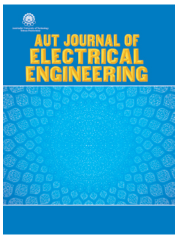


AUT Journal of Electrical Engineering

AUT J. Elec. Eng., 56(2) (2024) 213-226
DOI: 10.22060/eej.2024.22903.5570

# A Novel Multi-Task and Ensembled Optimized Parallel Convolutional Autoencoder and Transformer for Speech Emotion Recognition

Z. Sharifzadeh Jafari, S. Seyedin*

Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

**ABSTRACT:** Recognizing the emotions from speech signals is very important in different applications of human-computer-interaction (HCI). In this paper, we present a novel model for speech emotion recognition (SER) based on new multi-task parallel convolutional autoencoder (PCAE) and transformer networks. The PCAEs have been proposed to generate high-level informative harmonic sparse features from the input. With the aid of the proposed parallel CAE, we can extract nonlinear sparse features in an ensemble manner improving the accuracy and the generalization of the model. These PCAEs also address the problem of the loss of initial sequential information during convolution operations for SER tasks. We have also proposed using a transformer in parallel with PCAEs to gather long-term dependencies between speech samples and make use of its self-attention mechanism. Finally, we have proposed a multi-task loss function made up of two terms of classification and AE mapper losses. This multi-task loss tries not only to reduce the classification error but also the regression error caused by the PCAEs which also work as mappers between the input and output Mel-frequency-cepstral-coefficients (MFCCs). Thus, we can both focus on finding accurate features with PCAEs and improving the classification results. We have evaluated our proposed method on the RAVDESS SER dataset in different terms of accuracy, precision, recall, and f1-score. The average accuracy of the proposed model on eight emotions outperforms all the recent baselines.

**Review History:**

Received: Dec. 30, 2023
Revised: Jan. 14, 2024
Accepted: Jan. 15, 2024
Available Online: Mar. 01, 2024

**Keywords:**

Speech emotion recognition
Mel Frequency Cepstral Coefficients
Autoencoder
Transformer
Multi-Task deep learning

## 1- Introduction

Emotions serve as a window into human psychology, revealing both their underlying mental state and the true intent behind their words. Recognizing this is a critical step in cultivating more meaningful and harmonious interactions between humans and machines[1]. Speech, facial expressions[2], and EEG signals[3] are all forms of expressing emotion. Among these, EEG signals provide the clearest indication of emotion[4]. However, they cannot be used in everyday human-computer dialogue. Also, appearance alone cannot reliably reveal a person's true emotions, as someone may outwardly seem happy but inwardly be upset or sad. Despite the fact that speech is the most natural and widely used form of communication, a person's mental and psychological state can significantly impact their speech. This is why speech-based systems have gained particular importance in the field of emotion recognition.

Speech emotion recognition (SER) systems have a wide range of applications across diverse domains, including healthcare[5], cognitive science[6], psychology[7], marketing[8], call centers[9], lie detectors, voice assistants and the entertainment industry. For instance, SER systems can be utilized to identify the emotions of pilots in the cockpit or drivers on the road to enhance safety[10, 11], monitor patient well-being and detect early signs of distress in healthcare settings, develop more effective treatments for mental health disorders, create more engaging, personalized marketing campaigns and utilizing emotions detected from an actor's speech to convey the emotional state of a particular scene in subtitles and suggest background music.

There are considerable challenges and complexities in enhancing the performance of these systems. First, emotions are inherently subjective and multifaceted, often perceived and expressed differently by individuals. Additionally, the vocal anatomy of the larynx and mouth varies significantly among individuals, further complicating accurate emotion recognition. Moreover, factors like age, gender, language, and dialect introduce even greater variations in vocal patterns. Furthermore, humans rarely express emotions in a purely basic, one-dimensional manner. Instead, emotions often manifest as complex blends of multiple emotions, making it difficult for SER systems to accurately identify the predominant emotion. Emotional speech data is another stumbling block in this field. In the field of emotion recognition from speech, there are two types of databases: natural and acted [12]. Natural databases are people's

*Corresponding author's email: sseyedin@aut.ac.ir everyday conversations. Analyzing natural databases has yielded positive results. Nevertheless, accurately capturing everyday speech is challenging. Actors in acted databases are asked to say different sentences with different emotions. As a result, training systems on acted data is not as accurate as training them on real-world data because it fails to capture the natural variation of human speech [13]. Moreover, the development of SER systems is hampered by the need for extensive labeled data, which requires a significant amount of time and effort from professional annotators to classify and label audio recordings with corresponding emotions.

In this paper, we propose a novel method to learn a representation of Mel-Frequency Cepstral Coefficient (MFCC) features for speech emotion recognition. Speech signals are composed of both temporal and spatial information, and both types of information must be considered to achieve an effective representation. To consider the limited amount of available data and the need for generalizable models, we propose employing a convolutional autoencoder with a transformer to extract robust representations from speech data. We have proposed the transformer to enable the model to access both temporal features and the auxiliary task of reconstructing the MFCC from the feature vector which enhances the system's generalization. We have proposed the autoencoder to provide the system with multi-task training ability. To this end, a multi-task loss function composed of two tasks of emotion classification and accurately mapping speech information in the autoencoder (AE) is proposed. Simple convolutional models often struggle to achieve the required accuracy in speech-emotion classification due to the loss of initial sequential information during convolution operations. To address this issue, we suggest two parallel convolutional networks to extract sequential information leading to high-level features. We extract ensembled high-level features by this proposed parallel network. In other words, we train each convolutional autoencoder network with different random variables. Then, we finally insert both inputs for the classifier to make use of the ensemble learning technique. It is worth mentioning that we extract appropriate non-linear sparse features by this autoencoder. Therefore, we propose using sparse features almost free of irrelevant information from the bottleneck layer of AE leading to informative features for emotion recognition. Suggesting these non-linear sparse features by ensemble learning in a multi-task manner for speech emotion recognition is an important contribution of this paper. Additionally, white noise is introduced to the training data to enhance the model's generalization and noise robustness.

The paper is structured as follows: Section 2 reviews related work. Section 3 describes the proposed method. Section 4 presents the experiments and results as well as discussions. Finally, we conclude the paper in Section 5.

## 2- Related Work of SER

A typical approach to human speech emotion recognition has three stages of feature extraction, feature selection or dimension reduction, and classification[14]. The first two stages of feature extraction and feature dimension reduction can represent the data appropriately, which would solve many problems in this field. MFCC coefficients were used as features in the earlier works [15]. Later, MFCC or some other features such as non-negative matrix factorization (NMFCC) were used as inputs to machine learning classifiers including Support Vector Machine (SVM); Linear Discriminant Analysis; Hidden Markov Models[16-19].

The new ideas and developments in the speech recognition industry and its subfields that followed the birth of deep neural networks have made a big revolution to improve the results in this field. A very simple type of neural network known as a Multi-Layer Perceptron was suggested in [20] for emotion recognition based on age and gender as auxiliary features for discovering emotions from labeled data. In [21], the authors used a type of Convolutional Neural Network (CNN) architecture. This method was significantly superior to the Support Vector Machine method as the experiment results showed. In [22], emotion classification from spectrograms is done with pre-trained convolutional architectures such as AlexNet and VGG by transfer learning. A CNN trained on the dataset in [23] is used to extract features from speech spectrograms. These features are then used to train an SVM classifier to recognize speech emotions. In [24], a CNN-based model was proposed that uses two convolutional layers with different kernel sizes to extract horizontal and vertical features from speech spectrograms. The output of these layers is then concatenated and fed to a fully connected layer for classification. The authors in [25] explored an alternative approach, utilizing a fully convolutional neural network (FCN) devoid of a dense layer, enabling the model to handle audio samples of varying durations. They trained the model using both Mel-spectrograms and MFCCs, discovering that MFCCs yielded superior performance.

Recently, Researchers have increasingly focused on methods that can automatically learn discriminative and high-level representations of speech data for emotion recognition. In [26], a sparse autoencoder-based method for transfer learning of features in the field of speech emotion recognition is proposed. After MFCC feature extraction, in [27], they implemented a simple autoencoder and stacked autoencoder model. In [28], the authors proposed the use of autoencoders for the harmonization of heterogeneous extracted features. They argued that using a wide range of features in emotion recognition reduces the final accuracy due to their heterogeneity. They presented a model that removes heterogeneous acoustic features that may contain redundant and irrelevant information.

After Google successfully employed attention mechanisms to enhance machine translation[29], attention has been incorporated into a wide range of deep learning models. Recent research [30] shows that the use of recurrent networks, such as LSTM (Long-Short-Term-Memory) with a directional attention layer, can perform well compared to other deep learning methods. Attention-based CNN models [31] have also shown good performance. In a recent study, Xu et al. [32] combined a multi-head attention-based approach

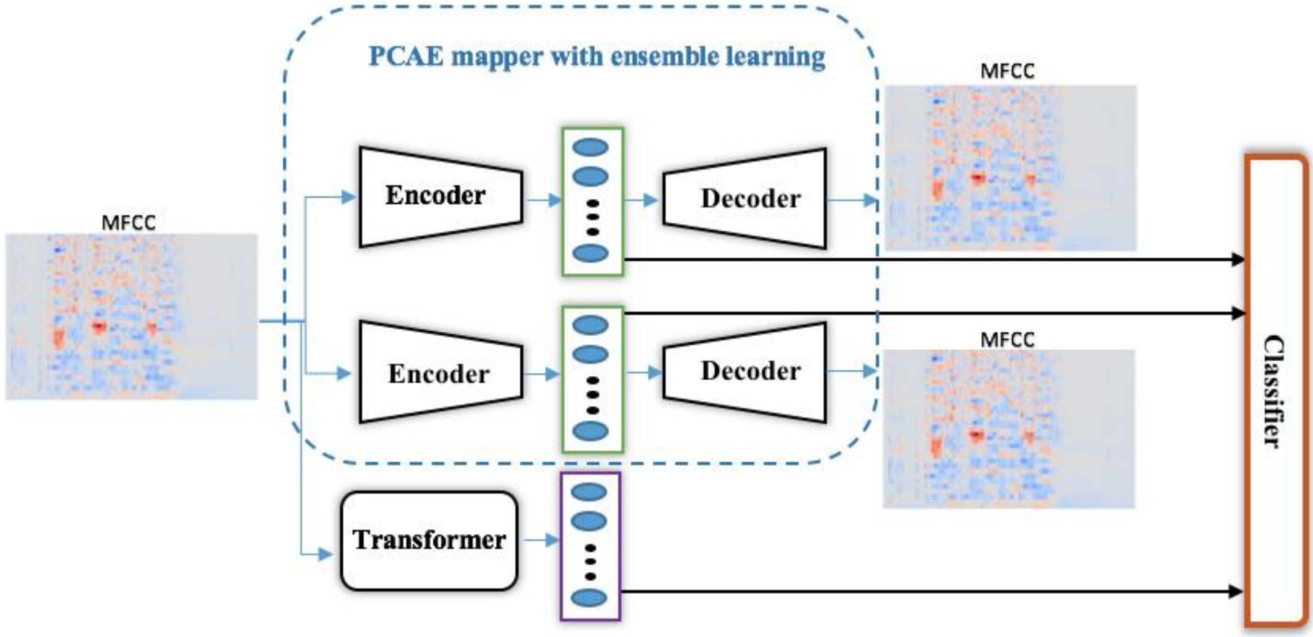


**Fig. 1. The proposed model composed of parallel convolutional autoencoders and transformer with multi-task training**

with a CNN. High accuracy rates were achieved using spectral features.

## 3- Proposed Model

In this section, we describe our proposed model as illustrated in Fig. 1, which is suggested for improving the performance of self-supervised convolutional autoencoder speech emotion recognition systems, that can select and extract higher-level features from MFCCs. This network is trained in a multi-task manner so that reconstructing the input spectrogram as an auxiliary task is defined along with feature selection and extraction for classification. Also, to extract proper temporal features from the input, we have proposed a transformer.

We explain the main blocks of our proposed model in the following subsections.

### 3- 1- Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients are one of the most widely used features in this field. In this research, we use this feature because MFCCs are less sensitive to changes in the environment such as background noise and channel distortion [33]. Also, the mel frequency bank is consistent with the human cochlear frequency response, which causes the system to become well acquainted with the way humans perceive and distinguish sounds.

### 3- 2- Parallel Convolutional Autoencoder Network with Ensemble Learning

Autoencoder networks are a type of unsupervised networks and do not require labeled data for their training. In this network, an output is tried to be made as similar as possible to the input. A key feature of autoencoders is their ability to reduce the dimensionality of the input data in their bottleneck layer. Convolutional autoencoders utilize convolutional layers to generate a compressed representation of the image [34]. Usually, convolutional autoencoders are used for reducing and compressing the input image and for eliminating noise while trying to retain important information [35-37]. More precisely, a convolutional autoencoder consists of two convolution-based models, the encoder and decoder, as depicted in Figure 2 The encoder is mainly used for encoding the initial input image into a hidden representation of smaller dimensions called the bottleneck layer. However, the decoder's job is to restore the compressed hidden representation into an output image of equal dimensions to the original image.

We consider the vector x as the input data of the convolutional autoencoder (CAE) network such that:

$$x = X^{M \times H \times W \times C} \tag{1}$$

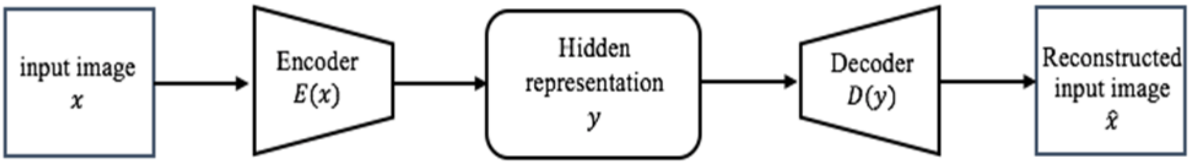


**Fig. 2. Convolutional autoencoder diagram**

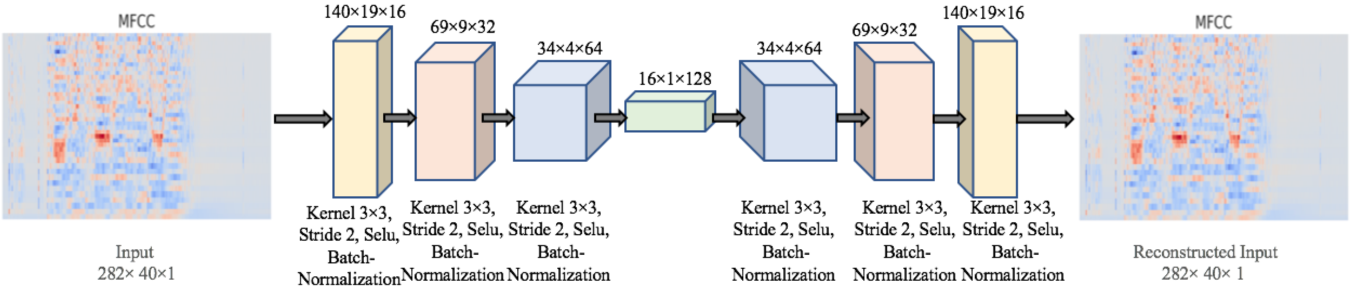

MFCC
Input
282× 40×1
140×19×16
69×9×32
34×4×64
16×1×128
34×4×64
69×9×32
140×19×16
Kernel 3×3, Stride 2, Selu, Batch-Normalization
Kernel 3×3, Stride 2, Selu, Batch-Normalization
Kernel 3×3, Stride 2, Selu, Batch-Normalization
Kernel 3×3, Stride 2, Selu, Batch-Normalization
Kernel 3×3, Stride 2, Selu, Batch-Normalization
Kernel 3×3, Stride 2, Selu, Batch-Normalization
MFCC
Reconstructed Input
282× 40× 1

**Fig. 3. Proposed convolutional autoencoder architecture**

where M is the number of samples, H is the height of a sample, W is the width of a sample and C is the number of channels in input samples. The encoder and decoder outputs are as follows:

$$\begin{cases} y = E(x) \\ \hat{x} = D(y) \end{cases} \tag{2}$$

The performance of the convolutional autoencoder can be measured by the reconstruction error, $\boldsymbol{e^{CAE}}$, which is calculated as follows:

$$e^{CAE} = L_{CAE}\left(\left(\hat{x}^{(k)}\right), x^{(k)}\right) \tag{3}$$

$\boldsymbol{L_{CAE}}$ denotes the squared Euclidean distance defined as follows:

$$L_{CAE}\left(\left(\hat{x}^{(k)}\right), x^{(k)}\right) = \frac{1}{2}\left\|\hat{x}^{(k)} - x^{(k)}\right\|^2 \tag{4}$$

Then the cost function can be shown in its general form as follows:

$$e^{CAE} = \frac{1}{M}\sum_{k=1}^{M} L_{CAE}\left(D\left(E\left(\left(\hat{x}^{(k)}\right)\right)\right), x^{(k)}\right) \tag{5}$$

In our proposed method, we use two convolutional autoencoders with architecture illustrated in Figure 3, in parallel to generate high-level features from the bottleneck layer of each AE without losing important information leading to a parallel CAE (PCAE). This parallel network extracts ensembled high-level features, i.e. we train each network with different random variables and finally use both as the inputs for the classifier making use of the ensemble learning procedure. In addition, the proposed network architecture can help reduce the local minimum problem by learning different features and combining them while it is being initialized by different values in the proposed PCAE. This autoencoder can also be considered as an appropriate network for extracting non-linear sparse features. Thus, its bottleneck layer represents the sparse features that have been proven to carry informative features and harmonic structure of speech signals for emotion recognition [19].

In other words, the bottleneck layer has much fewer features leading to more informative coefficients almost

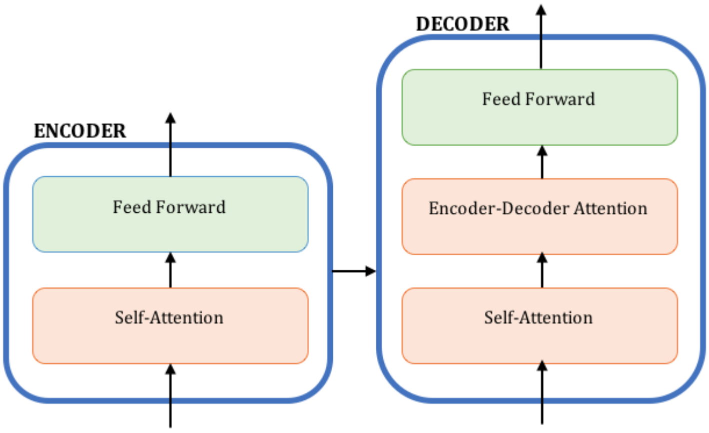


**Fig. 4. The block diagram of the transformer**

without irrelevant information. This is helpful for the final purpose of classification.

The cost function of the parallel autoencoder, $L_{PCAE}$, is as follows:

$$L_{PCAE}\left(\left(\hat{x}^{(k)}\right),x^{(k)}\right) = L_{CAE1}\left(\left(\hat{x}^{(k)}\right),x^{(k)}\right) + L_{CAE2}\left(\left(\hat{x}^{(k)}\right),x^{(k)}\right) \qquad (6)$$

$$e^{PCAE} = \frac{1}{M}\sum_{k=1}^{M} L_{CAE1}\left(D_1\left(E_1\left(\left(\hat{x}^{(k)}\right)\right)\right),x^{(k)}\right) + \frac{1}{M}\sum_{k=1}^{M} L_{CAE2}\left(D_2\left(E_2\left(\left(\hat{x}^{(k)}\right)\right)\right),x^{(k)}\right) \qquad (7)$$

where $L_{CAE1}$ and $L_{CAE2}$ are the cost functions of the first and second convolutional autoencoder networks.

### 3- 3- Transformer Encoder

The transformer is much more effective in speech emotion recognition than recurrent networks. Recurrent networks are not able to detect the subtle changes in tone, amplitude, and pitch which frequently express emotions, as well as the transformer can. It is also capable of measuring long-term dependencies in speech sequences, which are necessary for inferring emotion from a long sequence of time steps. Moreover, it can also take several steps of time at once along the speech series Thus, it is able to grasp how emotion is distributed over the whole segment of speech. It is also computationally more efficient. The first sublayer is a self-attention layer and the second is a feedforward neural network. As illustrated in Figure 4, Each encoder has two sublayers. This encoder input first travels through a self-attention layer which allows the encoder to concentrate on other parts at the same time that it is encoding one part. The output of the self-attention layer is passed to a feedforward neural network layer. The attention mechanism that allows the transformer to relate different elements of an input is computed as follows[29]:

$$Attention(Q,\mathrm{K},V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{n}}\right)V \qquad (8)$$

In this formula, $Q$, $K$, and $V$ are input matrices. $Q$ is the query, $K$ is the key, and $V$ is the value matrices. $n$ is a constant parameter which is the dimension of the key space. The Softmax function normalizes the attention values.

Multi-head attention is a technique to compute attention in a language model, letting the model understand relationships between different parts of a sequence from different points of view. The multi-head attention approach computes the attention mechanism several times using different weight matrices. These weight matrices enable the model to concentrate on different aspects of the relations. It is defined as follows[29]:

$$MultiHead(Q,K,V) = [head_1;\ head_2; \ldots\ head_m\ ]W^o \qquad (9)$$

$$head_i = Attention\left(QW_i^Q, KW_i^k, VW_i^V\right) \quad (10)$$

The $W$ matrices here are the trainable weight matrices generated by the model. The proposed method uses four stacked identical blocks of the transformer encoder for feature extraction. Each block includes a multi-head self-attention layer and a fully connected neural network feedforward layer.

### 3- 4- Multi-Task Learning

In multi-task learning, the model is trained simultaneously on multiple tasks. This shared representation allows the model to learn all of these tasks more effectively. Therefore, multi-task learning can improve the performance of the model on all the desired tasks. More importantly, multi-task learning reduces the number of training data samples required for each task,

which benefits generalization. This method can also be used to address the problem of overfitting.

In our proposed model, the model's primary task is to map emotions from features extracted from the input data, while the secondary task is to reconstruct the input data using the output of the bottleneck layer of the autoencoder network. Therefore, we define the multi-task cost function as follows:

$$Loss_{MTAE} = (1-\alpha) \times Loss_{PCAE} + (\alpha) \times Loss_{Classifier} \quad (11)$$

where $\alpha$ is a threshold parameter to control the importance of each cost function term. Here, MTAE stands for Multi-Task Autoencoder. Since the main task in this paper is emotion classification and the auxiliary task is autoencoder training, we consider a larger $\alpha$ so that the model's main focus is on classification accuracy. For the classification loss, we use cross-entropy loss, which is defined as follows:

$$Loss_{Classifier}(p,q) = -\sum_{i=1}^{N} p_i \, log q_i \quad (12)$$

where N specifies the number of classes, $q_i$ indicates the predicted probability distribution of category i, and $p_i$ is the actual probability distribution of category i. The sum of $p$ (or $q$ ) equals one.

## 4- Experiments

This section empirically examines the proposed PCAENet model for speech emotion recognition and demonstrates its efficiency. We conduct extensive experiments using the standard REVDESS dataset. The performance comparison of the proposed model with other advanced reported models is also reported. A complete description of speech emotion datasets and emotion recognition outputs along with discussion is provided in the following subsections.

### 4- 1- Dataset

The RAVDESS [38] dataset contains a total of 7356 audio-visual files (24.8 GB) of speech, song, and facial displays recorded at a sampling rate of 48000. It is a dynamic, multimodal set of emotional states and contents from North American English speakers. The database is gender balanced consisting of 24 professional actors, vocalizing two lexically matched statements in a neutral North American accent. It consists of 8 emotions calm, happy, sad, angry, fearful, surprised, disgusted, and neutral. Each expression is produced at two levels of emotional intensity (normal, and strong), with an additional neutral expression. All data is available in three modality formats: Audio-Video, Audio-only, and Video-only. In this research, only the audio data has been used. Table.1 provides a detailed breakdown of the emotions, audio files, and their respective percentage contributions.

### 4- 2- Speech Data Preprocessing and Feature Extraction

In this research, the RADESS dataset speech data was read at a sampling rate of 48000. For the experiment, we have used the 80% split method for training, 10% for validation and 10% for testing. Gaussian white noise is added to the RAVDESS dataset to improve model generalization and make it more noise-resistant. Adding noise creates new training samples that are more realistic and representative of real-world data, reducing overfitting and enhancing the model's ability to handle noisy inputs. In addition, it is an appropriate approach for data augmentation. This procedure and its effect on the data have been illustrated in Figure 5. We have used MFCCs as features to decrease the computational cost and also the redundant information. Here, we calculate 40 MFCCs for 282 time steps with a 512-length Hann window. This part was implemented using the librosa library [39].

### 4- 3- Model Preparation

The architecture of the proposed parallel autoencoder model in this study is shown in Figure 1. After each convolutional layer, a dropout layer with a dropout probability of 0.1 was placed to prevent overfitting. Also, the output of the bottleneck layer of the parallel autoencoder is given to a maximum pooling layer to generate a 128-dimensional feature vector. In the transformer model, first, the input is given to a maximum pooling layer and its output is fed to the transformer encoder. The transformer consists of 4 multi-head attention encoders and the dimension of its feedforward layer is 512. Also,

its activation function is a Rectified Linear Unit (Relu) with a 0.4 dropout layer. The outputs of the parallel encoders are concatenated with the mean output of the transformer and are given to a dropout layer with a dropout probability of 0.1. After that, it is given to the classification part which is a single-layer neural network to convert the created 296-dimensional vector to an 8-dimensional vector that is the number of labels (emotions). Next, it is given to a softmax layer to calculate

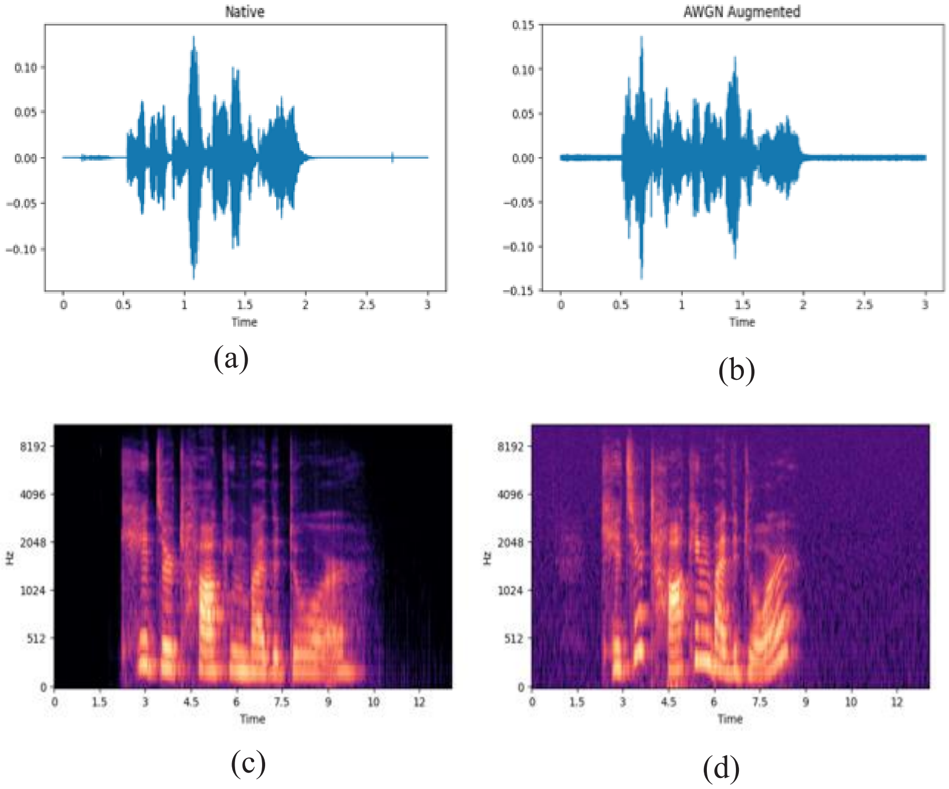


**Fig. 5. Data augmentation. a) Clean audio signal, b) Audio signal after adding white noise,c) clean audio signal spectrogram, d) noisy signal spectrogram**

**Table 1. A detailed breakdown of the emotions, audio files, and their respective percentage contributions from the RAVDESS dataset employed for the PCAENet model**

| Emotion | Audio files | contribution |
|---|---|---|
| **surprised** | 192 | **13.3%** |
| **angry** | 192 | **13.3%** |
| **calm** | 192 | **13.3%** |
| **happy** | 192 | **13.3%** |
| **sad** | 192 | **13.3%** |
| **neutral** | 96 | **6.67%** |
| **fearful** | 192 | **13.3%** |
| **disgust** | 192 | **13.3%** |

**Table 2. Impact of α on the proposed model's accuracy**

| $\alpha$ | Accuracy |
|---|---|
| **0.2** | **0.8198** |
| **0.3** | **0.8278** |
| **0.4** | **0.8527** |
| **0.5** | **0.8217** |

the probability. The model was trained with a batch size of 32 using the Adamw optimizer with a learning rate of 0.001. The model was trained for 600 epochs. The total trainable parameters are 360,682 for the PCAENet model.

The weight parameter α in (11) was empirically set to 0.4 to achieve the best results according to Table 2.

### 4- 4- Baseline models

For the comparison, we have selected the following state-of-the-art (SOTA) baseline models to evaluate the performance of the proposed PCAENet model. Att-Net [40] is a SOTA lightweight self-attention model for SER, where a CNN uses channel and spatial attention for the extraction of cues from the input tensors. The SVM ensemble model with a Gaussian kernel [41] is a standard benchmark that is utilized for SER comparison. The 1D-CNN [42] model is also used for comparison, which extracts MFCC features and uses the trained 1D-CNN for emotion identification. The other SOTA models are the CNN-BLSTM-based SER method from [43], A vector quantized masked autoencoder (VQ-MAE-S-12) [44], and CNN with a Convolutional Attention Block[45].

### 4- 5- Evaluation

The following evaluation measurements are used to investigate the emotion recognition performance of the proposed PCAENet model:

a) Accuracy is a metric that measures the proportion of correct predictions made by a machine learning model. It represents the percentage of correctly identified data points out of all the examples provided. The formulation is calculated as follows[46]:

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{13}$$

b) Precision is a metric that assesses the accuracy of positive predictions made by a machine learning model. It indicates the proportion of positive predictions that are actually correct. The formulation is calculated as follows[46]:

$$Precision = \frac{t_p}{t_p + f_p} \tag{14}$$

c) Recall, also known as true positive rate evaluates the completeness of positive predictions made by a machine learning model. It measures the proportion of actual positives that are correctly identified. The formulation is calculated as follows[46]:

$$Recall = \frac{t_p}{t_p + f_n} \tag{15}$$

d) The F1 score, also known as the F-measure, combines the precision and recall metrics to provide a single measure of the overall performance of a machine learning model. It aims to balance both precision and recall, making it suitable for situations where both are important. The formulation is calculated as follows[46]:

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

e) The confusion matrix is a table that summarizes the performance of a classification model by quantifying the true and false predictions it makes.

the true positives (TP) are instances where the model correctly identified as positive examples, while the true negatives (TN) are instances where the model correctly identified as negative examples. False positives (FP) occur when the model incorrectly identifies negative examples as positive, while false negatives (FN) occur when the model incorrectly identifies positive examples as negative.

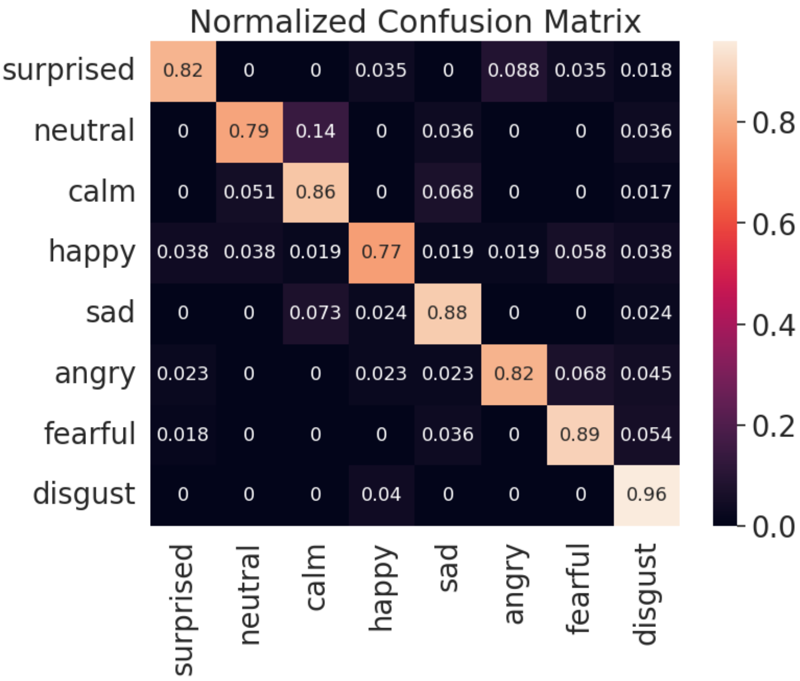


**Fig. 6. Confusion-matrixes of the proposed SER model using RAVDESS emotional speech data set with 85.27% average recognition rate among actual and predicted emotions.**

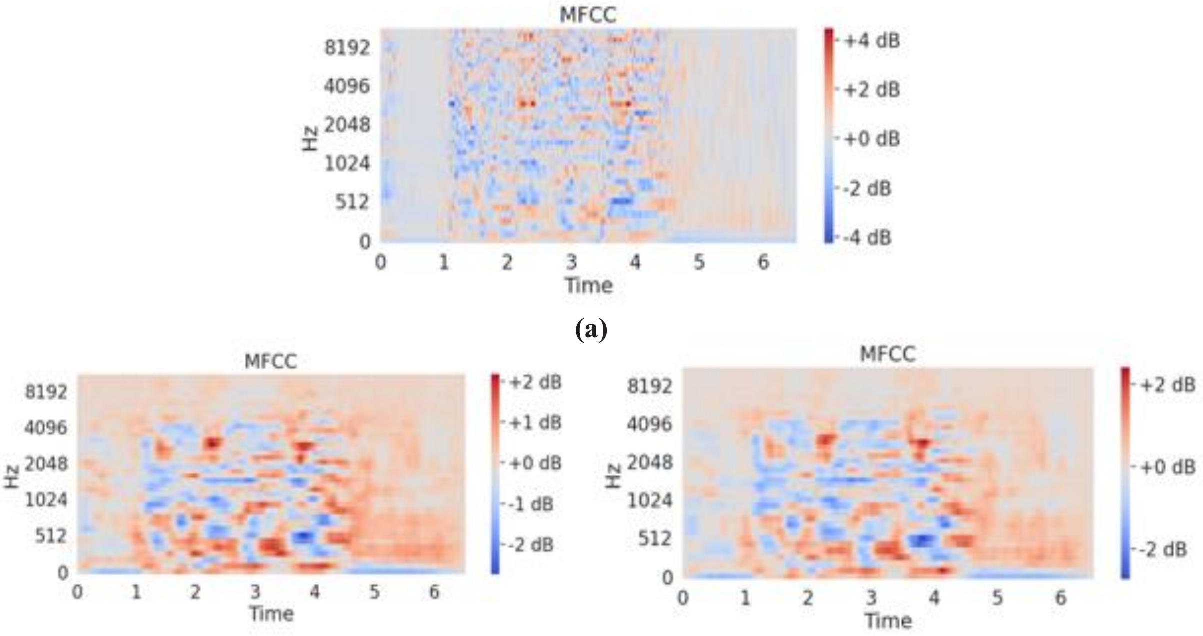
### 4- 6- Results

The recognition evaluations for each emotion class as well as the average values of the proposed PCAENet model for the RAVDESS dataset are shown in Table 3 The highest score belongs to the disgust emotion according to this table. The lowest result in terms of precision and F1-score refers to sad emotion because of its natural confusion by calm and disgust classes. These results correspond well to Fig. 6 showing the confusion matrix of the proposed model for 8 speech emotion classes.  As the confusion matrix shows in Fig. 6, the highest error is for the happy class, which is mistakenly taken as the fearful class the most. This happens due to the way it is expressed. After the fear class, the highest error is for the neutral class. Given that there are few neutral labeled data, and also the fact that calm and sad data are very similar to the neutral audio data, this result was expected. Then, the surprised and angry classes have the most errors, which have been mistakenly taken as anger and fearful classes due to their expression shape.Fig. 7 illustrates the correct mapping done by the autoencoder. To examine the performance of autoencoders, it should be noted that the main focus of the model is to improve classification accuracy. Also, due to the existence of the dropout layer in the encoder section, the suggested autoencoder acts as a denoiser too. The decoder parts of the autoencoders are shown in Fig. 7, and its output indicates that it has correctly learned a certain feature, and hence, can be used to reconstruct the input.

Table 4 compares the accuracy performance of the proposed PCAENet model with the baselines. According to this table, our proposed model has led to better results for SER. This has been achieved because of the following reasons:

- Suggesting parallel convolutional AEs for finding high-level and more informative features which also improve the generalization of the model. These PCAEs also take into account the ensemble learning technique to improve the random initializations in the training phase leading to higher accuracies,
- Proposing to concatenate and insert the bottleneck layers of the PCAEs including the nonlinear sparse features. These features have been proven to be effective for extracting the harmonic structure of speech signals and specifically useful for SER especially due to their informative structure almost without irrelevant features extracted in a nonlinear procedure,
- Proposing the multi-task loss function composed of both classification and regression terms. The regression term refers to accurately mapping the input MFCC features in the parallel autoencoders in an ensembled manner,
- Suggesting a transformer to take into account the long-term dependencies between speech samples parallel with the PCAEs. Their self-attention mechanism is effective here.

We have shown the loss curves of training and validation data for two CAEs and the whole proposed PCAENet model in Fig. 8. According to this figure, our proposed model



(a)

(b)

(c)

**Fig. 7. a) Parallel Autoencoders input, b) output of the decoder of the first CAE, c) output of the decoder of the second CAE**

**Table 3. The classification-report of the proposed SER system**

| | f1-score | recall | |
|---|---|---|---|
| surprised | 0.8704 | 0.8246 | 0.9216 |
| neutral | 0.8000 | 0.7857 | 0.8148 |
| calm | 0.8644 | 0.8644 | 0.8644 |
| happy | 0.8163 | 0.7692 | 0.8696 |
| sad | 0.7912 | 0.8780 | 0.7200 |
| angry | 0.8889 | 0.8182 | 0.9730 |
| fearful | 0.8772 | 0.8929 | 0.8621 |
| disgust | 0.8807 | 0.9600 | 0.8136 |
| Un-weighted Accuracy | 0.8486 | 0.8491 | 0.8549 |
| weighted Accuracy | 0.8532 | 0.8527 | 0.8601 |
| **Accuracy** | | **0.8527** | |

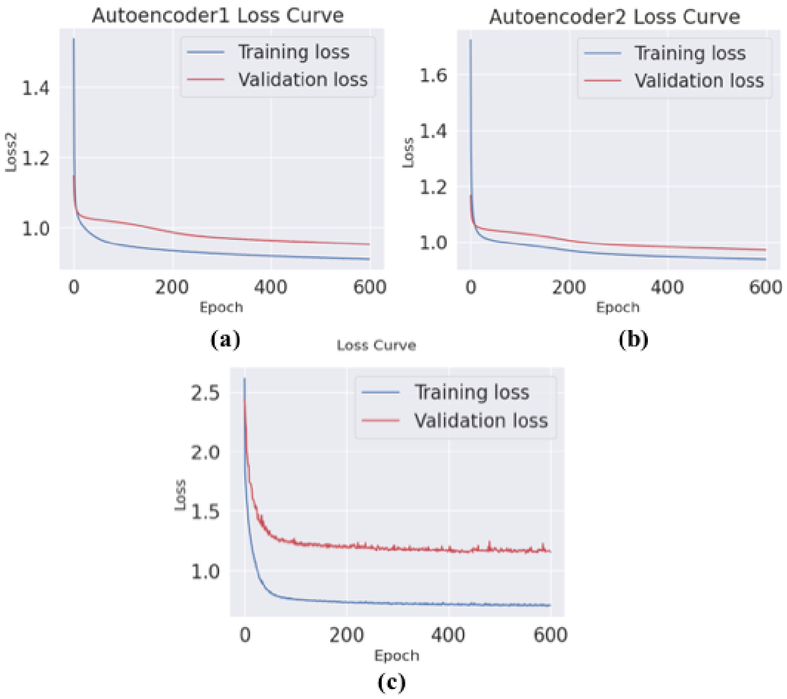


**Fig. 8. Loss curves. a) First autoencoder loss, curve b) second autoencoder loss curve, c) The proposed model loss curve**

**Table 4. Performance comparison between the proposed**

| models | Accuracy |
|---|---|
| **Att-Net[40]** | 80 |
| **BE-SVM[41]** | 75.69 |
| **1D-CNN[42]** | 71.61 |
| **Deep-BLSTM[43]** | 77.02 |
| **CNN-CBAM[45]** | 82.38 |
| **VQ-MAE-S-12[44]** | 84.1 |
| **PCAENet** | **85.27** |

behaves is learned well and no overfitting problem has been seen.

To assess the model's noise robustness, we also tested the network on noisy test data, achieving an accuracy of 83.29%. This small difference of less than 2% compared to clean test data (85.27%) proves the model's robustness to noise. The reason for this noise robustness is that the proposed parallel autoencoders have denoising properties, and the training data includes both noisy and clean data.

The proposed PCAENet model demonstrated improved generalization during the experiments and evaluations for the RAVDESS dataset, and it obtained better emotion recognition accuracy.

## 5- Conclusion

In this paper, we proposed a novel model for SER including new multi-task parallel convolutional autoencoder (PCAE) and transformer networks. We proposed the PCAEs to generate informative harmonic sparse and high-level features from the input in a non-linear manner. In other words, we extracted nonlinear sparse features from the bottleneck layer of each convolutional AE having much fewer dimensions compared to the input. These lower-dimension features extracted from the MFCC input in two parallel networks in an ensemble manner can be considered as informative nonlinear sparse features that are almost free of irrelevant information. Thus, they can be helpful for the final classification purpose. The proposed PCAEs designed taking into account the ensemble learning procedure could improve the accuracy and the generalization of the model because they solve the problem of initial sequential information loss during convolution operations. In order to acquire long-term dependencies between speech samples, we also proposed using a transformer in parallel with PCAEs making use of its self-attention mechanism. Finally, we proposed a multi-task loss function composed of two terms of classifications and regression which works as AE mapper. This multi-task loss makes a new model that not only reduces the classification error but also decreases the regression error caused by the PCAEs which are also considered as mappers between the input and output MFCCs. Therefore, the proposed model is capable of finding appropriate features with PCAEs leading to improved classification results. We used RAVDESS SER dataset including eight emotions in this paper to evaluate our proposed model. Our model outperformed all previous SER methods in terms of average accuracy.

To further enhance the performance of the proposed SER model, the use of autoencoders with more sophisticated architectures such as variational and adversarial autoencoders could be explored.

## References

[1] C. Zhu, W. Ahmad, Emotion Recognition from Speech to Improve Human-Robot Interaction, in: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2019, pp. 370-375.

[2] P. Tarnowski, M. Kołodziej, A. Majkowski, R.J. Rak, Emotion recognition using facial expressions, Procedia Computer Science, 108 (2017) 1175-1184.

[3] C. Yu, M. Wang, Survey of emotion recognition methods using EEG information, Cognitive Robotics, 2 (2022) 132-146.

[4] Q. Wang, M. Wang, Y. Yang, X. Zhang, Multi-modal emotion recognition using EEG and speech signals, Computers in Biology and Medicine, 149 (2022) 105907.

[5] N. Azam, T. Ahmad, N. Ul Haq, Automatic emotion recognition in healthcare data using supervised machine learning, PeerJ Comput Sci, 7 (2021) e751.

[6] G. Liu, S. Cai, C. Wang, Speech emotion recognition based on emotion perception, EURASIP Journal on Audio, Speech, and Music Processing, 2023(1) (2023) 22.

[7] J. de Lope, M. Graña, An ongoing review of speech emotion recognition, Neurocomputing, 528 (2023) 1-11.

[8] X. Li, R. Lin, Speech Emotion Recognition for Power Customer Service, in: 2021 7th International Conference on Computer and Communications (ICCC), 2021, pp. 514-518.

[9] M. Bojanić, V. Delić, A. Karpov, Call Redistribution for a Call Center Based on Speech Emotion Recognition, Applied Sciences, 10(13) (2020) 4653.

[10] E. Andre, M. Rehm, W. Minker, D. Bühler, Endowing Spoken Language Dialogue Systems with Emotional Intelligence, 2004.

[11] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, R.W. Picard, Driver Emotion Recognition for Intelligent Vehicles: A Survey, ACM Comput. Surv., 53(3) (2020) Article 64.

[12] B. Schuller, Towards intuitive speech interaction by the integration of emotional aspects, in: IEEE International Conference on Systems, Man, and Cybernetics, 2002, pp. 6 pp. vol.6.

[13] P. Jamshidlou, N. Keshtiari, M. Eslami, M. Bahrani, Acoustic Representation of Intonational Elements in Persian Emotional Speech, 2013.

[14] F. Daneshfar, S. Kabudian, Speech Emotion Recognition Using Deep Sparse Auto-Encoder Extreme Learning Machine with a New Weighting Scheme and Spectral/Spectro-Temporal Features Along with Classical Feature Selection and A New Quantum-Inspired Dimension Reduction Method, (2021).

[15] G. Drakopoulos, G. Pikramenos, E. Spyrou, S. Perantonis, Emotion Recognition From Speech: A Survey, 2019.

[16] A. Satt, S. Rozenberg, R. Hoory, Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms, 2017.

[17] C.H. Wu, W.B. Liang, Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels, IEEE Transactions on Affective Computing, 2(1) (2011) 10-21.

[18] S. Majuran, A. Ramanan, A feature-driven hierarchical classification approach to emotions in speeches using SVMs, in: 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), 2017, pp. 1-5.

[19] M. Lashkari, S. Seyedin, NMF-based Cepstral Features for Speech Emotion Recognition, 2018.

[20] A. Guerrieri, E. Braccili, F. Sgrò, G.N. Meldolesi, Gender Identification in a Two-Level Hierarchical Speech Emotion Recognition System for an Italian Social Robot, Sensors, 22(5) (2022) 1714.

[21] W.Q. Zheng, J.S. Yu, Y.X. Zou, An experimental study of speech emotion recognition based on deep convolutional neural networks, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 827-831.

[22] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, Transfer learning for improving speech emotion classification accuracy, arXiv preprint arXiv:1801.06353, (2018).

[23] B. Xie, M. Sidulova, C.H. Park, Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion, Sensors, 21(14) (2021) 4913.

[24] M. Xu, F. Zhang, W. Zhang, Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset, IEEE Access, 9 (2021) 74539-74549.

[25] M.T. García-Ordás, H. Alaiz-Moretón, J.A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, C. Benavides, Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network, Biomedical Signal Processing and Control, 69 (2021) 102946.

[26] J. Deng, Z. Zhang, E. Marchi, B. Schuller, Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 511-516.

[27] H. Aouani, Y.B. Ayed, Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder, in: 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2018, pp. 1-5.

[28] W. Jiang, Z. Wang, J.S. Jin, X. Han, C. Li, Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network, Sensors, 19(12) (2019) 2730.

[29] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.

[30] D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Systems with Applications, 173 (2021) 114683.

[31] M. Chen, X. He, J. Yang, H. Zhang, 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition, IEEE Signal Processing Letters, 25(10) (2018) 1440-1444.

[32] M. Xu, F. Zhang, S.U. Khan, Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion, in: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020, pp. 1058-1064.

[33] L. Rabiner, R. Schafer, Theory and Applications of Digital Speech Processing, Pearson Education, 2011.

[34] E. Pintelas, I. Livieris, N. Barotsis, G. Panayiotakis, P. Pintelas, An autoencoder convolutional neural network framework for Sarcopenia detection based on multi-frame ultrasound image slices, 2021.

[35] C.S. Wickramasinghe, D.L. Marino, M. Manic, ResNet Autoencoders for Unsupervised Feature Learning From High-Dimensional Data: Deep Models Resistant to Performance Degradation, IEEE Access, 9 (2021) 40511-40520.

[36] Y. Sun, H. Mao, Q. Guo, Z. Yi, Learning a good representation with unsymmetrical auto-encoder, Neural Comput. Appl., 27(5) (2016) 1361–1367.

[37] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, Q. Du, Unsupervised Spatial–Spectral Feature Learning by 3D Convolutional Autoencoder for Hyperspectral Classification, IEEE Transactions on Geoscience and Remote Sensing, 57(9) (2019) 6808-6820.

[38] S.R. Livingstone, F.A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PLoS One, 13(5) (2018) e0196391.

[39] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and Music Signal Analysis in Python, in: SciPy, 2015.

[40] Mustaqeem, S. Kwon, Att-Net: Enhanced emotion recognition system using lightweight self-attention module, Applied Soft Computing, 102 (2021) 107101.

[41] A. Bhavan, P. Chauhan, Hitkul, R.R. Shah, Bagged support vector machines for emotion recognition from speech, Knowledge-Based Systems, 184 (2019) 104886.

[42] D. Issa, M. Fatih Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control, 59 (2020) 101894.

[43] Mustaqeem, M. Sajjad, S. Kwon, Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, IEEE Access, 8 (2020) 79861-79875.

[44] S. Sadok, S. Leglaive, R. Séguier, A Vector Quantized Masked Autoencoder for Speech Emotion Recognition, in: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2023, pp. 1-5.

[45] F. A. Dal Ri, F.C. Ciardi, N. Conci, Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks, IEEE Access, 11 (2023) 116638-116649.

[46] D. Powers, Evaluation: From Precision, Recall, and F-Factor to ROC, Informedness, Markedness & Correlation, Mach. Learn. Technol., 2 (2008).

**HOW TO CITE THIS ARTICLE**

*Z. Sharifzadeh Jafari, S. Seyedin. A Novel Multi-Task and Ensembled Optimized Parallel Convolutional Autoencoder and Transformer for Speech Emotion Recognition. AUT J. Elec. Eng., 56(2) (2024) 213-226.*

**DOI:** 10.22060/eej.2024.22903.5570