improved SOAR Model

Matin Ramzani Shahrestani [a,*] Sara Motamed [b,*] Mohammadreza Yamaghani [c]

[a] *Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran.*
[b] *Department of Computer Engineering, Fouman and Shaft Branch, Islamic Azad University, Fouman, Iran.*
[c] *Department of Computer Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran.*

**ARTICLE INFO.**

*Article history:*
**Received:** 18 December 2023
**Revised:** 26 June 2024
**Accepted:** 19 August 2024
**Published Online:** 19 August 2024

*Keywords:*
Speech Emotion Recognition, Convolutional Neural Network (CNN), Learning Automata, Improved SOAR Model.

**ABSTRACT**

In recent years, emotion recognition as a new method for human-computer interaction has attracted the attention of researchers. Automatic speech emotion recognition has become one of the practical methods to increase engagement in most industries. It is expected that emotion recognition based on audio information can result in better accuracy. The purpose of this article is to present an efficient method for recognizing emotional states from speech signals, based on a new cognitive model. Due to the importance of the topic, this article presents an efficient method for recognizing emotional states from speech signals based on a mixed deep learning and cognitive model called SOAR. To implement each part of this model, two main steps have been introduced. The first step is reading the video and converting it to images and preprocessing it. The next step is to use the combination of convolutional neural network (CNN) and learning automata (LA) to classify and detect the rate of facial emotional recognition. The reason for choosing CNN in our model is that no dimension is removed from the speech signal and considering the temporal information in dynamic speech leads to more efficient and better classification. Also, the training of the CNN network in calculating the backpropagation error is adjusted by LA so that the efficiency of the proposed model is increased and the working memory part of the SOAR model can be implemented. In the proposed model, audio databases available in the field of multimodal emotion recognition eNTERFACE' 05 and SAVEE have been used for various experiments. The recognition accuracy of the presented model in the best case from eNTERFACE' 05 and SAVEE databases is equal to 85.3% and 84.5%, respectively.

## 1 Introduction

Although Speech Emotion Recognition (SER) is becoming an active exploration area, the state-of-the-art performance is limited by the failure of emotional datasets. The preface of Multi-Task Learning(MTL), which concertedly trains multiple relative tasks, can use further emotion-related information within limited data, enhancing the SER performance but also adding the training difficulty. The emotion recognition system from the speech from the point of view of pattern recognition includes 3 corridors: point of birth, point selection, and bracket. Despite numerous advances in this field, there's a long gap between

* Corresponding author.
Email address: motamed.sarah@gmail.com (S. Motamed)
https://dx.doi.org/10.22108/jcs.2024.140141.1138
ISSN: 2322-4460

natural mortal-computer relations. The main reason for this issue is the computer's incapability to understand the stoner's passions. Thus, in the last many times, emotion recognition from speech is one of the grueling issues in the field of speech processing. Also, emotion recognition from speech can be used to prize useful meanings from speech. In addition, it can be used to fête the stoner's personality, which is used in numerous associations, including political, military, etc. Emotion recognition from speech has colorful operations that have been addressed in different ways. Including the development of automatic speech recognition systems, textbook-to-speech conversion, motorist internal state reports, and computer games [1, 2], as a tool for opinion in medical lores [3, 4], for impaired people or autistic children to communicate with others [5, 6], and it's also used as an operation in mobile phones [7].

The most important challenges of emotional recognition from speech are substantially at the birth stage. The main reason for the query about the effective features in feting the emotion of speech is the voice variety, which itself is caused by the actuality of colorful words, different speakers, and speaking styles. These factors also affect features uprooted from speech, similar to pitch wind and energy. The frequency of the main signals depends on the way the oral cords are stimulated, the gender of the speaker, and the physiological structure of the larynx.

In [8], we propose a primary-task driven adaptive loss function named PTDA- Loss to adaptively acclimate the task weights in the loss function, aiming to more nicely allocate the literacy coffers in the training process and further ameliorate the SER performance. We observe that inordinate concern with the supplementary task can lead to inadequate literacy about coffers for SER, which is not conducive to SER performance enhancement. Meanwhile, the literacy coffers for the supplementary task will be metal but not over-compressed, using the effective literacy of useful information from the supplementary task. To ameliorate the model's literacy capability, likewise, we first employ several dynamic convolutional layers as the backbone network. Regarding gender recognition and speaker recognition as the supplementary task independently, ablation trials on the IEMOCAP, MSP-IMPROV, and immingle datasets prove the rationality and effectiveness of PTDA- Loss. Also, relative trials demonstrate that our system achieves state-of-the-art performance on the three datasets, further enhancing the SER delicacy.

In [9], To validate our suppositions, we carry out an expansive experimental study fastening on the development of a model characterized by speaker/ gender independence, robustness to noise, and robustness against multiple voices. We probe a variety of audio features, classifiers, datasets, and data addition styles in the problem to define effective ways to address this under-delved yet socially significant problem. Our trials show that the proposed systems attain an F1 score of over 90 in the disruptive class, indeed, when introducing noisy rudiments similar to environmental noise or multiple lapping voices. Similar promising results indicate that framing disruptive situation discovery as a speech emotion recognition task could pave the way to the relinquishment of new types of intelligent systems with a positive impact on public safety.

In [10], a brain functionality-inspired model is presented for Face Image Compliancy verification (FICV) using the Haxby model, which is a face visual perception consistent model containing three bilateral areas for three different functionalities. As a result, the contribution of this work is presenting a new model, based on human brain functionality, improving the compliance verification of face images in the FICV context. Perceptual understanding of an image is the motivation of most of the quality assessment methods, i.e., human quality perception is considered a gold standard and a perfect reference for recognition and quality assessment. The model presented in this work aims to make the operational process of a face image quality assessment system closer to the performance of a human expert. Three basic modules have been introduced. Face structural information, for initial information encoding, is simulated by an extended Viola-Jones model. Face image quality assessment is presented by the International Civil Aviation Organization (ICAO), in the ICAO (ISO / IEC19794 -11) requirements compliance assessment document. Like the Haxby model, perception is performed through two distinct functional and neurological pathways, using Hierarchical Maximum pooling (HMAX) and Convolutional Deep Belief Networks (CDBN). The information storing and fetching for training are similar to their corresponding modules in the brain. For simulating brain decision-making, the final results of two separate paths are integrated by the weighting sum operator. Nine ISO / ICAO requirements were used for testing the model. The simulation results, using AR and PUT databases, show improvements in six requirements using the proposed method, in comparison with the FICV benchmark.

The model presented in this composition has two main features: a model for better representation of audio features and learning the time sequence of signals in trains, accompanying and introducing the ImprovedSOAR model for audio processes. The method of the proposed model is that first, pre-processing

is done on the dataset introduced in the textbook to convert the collection into a delivery and usable data. Also, on the regularized emotional speech signals, point birth is performed with MFCC speech signals.

Choosing MFCC on speech signals results in the necessary contraction and understanding of the speech diapason of the speech source and helps in the necessary identification and opinion. Next, to reduce the dimension and increase the effectiveness of the proposed system, the selection of features is grounded on a Convolutional neural network (CNN). CNN networks have been used to select the stylish features of speech because they're veritably resistant to changing the scale and size of lines and give a suitable response to the affair, indeed, in noisy signals. Experimenters proved that in the bracket of features, general features that have analogous arousal lead to failure in recognition [7]. For illustration, wrathfulness, fear, and pleasure have a high thrill, and sadness has a low thrill. The purpose of rooting pitch frequency details, including minimum, outside, standard, mean, and standard division, is to directly model the pitch wind, which seems to be veritably effective in distinguishing high-arousal feelings similar to wrathfulness from other low-arousal feelings. For this reason, LA has been used in the proposed system for accurate identification along with CNN.

To implement the SOAR cognitive science model, learning automata have been used. The learning automata (LA) has the capability of continuous learning throughout its life cycle and can improve learning with the considered reward mechanism and can learn through closely related examples. The main reason for using the LA in the proposed model is that this algorithm can improve its behavior by using its previous experience. The first position and action in this algorithm is random, and it will perform the next movements based on the response of the environment it receives.

The state changes in the symbolic long-term memory and short-term working memory sections based on environmental changes. So, learning automata is employed in the proposed method of its learning part to improve the Soar model.

The main parts of this article are organized in such a way that Section 2 is related to the review of works done on the recognition of emotional states of voice. The Section 3 presents the proposed ImprovedSOAR method and, finally, the test results and conclusions are presented in Sections 4 and 5 respectively.

## 2 Emotional States Recognition

### 2.1 Speech Emotion Recognition

The speech signal is the fastest and most natural way to show mutual communication between people. This fact has prompted researchers to use speech as a fast method of communication between machines and humans [11, 12]. Today, a lot of research has been conducted in the field of speech signal processing and its classification to recognize emotional speech based on machine learning models. For example, neural network (NN) [13], support vector machine (SVM) [14], k-nearest neighbor (KNN) [15, 16], hidden Markov model (HMM) [7], Gaussian mixture model (GMM) [17, 18], etc. For example, [19, 20] used the multi-layer perceptron (MLP) neural network model and seven types of emotions, and the highest recognition rate obtained in their experiments was 53%. Also, [4, 5] used classification models of SVM and NN on four types of emotions: anger, natural, happiness, and fear. The recognition rate in their experiments for SVM and NN models was 71% and 42% respectively. In the article [21], convolutional neural networks (CNN) have been used to extract speech features. The first layer of CNN is used for speech recognition and to extract features of emotional signals.

In the research presented [22], DCNNs, which consist of multilevel deep convolutional layers and pooling layers, were used. Because DCNNs include more parameters for time-frequency extraction, they are more accurate. Also, correlation information and learning ability with their strong features have led to better performance in shallow CNNs. In research [23], DCNN, AlexNet, was used to recognize emotional speech. The AlexNet model is trained using the large-scale ImageNet database [24]. The results showed less training time consumption with many different object classes. The trained structure includes AlexNet, consisting of an input layer, convolutional layers, pooling layers, and fully connected layers (FCLs). In the article [25], the convolution layer (Conv4) of the model was used to obtain features at a lower level.

### 2.2 Learning Automata (LA)

The LA, as one of the techniques of artificial intelligence, is a stochastic model that works in the framework of reinforcement learning. Automatic learning uses the selected action as input to its stochastic and operational environment. Reaching the ideal solution, the environment uses reinforcement feedback to respond to the approach of the adopted action to the ideal goal.

The action probability vector is updated using reinforcement feedback. Automatic learning tries to find

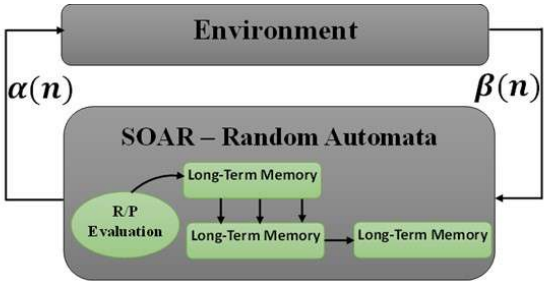


**Figure 1**. A View of the Interaction of the Environment And the Automatic Learning of Automata.

the optimal action from the set of actions to minimize the average penalty for the environment. In systems where complete information about the environment is not available, automatic learning can be useful [26].

Learning automata are classified into two main categories with a fixed and variable structure. Variable structure learning automata are represented by the triple $\{\beta, \alpha, T\}$, where $\beta$ is the set of inputs, $\alpha$ is the set of actions, and $T$ is the learning algorithm. As an iterative relation, the learning algorithm is used to modify the action probability vector. Figure 1 shows an interaction between the environment and the considered automatic learning.

Let $\alpha_i(n) \in \alpha$ and $P(n)$ represent the action selected by the learning automaton and the probability vector defined on the action set at moment $n$, respectively. Let $a$ and $b$ denote the reward and penalty parameters, and let the penalty determine the increase and decrease of action probabilities, respectively.

Also, consider $r$ as the number of actions that may be performed by machine learning. At every $n$ moment, the action probability vector $P(n)$ is updated by the linear learning algorithm presented in eq. (1) if the environment rewards the chosen action $\alpha_i(n)$. If the action taken is penalized, it will be updated according to eq. (2).

$$\begin{aligned} P_i(n+1) &= P_i(n) + a[1 - P_i(n)] \quad \text{if } j = i \\ P_j(n+1) &= (1-a)P_j(n) \qquad\qquad \forall j \neq i \end{aligned} \tag{1}$$

$$\begin{aligned} P_i(n+1) &= (1-b)P_i(n) \qquad\qquad \text{if } j = i \\ P_j(n+1) &= \tfrac{b}{r-1} + (1-b)P_j(n) \quad \forall j \neq i \end{aligned} \tag{2}$$

Repetition (1) and (2) are called linear reward-penalty (LR-P) algorithms if $a = b$, and the given equations are called linear-epsilon reward-penalty (LR-$\epsilon$P), where $a > b$. Finally, they are called linear reward-inaction (LR-I) where $b = 0$.

If the learning automaton chooses an action like $\alpha_i$ in the $n$th repetition and receives a favorable response from the environment, $p_i(n)$ (probability of action $\alpha_i$) increases and the probability of other actions decreases. On the contrary, if the response received from the environment is unfavorable, the probability of action $\alpha_i$ decreases and the probability of other automata actions increases. In any case, changes are made in such a way that the sum of $p_i(n)$ always remains constant and equal to one.

## 3 The Proposed Model

In this section, the proposed model is explained by providing details of the substructure of each component of it.

### 3.1 Improved SOAR Model

In Soar, all goal-oriented symbolic tasks are formulated in problem spaces. A problem space consists of a set of states and a set of operators. States represent states and operators represent actions that, when applied to states, yield other states. Each functional context includes a goal, plus roles for a problem state, a mode, and an operator. Problem-solving is driven by decisions that lead to the selection of problem spaces, states, and operators for the respective context roles. Given a goal, a problem space must be selected in which to pursue the goal. An initial state should then be chosen to represent the initial state. An operator must then be selected to apply to the initial state. Then another state must be selected (most likely the result of applying the operator to the previous state). This process continues until a sequence of operators is discovered that transforms the initial state into a state in which the goal is achieved. One of the subtle consequences of using problem spaces is that each one implicitly defines a set of constraints on how to do the work. For example, if eight puzzles are attempted in a problem space that contains only one-slide tile operator, all solution paths will obey the constraint that tiles are never removed from the board. Therefore, such conditions do not require explicit testing in the desired states [27, 28].

The input of the proposed model includes the emotional speech signals read from the input. Noise removal and resizing are performed on speech signals so that the pre-processed inputs enter the proposed model. In the memory section, a procedure from the

SOAR model will be applied to all normalized signals which MFCC to extract the best features.

All the outputs obtained from the previous step are entered into the symbolic memory where the best features will be selected. For this purpose, CNN-LA applied the features obtained from speech signals. The training of these networks, including CNN, in calculating the backward propagation error, is adjusted by LA so that the working memory part can be implemented. In the general description of LA, according to the error rate of CNN, it improves the updating parameters of the neural network weights in the backpropagation algorithm (Figure. 2).

Figure 2 shows the Improved SOAR model. As explained, the proposed model includes two inputs, and each network, with $W$ parameters, produces class membership probabilities $P(C \mid x, W)$ ($C$ classes according to observation $x$ of emotional state). All network layers, except the Softmax layers, had linear unit activation (ReLU) functions (eq. 3):

$$f(z) = \max(0, z) \tag{3}$$

We calculated the output of the Softmax layers as follows (eq. 4):

$$P(C \mid x, W) = \frac{\exp(z_C)}{\sum_q \exp(z_q)} \tag{4}$$

Where $z_q$ was the output of neuron $q$. The outputs of the network were $P(C \mid x, W)$. The training process of convolutional deep learning consists of optimizing network parameters $W$ to minimize a cost function for a dataset $D$. We chose the negative log probability as the cost function (eq. 5):

$$L(W, D) = -\frac{1}{|D|} \sum_{i=0}^{|D|} \log(P(C^{(i)} \mid x^{(i)}, W)) \tag{5}$$

However, the problem in learning neural networks is that this optimization problem is no longer convex. For this reason, one of the common methods of solving the optimization problem in neural networks is backtracking.

To improve the CNN, the probability theory of the LA in backpropagation error is used to train the Improved SOAR model to reduce the complexity of calculations and increase the speed compared to the traditional gradient descent method.

The output of LA, $\alpha_i$, is the possible actions that are used as the momentum factor to update the weights of the neural network. In this way, after several iterations, a strong connection between the best input and output features is created. This action adjusts the parameters adaptively in the gradient descent method of calculating the backpropagation error. In the following, the details of the implementation of the proposed model are described.

### 3.2 Preprocessing

To recognize emotion from speech signals in the video file, database preparation, and pre-processing operations are performed in the first stage. These steps are as follows: in the database preparation phase, the database set is divided into two training and testing sections. So, 70% of the database set is considered for training and 30% for testing. In the next step, speech signals are extracted from each video sample and stored in the folders related to speech frames, respectively. To extract the image frames and speech signal from the video, the separation algorithm with a certain frame rate was applied to all the samples. In these tests, on average, the length of eNTERFACE' 05 database videos is 3 to 4 seconds. Its speech signal is recorded at a sampling rate of 48000 Hz with a resolution of 16 bits and one channel [29].

## 4 Results and Experiments

All the tests of this article are implemented by Python software and the accuracy test is done by the K-Fold method with K=10. The measurement criterion of each part is recognition confusion matrices [30]. In the speech section, 10 sequences of 2 milliseconds were separated from the clip, and each of them was converted into a 100x96 image using the Mel Spectrogram function and fed to the CNN-LA neural network.

In the generated convolutional neural network CNN-LA, the size of each batch is 32, and the number of repetition steps is 500. The learning rate is 0.001 and the momentum rate is 0.0001. In this method, the early stopping technique is used during training.

### 4.1 Speech Emotion States Recognition

Table 1 shows the confusion matrix of the recognition rate of voice emotional states using ImprovedSOAR on the eNTERFACE' 05 database.

As can be seen in Table 1, the recognition rate of voice emotions using the ImprovedSOAR model is completely different. Examining the results revealed that the emotion "Disgust" has the highest recognition rate and the emotion "fear" has the lowest recognition rate with values of 89% and 80%, respectively.

The reason for such results can also be that the

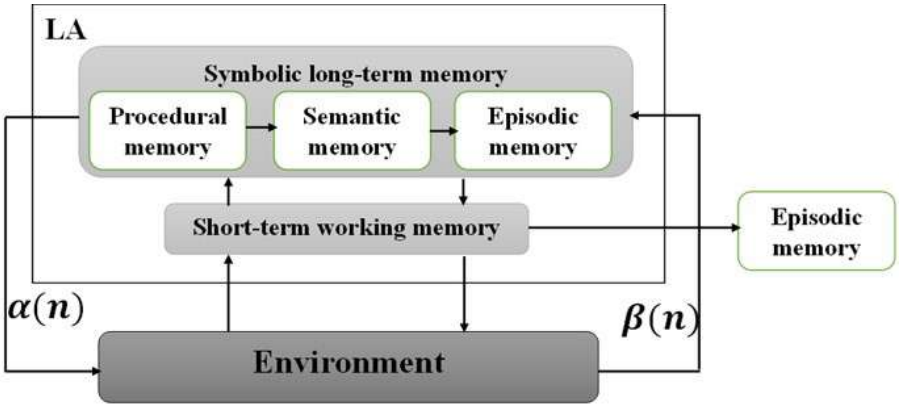


**Figure 2**. Overview of the Improved SOAR Model.

**Table 1**. Confusion Matrix of Recognition of Auditory Emotional States Using Improved SOAR.

| | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| Anger | 85% | 2% | 3% | 2% | 3% | 5% |
| Disgust | 3% | 89% | 3% | 1% | 2% | 2% |
| Fear | 2% | 3% | 80% | 6% | 7% | 2% |
| Happiness | 3% | 2% | 3% | 88% | 2% | 2% |
| Sadness | 2% | 4% | 2% | 2% | 87% | 3% |
| Surprise | 2% | 4% | 3% | 5% | 3% | 83% |

above recognizable cases are mostly an obvious emotion and these cases are visible in speech. Like Disgust, which has a high percentage of detection rate. But an emotion like fear, which has the lowest value, is not visible in speech, and many people can hide this type of emotion behind their voices.

In the evaluation of the proposed method, in addition to the eNTERFACE' 05 database, the SAVEE database has been used. The SAVEE database recruited four male native English speakers (identified as DC, JE, JK, and KL), postgraduate students, and researchers at the University of Surrey aged 27–31 years [29].

Table 2 shows the confusion matrix based on ImprovedSOAR to recognize audio-emotional states in six basic emotion classes on the SAVEE database.

As can be seen in Table 2, the emotional state of "surprise" has the highest, and the emotional state of "Sadness" has the lowest detection accuracy using

**Table 2**. Confusion matrix of the recognition of auditory emotional states using the proposed model on the SAVEE database.

| | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| Anger | 88% | 0% | 4% | 4% | 4% | 0% |
| Disgust | 0% | 89% | 0% | 5% | 6% | 0% |
| Fear | 2% | 6% | 82% | 6% | 2% | 2% |
| Happiness | 4% | 5% | 5% | 81% | 5% | 0% |
| Sadness | 10% | 8% | 4% | 4% | 74% | 0% |
| Surprise | 0% | 7% | 0% | 0% | 0% | 93% |

the proposed model. Figure 3 shows the results of the eNTERFACE' 05 and SAVEE databases together.

As shown in Figure 4, in the eNTERFACE' 05 database, the emotional state of disgust with a detection rate of 89%, and in the SAVEE database the emotional state of surprise with a detection rate of 93% have the highest recognition rates. Also, in the eNTERFACE' 05 database, the emotional state of fear with a detection rate of 80%, and in the SAVEE database, the emotional state of discomfort with a detection rate of 74% shows the lowest recognition rate. Another experiment that has been done in this article is applying different fusion methods to fuse the selected features at the decision-making level.

As the details of the proposed method are shown in Figure 2, the proposed method uses decision-level fusion to maintain the independence of each of the methods and uses the advantages of each at the final decision level. This causes the separate benefit of each

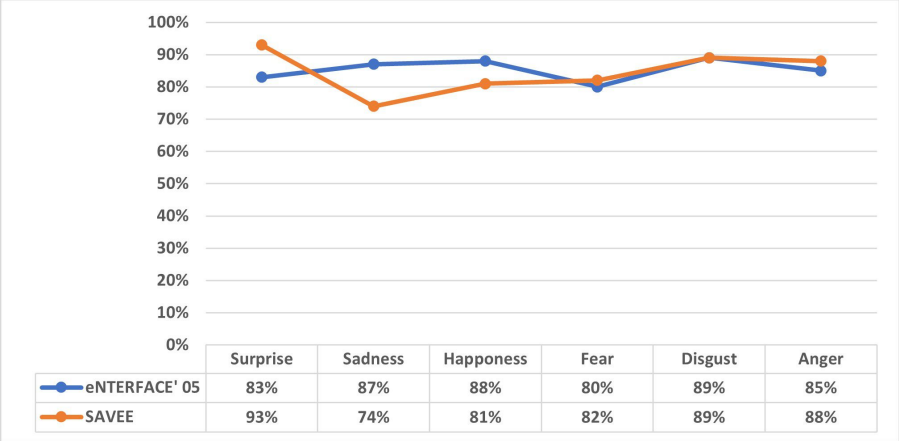


**Figure 3**. Comparison of recognition of auditory emotional states using the proposed model on eNTERFACE' 05 and SAVEE databases.

classification and the final improvement to produce better answers.

For integration at the decision level, various methods of combination rules have been tested, including minimum, maximum, average, product, and combination. Each of the aforementioned combination rules is tested for integration at the decision level in the proposed model, and the performance results of each are shown in Table 2. According to Table 2, the "product" combination rule has the highest recognition accuracy due to how the product of the output points of different funds obtains the maximum class score.

As shown in Figure 4, the fusion method at the level of decision-making shows a minimum accuracy of 71.83%, with a maximum of 85% in total combinations, and an average value of 78.4%. Among the presented methods, the best fusion detection rate at the decision-making level is related to the product method with a rate of 85.3%, which we will choose for the proposed method.

The next stage of the experiments is to apply the proposed model, without considering the phase of extracting the features of the evidence. Table 3 shows the confusion matrix of recognition of emotional states without PCA and MFCC feature extraction.

According to Table 3, the highest recognition rate belongs to the emotional state of Disgust (85.1%), whereas the emotional state of Fear had the lowest recognition rate (70.7%). In comparison to Table 3, the recognition rate obtained is much lower without feature extraction.

## 4.2 Emotion States Recognition

To check the performance of the proposed model, all tests on the single modalities of facial emotional states and speech emotional states are conducted separately and the results are discussed.

**Table 3**. Confusion matrix for the recognition of audio emotional states using the proposed model without considering the feature extraction stage on the eNTERFACE' 05 database.

| | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| Anger | 82% | 7.1% | 2.3% | 2.3% | 3% | 3.3% |
| Disgust | 2.3% | 85.1% | 2.3% | 2.3% | 4.7% | 3.3% |
| Fear | 7.3% | 6% | 70.7% | 6% | 6% | 4% |
| Happiness | 7.1% | 3.3% | 4.7% | 75.6% | 3.6% | 5.7% |
| Sadness | 2% | 1% | 7.5% | 2% | 78% | 9.5% |
| Surprise | 2.3% | 3.8% | 4.7% | 7.1% | 7.1% | 75% |

## 4.3 Speech Emotional States Recognition

To recognize the emotional states of speech, first in the pre-processing stage, the speech signal is extracted from the video, and noise removal is done on it. The output of this step is entered into the next step and will be applied in the feature extraction step from MFCC. CNN-LA is then applied to calculate the recognition rate of speech-based emotional states. Table 4 shows the confusion matrix of the presented model to recognize the emotional states of speech in six basic emotion classes from the eNTERFACE' 05 database.

As can be seen in Table 4, the emotional states of "anger" and "fear" show the highest and lowest recognition rates, respectively, for recognizing emotional speech. According to studies, factors affecting vocal expressions, such as anger, tend to have a higher recognition rate, while internal emotional states, such as fear, have a significantly lower recognition rate.

The general knowledge added to BERT, SentiGAN, and CGAN increases diversity. But, as mentioned in

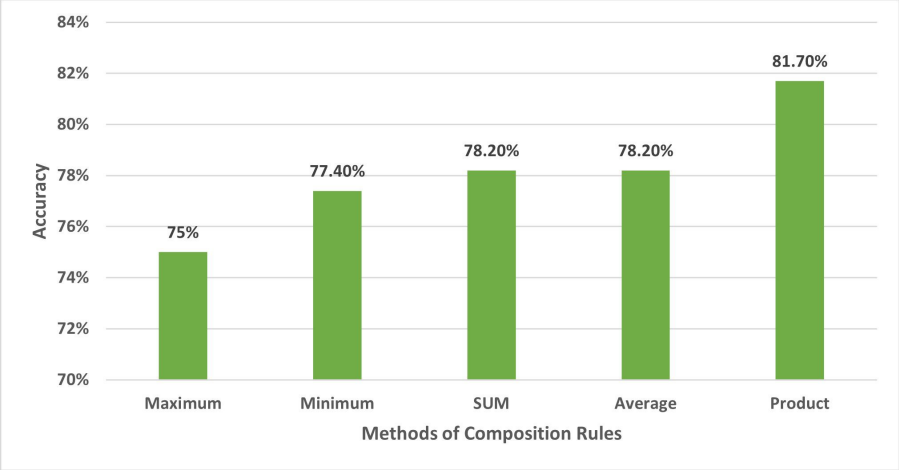


**Figure 4**. Comparison of the accuracy of bimodal emotion recognition based on integration at the decision level between different methods of combined rules.

**Table 4**. Confusion matrix of recognition of speech emotional states using the proposed model.

| | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| Anger | 83% | 6% | 0% | 3% | 3% | 5% |
| Disgust | 10% | 65% | 8% | 5% | 6% | 6% |
| Fear | 3% | 4% | 64% | 12% | 12% | 5% |
| Happiness | 10% | 2% | 3% | 74% | 3% | 8% |
| Sadness | 2% | 4% | 8% | 0% | 76% | 10% |
| Surprise | 4% | 8% | 3% | 11% | 5% | 69% |

previous sections, the general dataset significantly reduces the quality of the generated sentences and their relation to the desired aspects.

## 5 Conclusions

With the increasing interaction between machines and humans and the need to understand human emotions to advance the goals of industries, the discovery of human emotional states plays a fundamental role. According to studies, the recognition of human emotions is not reliable without voice processing. As a result, several methods should complement each other to increase the detection performance. In this paper, a proposed method using two-channel verbal signals is used. CNN-LA has been used in SOAR to increase the recognition rate. The biological structure of SOAR combined with the improvement of CNN-LA learning automata provided more learning power for understanding complex knowledge such as facial expressions. The proposed method in this paper classified each of the verbal channels using Improved SOAR classification. In the last stage, this classification produced an output using the fusion of information of the decision level and presented the amount of recognition of the expression of auditory emotional states calculated. The highest and lowest rates of recognition, based on voice, respectively belonged to the expression of hatred with 89% and fear with 80%. The proposed model has recorded an average improvement of 4.2% compared to the closest model.

## References

[1] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–577. IEEE, 2004. doi:10.1109/ICASSP.2004.1326051.

[2] Zengzhao Chen, Mengting Lin, Zhifeng Wang, Qiuyu Zheng, and Chuan Liu. Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms. *Knowledge-Based Systems*, 281:111077, 2023. doi:10.1016/j.knosys.2023.111077.

[3] M. J. Landau. Acoustical properties of speech as indicators of depression and suicidal risk. *Vanderbilt Undergraduate Research Journal*, 4, 2008. doi:10.1109/10.846676.

[4] V. Singh and S. Prasad. Speech emotion recognition system using gender dependent convolution neural network. *Procedia Computer Science*, 218:2533–2540, 2023. doi:10.1016/j.procs.2023.01.227.

[5] D. Ververidis and C. Kotropoulos. A review of emotional speech databases. 2003. doi:10.26634/jcom.10.3.19103.

[6] J. de Lope and M. Graña. An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11, 2023. doi:10.1016/j.neucom.2023.01.002.

[7] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011. doi:10.1016/j.patcog.2010.09.020.

[8] L. Y. Liu, W. Z. Liu, and L. Feng. A primary task driven adaptive loss function for multi-task speech emotion recognition. *Engineering Applications of Artificial Intelligence*, 127:107286, 2024. doi:10.1016/j.engappai.2023.107286.

[9] E. Mancini, A. Galassi, F. Ruggeri, and P. Torroni. Disruptive situation detection on public transport through speech emotion recognition. *Intelligent Systems with Applications*, page 200305, 2023. doi:10.1016/j.iswa.2023.200305.

[10] A. Nourbakhsh, M. S. Moin, and A. Sharifi. An inspired haxby brain perceptual model for facial images quality assessment. *Journal of Intelligent and Fuzzy Systems*, 39(6):8543–8555, 2020. doi:10.3233/JIFS-189171.

[11] C. Gan, K. Wang, Q. Zhu, Y. Xiang, D. K. Jain, and S. García. Speech emotion recognition via multiple fusion under spatial–temporal parallel network. *Neurocomputing*, 555:126623, 2023. doi:10.1016/j.neucom.2023.126623.

[12] M. R. Falahzadeh, F. Farokhi, A. Harimi, and R. Sabbaghi-Nadooshan. Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition. *Circuits, Systems, and Signal Processing*, 42(1):449–492, 2023. doi:10.1007/s00034-022-02130-3.

[13] A. Harimi, A. Shahzadi, A. Ahmadyfard, and K. Yaghmaie. Classification of emotional speech using spectral pattern features. *Journal of AI and Data Mining*, 2(1):53–61, 2014. doi:10.22044/jadm.2014.150.

[14] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson. Speech emotion recognition in emotional feedback for human-robot interaction. *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, 4(2):20–27, 2015. doi:10.14569/IJARAI.2015.040204.

[15] J. Winter, Y. Xu, and W. C. Lee. Energy efficient processing of k nearest neighbor queries in location-aware sensor networks. In *The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, pages 281–292. IEEE, 2005. doi:10.1109/MOBIQUITOUS.2005.28.

[16] E. Jing, Y. Liu, Y. Chai, J. Sun, S. Samtani, Y. Jiang, and Y. Qian. A deep interpretable representation learning method for speech emotion recognition. *Information Processing and Management*, 60(6):103501, 2023. doi:10.1016/j.ipm.2023.103501.

[17] M. Sheikhan, M. Bejani, and D. Gharavian. Modular neural-svm scheme for speech emotion recognition using anova feature selection method. *Neural Computing and Applications*, 23:215–227, 2013. doi:10.1007/s00521-012-0814-8.

[18] Z. T. Liu, B. H. Wu, M. T. Han, W. H. Cao, and M. Wu. Speech emotion recognition based on meta-transfer learning with domain adaption. *Applied Soft Computing*, 147:110766, 2023. doi:10.1016/j.asoc.2023.110766.

[19] S. Haq, P. J. Jackson, and J. Edge. Audio-visual feature selection and reduction for emotion classification. In *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08)*, Tangalooma, Australia, 2008.

[20] R. V. Darekar, M. Chavan, S. Sharanyaa, and N. M. Ranjan. A hybrid meta-heuristic ensemble based classification technique speech emotion recognition. *Advances in Engineering Software*, 180:103412, 2023. doi:10.1016/j.advengsoft.2023.103412.

[21] P. Tzirakis, J. Zhang, and B. W. Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093. IEEE, 2018. doi:10.1109/ICASSP.2018.8462677.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. doi:10.1145/3065386.

[23] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5. IEEE, 2017. doi:10.1109/PlatCon.2017.7883728.

[24] 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*, 20(21):6008, 2020. doi:10.3390/s20216008.

[25] Z. Farhoudi, S. Setayeshi, and A. Rabiee. Using learning automata in brain emotional learning for speech emotion recognition. *International Journal of Speech Technology*, 20:553–562, 2017. doi:10.1007/s10772-017-9426-0.

[26] S. Wongthanavasu and J. Ponkaew. A cellular automata-based learning method for classification. *Expert Systems with Applications*, 49:99–

111, 2016. doi:10.1016/j.eswa.2015.12.003.

[27] M. Ramzani Shahrestani, S. Motamed, and M. Yamaghani. Recognition of facial emotion based on soar model. *Frontiers in Neuroscience*, 18:1374112, 2024. doi:10.3389/fnins.2024.1374112.

[28] M. Ramzani Shahrestani, S. Motamed, and M. Yamaghani. Recognition of facial and vocal emotional expressions by soar model. *Journal of Information Systems and Telecommunication (JIST)*, 3(43):209, 2023. doi:10.61186/jist.39828.11.43.209.

[29] A. S. Cecchi. Cognitive penetration of early vision in face perception. *Consciousness and Cognition*, 63:254–266, 2018. doi:10.1016/j.concog.2018.06.005.

[30] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006. doi:10.1016/j.specom.2006.04.003.

**Matin Ramzani Shahrestani** received her bachelor's degree in computer software engineering in 2014 from Islamic Azad University, Lahijan branch, and her master's degree from Gilan University in 2014. Her favorite research fields are: cognitive science, data analysis, analytical database, and data mining. Her email address is: matinramzani@gmail.com

**Sara Motamed** received her PhD in Computer Engineering in 2016 from Islamic Azad University, Science and Research Branch of Tehran. She is currently an assistant professor in the computer engineering department of Islamic Azad University, Fuman and Shaft branch. Her research areas are artificial intelligence, image processing, machine vision and data science. Her email address is: motamed.sarah@gmail.com

**Mohammad Reza Yamaghani** received his Ph.D in the field of artificial intelligence in 2015 from Islamic Azad University, Sciences and Research University of Tehran. He is currently an assistant professor in the computer engineering department of Islamic Azad University, Lahijan branch. His research areas are artificial intelligence, image processing, compression and data mining. His email address is: O_yamaghani@iau.ac.ir