International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Automatic Speech Emotion Recognition Using Hybrid Deep Learning Techniques

Bilal Hikmat Rasheed[1*], D. Yuvaraj[2], Saif Saad Alnuaimi[3], S. Shanmuga Priya[4]

Submitted: 13/12/2023 **Revised:** 25/01/2024 **Accepted:** 04/02/2024

Abstract: An emerging field of research is the advancement of deep learning techniques for speech emotion recognition. The current scenario of human-computer interaction is being significantly impacted by and altered by speech recognition technologies. In human-computer interaction, developing an interface that can sense and react accurately like a human is one of the main crucial challenges. As a result, the Automatic Speech Emotion Recognition (ASER) system has been developed. It extracts and identifies important data from voice signals to classify various emotional categories. The novel advancements in deep learning have also led to a major improvement in the ASER system's performance. Numerous methods, including some well-known speech analysis and classification approaches, have been used to derive emotions from signals in the literature on ASER. Recently, deep learning methods have been suggested as an alternative to conventional methods in ASER. The main goal of this research is to use deep learning techniques to analyze different emotions from speech. Because deep learning networks have sophisticated feature extraction processes, they are frequently utilized for emotional classification, in advance of traditional/machine learning systems that depend on manual feature extraction before classifying the emotional state. To extract features and identify different emotions depending on input data, the authors have implemented the most efficient hybrid deep learning algorithms, CNN+LSTM. By training and testing the suggested network algorithm with the standard dataset, the authors, accordingly, achieved the highest accuracy.

Keywords: *Automatic Speech Emotion Recognition, Deep Learning; Human-Computer Interaction, Convolutional Neural Network, Long Short Term Memory*

1. Introduction

Speech has a significant impact on recognizing people's emotional states when it comes to emotion recognition. With the development of voice-based HCI and the introduction of smart devices such as the Amazon Echo and Apple HomePod, speech data has become a significant component of big data [1]. The advancement of technology has led to a rise in human-computer connection, and verbal communication between humans and machines represents an approach to enhance and simplify this interaction [2]. Because of this, authors have been researching various ways to improve spoken communication efficiency in systems like speaker and speech recognition for the past few decades. Creating a technology that can recognize and communicate with humans constitutes a few of the main goals of voice emotion recognition systems. This has given development to the significant and difficult subject of Automatic Speech Emotion Recognition (ASER) in recent times. In everyday human-machine interactions, speech-emotion recognition proves to be an advantageous feature that improves the connection, accuracy, and speed of communication. SER has been used in numerous fields recently; including web call centers, computer gaming, smartphones, automobile engineering, and medical crises. The speech emotion recognition field experiences numerous obstacles despite the wide range of techniques available. For example, in speech emotion recognition tasks, the collection and decision on useful elements usually influence classification accuracy.

Deep learning methods are currently used to address recognition challenges in speech emotion recognition, image recognition, voice recognition, and face recognition [3]. The ability to automatically choose features is one of the primary benefits of deep learning approaches [4]. This capability could be used, for instance, to identify essential qualities intrinsic to sound files containing a specific emotion in the process of speech emotion recognition [5]. Several deep neural network-based approaches for speech emotion identification have been released in recent years. One set of these algorithms develops the neural network to directly identify important features from unprocessed audio samples [6], whereas the other collection of algorithms only requires an individual sound file format as input.

Using mean values along the time axis, previously constructed a one-dimensional array by stacking the five distinct features that we extracted from a sound sample. Afterwards, the "Convolutional Neural Network (CNN)" model receives this array as input. The Proposed System shows that combining these elements in the input data yields a more varied representation of a sound file, improving both the classification and generalization of speech recognition systems for emotion identification. Furthermore, to increase the classification accuracy of our baseline approach, researchers implement an incremental approach [7, 8]. The suggested feature combination, which uses five distinct spectral representations of the same sound file, has not yet been explored by researchers, even though multiple speech-emotion detection approaches in the literature integrate numerous feature types. More features are included in the

[1]*Department of Computer Science, Cihan University-Duhok, Iraq.*
Email: bila.rasheed@duhokcihan.edu.krd
**(Corresponding Author)*
[2]*Department of Computer Science, Cihan University-Duhok, Iraq.*
Email: d.yuvaraj@duhokcihan.edu.krd
[3]*Department of Computer Science, Cihan University-Duhok, Iraq.*
Email: saif.saad@duhokcihan.edu.krd
[4]*Department of Computer Science Engineering,*
SRM Institute of Science and Technology, Trichy, India
Email: priya501@gmail.com

combination of features that produce strong audio fluctuation identification and tracking but weak pitch class and harmony distinct representations, hence enhancing the representational capacity of the composition.

Thus, the highest-performing system becomes the new state-of-the-art by outperforming all previous systems that use audio characteristics and provide their classification accuracies on the same emotion classes for both “RAVDESS [9] and the IEMOCAP” datasets [10]. Our highest performance system compares to all prior efforts for the EMO-DB dataset [11], except the Zhao et al. study [12]. On the other hand, our model compares well with that one in the categories of simplicity, consistency, and application. Furthermore, to address them in Section 4.4, we have observed a few discrepancies concerning [12]. Given the effectiveness of these DL-based models in the situations in which they are used, there continue to be several major obstacles that need to be addressed. These issues are as follows:

- The primary obstacle in identifying emotions from speech is the variation in speech length. The conventional approach involves reducing and padding the data to provide the fixed-shape features that are needed as the neural network's input. This method results in information loss and increases computation costs.
- Convolutional neural networks (CNNs) can find local features that are not globally contextually relevant, and specific emotional content could be hidden in long speeches.
- To independently record both local and global emotional information in a long speech, an innovative architecture must be developed.

The automatic speech emotion identification framework used in this article combines CNN in conjunction with Bi-LSTM networks, emphasizing the relation-based method of attention weight calculation. In this case, short-term temporal information was modeled using CNN using a kernel window for each response. In the end, the long-term temporal emergence of the emotional characteristic for each speech response was characterized using attention-based Bi-LSTM. The following section of the article examines the relevant work. Section 3 discusses the main processes along with the suggested solution for ASER. The Multimedia Devices and Applications evaluation findings and implementation are discussed in detail, along with pertinent discussions, in the section that follows. This section also compares the suggested solution's accuracy and processing time to the state-of-the-art solution. After providing a brief synopsis of the study, the conclusion includes a briefing on the next work.

2. Related Works

The process of automatically identifying a speaker from a recording of their voice or speech utterance is known as speaker recognition. Using four classifiers “Support vector machines (SVM), K-Nearest Neighbours, Multilayer Perceptron (MLP), and Random Forest (RF)” that are trained by the WEKA data mining application they have shown automatic speaker detection of Sepedi home language speakers in [6]. Through the use of 10-fold cross-validation, each algorithm's efficacy was determined. Two classifiers, RF with 99.9% accuracy and MLP with 97% accuracy scored the highest precision. Ultimately, a GUI was used to execute the RF framework for development testing.

Spoken language comprehension and automatic speech recognition (ASR) are two of the most crucial aspects of modern machine intelligence. The focus of [7] was on isolated voice command recognition for intelligent robotics and automatic HCI systems. They developed a grammar framework for small-scale command set testing with self-loops for every condition, returning blank representations for noise and vocabulary-out phrases. Experimental data shows that this method succeeded with 60% higher accuracy than traditional offline language model-based speech recognition approaches. They compared the privilege of the technique's recognition accuracy and average decision-making time with the most developed continuous speech recognition engines.

The efficiency of automated speech recognition (ASR) has significantly increased with the development of DL [7]. They've described a technique in [8] for improving an early pre-trained model and script for automatic voice recognition. They created a pair of audio files and their corresponding scripts by comparing portions from the pre-trained model with the ground truth data. Because they employed human-written scripts, each pair of audio files has beginning and complete perspectives for each utterance. Through precise and efficient technique extraction, they were able to obtain an autonomous speech recognition dataset.

Of all the natural means of communication among people, speech is the most significant. Since the majority of people have been employed in the voice recognition and communication fields for a longer period, [9] has analyzed methods and strategies related to speech recognition. In the realm of speaker-dependent and speaker-independent voice recognition, a significant number of researchers have also made contributions, according to [9]. After applying the many methods for creating an ASR framework, each with pros and cons, it became apparent that the two main things influencing the effectiveness of the system were the methods for extracting features from the language and the speech recognition strategy.

An effective source filter modeling framework-based neural vocoder test-to-speech (TTS) system has been proposed in [10]. Although SampleRNN and WaveNet are well-known for producing high-quality speech, their generation speed was too weak for practical usage. Instead, they employed two neural vocoder methods. To train the excitation component, they used a SampleRNN-based generative model, and for the spectrum or acoustic characteristics, they used a long short-term memory network. At last, the findings validate the superiority of the suggested method over the initial SampleRNN-based speech synthesis approach and a parametric glottal modeling-based method.

One of the fundamental methods for transferring voice commands to text in machine-to-machine and machine-to-machine (M2M) communications is automatic speech recognition (ASR) [1]. The speech aid of Apple Siri, the home management aid of Amazon Alexa/Echo, the intelligent search and service assistant of Google Home, and the personal assistant of Microsoft Cortana are just a few examples of the many applications of ASR methods in recent times in information retrieval and speech-to-text services. The effectiveness and intelligence of ASR's interface are crucial to the eligibility of those applications, where it performs an important function.

ASR is essentially a method of mapping a spoken audio sequence to a word sequence. Due to its capacity for extrapolating significant features in a way that is comparable to the human ear, the “Mel-frequency Cepstrum Coefficient (MFCC)” is the most often utilized feature for ASR [1]. After that, the system obtains the feature vector and uses it for either training or inference to determine whether the text is recognized. The issue formulation between the statistical techniques is to maximize the posterior probability of a word sequence while maintaining an eye on an audio series. Hybrid algorithms, such as the “Gaussian mixture approach with hidden Markov framework (GMM-HMM) or the deep neural network with hidden Markov designs (DNN-HMM),” are examples of conventional frameworks. The acoustic and linguistic systems are the two separately optimized factors that make together any of these hybrid approaches.

According to many studies [11-13], modern ASR systems correspond to an end-to-end structure that merges the two features into a single trainable network. These frameworks may be applied when considering the output words or characters as identifiers. It does, however, depend on the attention framework or “bidirectional long short-term memory (BLSTM) recurrent neural network”, which traverses an entire phrase from both sides to determine frame-wise outputs. The framework becomes unfeasible for generating adversarial examples and causes significant temporal lag. These frameworks also cannot specify the locations of hostile cases.

In this type, Attention-Based Convolution is implemented. A deep-learning audio framework called a SCBAMM, or skip convolution dependent on attention long short-term memory which may be utilized in both directions, is used to identify speech emotion. Among the thick layers are a Bi-LSTM layer, a convolutional layer, a pooling layer, a skip layer, a mask layer, and an attention layer. With SCBAMM, spatiotemporal data may be used more effectively, and elements related to emotion are captured more effectively. Moreover, it examines some of the deep learning issues of gradient bursting and gradient fading [14]. The machine learning framework used in this technique is “Sparse Kernel Reduced-Rank Regression (SKRRR) for Speech Emotion Recognition”. GRIN Several methods were used in this assessment to extract speech emotion features, including functional extraction feature descriptors transformed features and scale symmetry [14]. (SKRRR) was used to categorize and forecast the extracted emotions to get the optimal solution for the coefficient matrices with a high recognition rate. For training, it generates the five classes using two different classifier datasets.

According to [15], mel-spectrograms containing deltas and delta-deltas are utilized as input for 3-D attention-based R-CNN, which are designed to develop distinguishing characteristics for SER. [16] develop a 3-D CRNNs-based temporal modulation cue-based end-to-end SER solution following this methodology. Subsequently, they suggest using a BLSTM and dilated CNN with a residual block that depends on the attention strategy as a basis for SER [17]. To address the drawbacks of using alone, it may exploit the benefits of various frameworks. By utilizing attention-based BLSTM with fully convolutional networks, [18] offers an approach to the challenge of emotionally significant feature extraction from speech. This enables the system to learn on its own the optimal spatio-temporal representations of speech signals. To forecast the emotional expression of a spoken utterance, the learned high-level features are subsequently fed into a DNN. The attention-BLSTM-FCN architecture was introduced by [19] in the following study. This involves integrating the spectrograms into two separate neural networks in parallel: an attention-based BLSTM network for temporal feature extraction and an attention-based FCN network for spatial feature extraction. These network outputs are combined to create a composite spatial-temporal feature vector. These research findings have shown how important it is to use discriminative spatiotemporal variables to explain how various emotions change over time. [20] Presented a system that uses low-level descriptors (LLDs) as input, constructs utterance-level representation with LSTM, uses Dilated Residual Network (DRN) with multi-head self-attention for feature learning, and uses a DNN-based classifier for determining emotion.

3. Proposed System

Our analysis of several standard techniques brings us to the conclusion that speech emotion recognition could be achieved with CNN Bi-LSTM networks and a few more layers. Long-term contextual interconnections are the primary goal of the Bi-LSTM layer. According to [21], the Bi-LSTM network collects local as well as global contextual information, illustrating how a signal's frequency content changes over time, whereas the CNN network may extract emotional elements from raw audio sounds. The primary purpose of the “Local Feature Learning Block (LFLB)” in this case is to gather local features, which are subsequently paired with LSTM to acquire global and local features from raw audio samples, accordingly. Even though this algorithm's recognition accuracy is better than that of previous approaches, it has a black box issue where it is easy to manipulate the algorithm with particular speech inputs while the gathered characteristics in the SER cannot be predicted. While batch normalization has been included in the CNN LSTM network, this framework fails to address the overfitting problem.

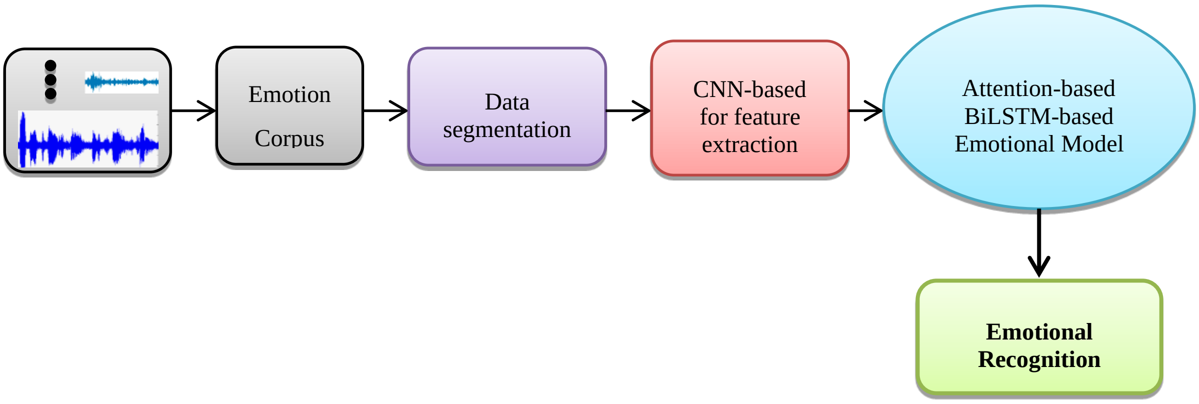
Emotion
Corpus
Data segmentation
CNN-based for feature extraction
Attention-based BiLSTM-based Emotional Model
Emotional Recognition

Fig. 3.1. Proposed Framework for Automatic Speech Emotion Recognition

Proposed an attention-based system that uses multimedia devices and applications to measure each voice response, produce an accurate emotional representation of the response, and focus on relevant responses to enhance performance. With this method, voice replies found in the databases frequently exhibit many emotional characteristics; listeners' evaluations result in the identification of multiple emotional classes per instance. The author of [22] suggested an over-sampling strategy that takes into account the traditional perceptual evaluation procedure to solve this issue. The proposed framework for Automatic Speech Emotion Recognition is shown in Figure 3.1.

A. Pre-Processing

Silence segments are identified during the pre-processing phase using the Praat Silence Detection analyzer. The silent threshold (dB), min-quiet interval, and min-sounding interval are the three primary features of this method. This program detects the sound and silent intervals and evaluates the intensity contour accordingly. The sound intervals with durations lesser than the minimum sound interval duration are then eliminated. Using an IIR band pass filter with a lower cut-off frequency of 8 KHz, the disturbances that is, sounds captured in surroundings that are sensitive to high-frequency noise and DC offsets at lower frequencies are eliminated. By allowing for a 10-millisecond overlap between frames, hamming windowing is also used in this instance to deal with the non-stationary speech challenge and prevent information loss.

After that, the audio signal is transformed into a dimensional model to use CNNs. Each audio signal is subsequently illustrated by a dimension tensor of size n x H x V, where n is the number of consecutive spectrograms that define the keyframes, H is the spectrogram's horizontal dimensions, and V is its vertical dimensions.

B. Feature Extraction and Classification

For emotion research, prosodic, cepstral, and spectral features are typically used. It primarily uses prosodic and spectral characteristics in this paper. These acoustic features, which can be distinguished between low-level and high-level features in the ASER, are utilized to classify emotions. In this case, the CNN which has convolutional layers, a single batch normalization (BN) layer, a pooling layer, and a fully connected layer is made for feature extraction, whereas the Bi-LSTM network is specifically made for long-term sequence dependencies. In this instance, features are extracted from temporal and spatial dimensions using CNN.

C. Deep Features Extracted by Convolutional Neural Network (CNN)

CNN has demonstrated amazing capabilities in the past few years when it comes to the extraction of deep features from log mel-spectrograms [23] [24]. Overcoming every kind of speech signal which includes a range of speakers, diverse environments, and so forth is essential to improving speech emotion recognition performance. CNN offers an interpretation that is independent of geographical studies and time. Translation invariance refers to the ability of the CNN to determine the class to which the input corresponds even after translation. The pooling process yields the translational invariance. The variability of speech signals may be addressed by using the convolution's invariance when the concept of a convolution neural network is implemented in the acoustic modeling of SER. Thus consider the mel-spectrogram that is obtained from the speech signal as an image, and then utilize a deep convolution network that is frequently applied in images to extract features.

The convolution layer, pooling layer, and fully connected layer constitute the computational framework of a CNN in the majority of instances. Fully connected layers provide the retrieved features to the final output for classification after the features have been determined by the convolution and pooling layers. Convolution layer filters are shifted from left to right throughout the input image according to their height and width. Equation (i) provides an equation for convolution.

$$C\,(a, b) = (F*d)\,(a, b) = \sum_m \sum_n F(a-m, b-n)d(m, n) \qquad (i)$$

The convolution component of (m, n) size is denoted by d, and C represents the characteristic mapping provided in the equation.

CNN also includes the pooling layer that is introduced following the convolution method. Its primary goal is to reduce the convolutional network's computation load and variable load. The most used pooling strategy is maximum pooling. In general, the output features mapping via the pooling or final convolution layer are flattened and transformed into a number sequence or dimensional vector. The completely linked layers then get these features as input. Here, each input and output is connected, allowing it possible to determine each neural node's weight. The fully connected layer completes the classification; typically, its output node count corresponds to its class number. The attention-BLSTM network receives deep features; hence the only layers that are utilised to extract deep features from log mel-spectrograms are the convolution and pooling layers.

D. Attention-based Bi-LSTM Approach for Automatic Speech Emotion Detection

The purpose of long-term dependency problem-solving and long-term information archives in recurrent neural networks (RNNs) is achieved by LSTM networks. In the proposed analysis, the evoked response's the Speech Emotion Analysis (SEs) path is modeled using LSTM during the interview process. Here, the introduction of attention mechanisms served to highlight the crucial speech responses and remove some unnecessary information. With h_a and l_a representing the attention weight and ath hidden outputs of the LSTM network, correspondingly, the concatenated SEs are implemented to train the LSTM-based emotion detection system. From the output of the fully connected layer, the emotion value is derived.

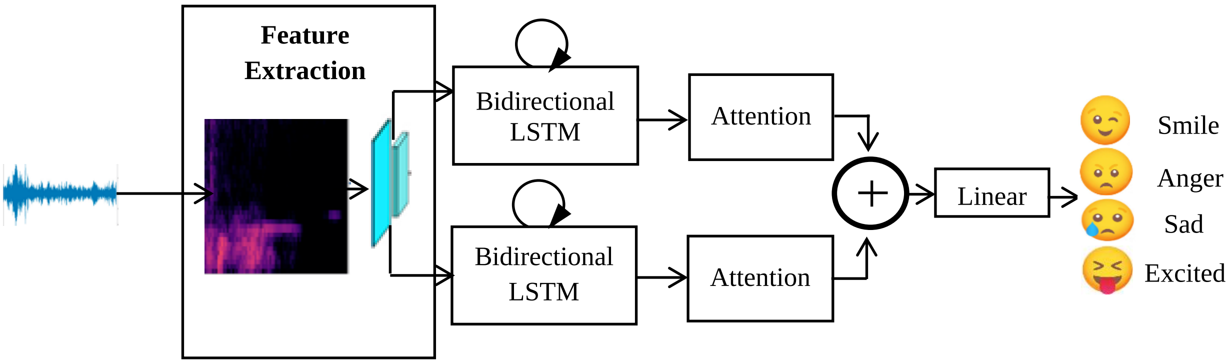

Fig. 3.2. Attention-based Bi-LSTM Approach for Automatic Speech Emotion Detection

Using the MLP-based classifier as the basis, an LSTM-based system that includes a technique for attention was created for recognizing mood disorders. The suggested approach outperformed CNN and SVM-based techniques in terms of detection accuracy. In this investigation, the evaluation was performed via the RAVDESS and IEM databases. According to the proposed approach, the attention operation yields a unique vector that, while providing varying attention weights, reflects every hidden state. Attention-based Bi-LSTM Approach for Automatic Speech Emotion Detection is shown in Figure 3.2. Consider the attention weights to be the relevance scores of the hidden states of the input encoder when executing the queries.

$$SE_{a,b} = \partial_{ab} \times N_{a,b} \otimes \mathrm{D} \qquad \text{(ii)}$$

The attention mechanism-based EP was computed by the attention weight sequence, $\partial_{a,1}, \partial_{a,2}, \ldots, \partial_{a,5}$. The SoftMax activation function is used to determine and normalize this attention weight.

$$\partial_{a,b} = \frac{X(x_{a,b})}{\sum_{b=1}^{zh} X(x_{a,b})} \qquad \text{(iii)}$$

To estimate attention weights, the similarity between the ith video-based response vector M_a and the question-based vector $M_{a,b}$was calculated using cosine similarity in the following way:

$$x_{a,b} = \frac{M_a . M_{a,b}}{\|M_a\| \, \|M_{a,b}\|} \qquad \text{(iv)}$$

The proposed solution's basic drawback is that it is dependent on a tiny database. 45 participants 15 with bipolar disorder (BPD), 15 with unipolar depression (UPD), and 15 healthy controls were employed in Huang et al.'s study [5] to analyse speech responses. The attention system in CNN and BiLSTM models yields better accuracy but at the cost of computational complexity and the inability to parallelize CNN sequences. While the attention approach may effectively reduce wide-scale sequence retrieval searches, even with massive and higher-dimensional sequences, it ignores the computation of temporal ordering, which is an essential consideration.

Managing these issues is necessary to achieve higher performance accuracy in emotion detection, as the currently proposed approach is unable to contain relative and absolute position information in its attention analysis of CNN and BiLSTM frameworks. Firstly, it can remove noise for the initial step of collecting an emotion corpus before obtaining an emotion identity that contains irrelevant information. Next, the relation-aware self-attention emulation might replace the attention-based models without the use of position principles. Finally, the recently proposed mechanism enables parallelization of the computation that follows the elimination of complexities in large sequences.

E. Proposed CNN-BiLSTM System

The proposed system introduces the SIEOS approach as a means of oversampling. Using a repeated sample from the targeted emotional class, this offered an equation for KNN, ‘a’ phrases. The interpolation of the feature vectors of the repeated sample and the chosen neighbor sentence yields the recommended feature vector of the resulting analysis,

$$C_{SIEOS} = \partial * C_{rep} + (1 - \partial) * C_{NN} \qquad \text{(v)}$$

Here, let's compute the section that separates the sentences and run the example again in feature space. Next, by choosing a random point along the chosen segment, the final location of the repeated sample is computed. The method by which the synthetic sample is interpolated is determined by the random selection ∂, which is set between 0 and 1, in the equation above. It chooses the settings at random as a maximum of two artificial samples with identical feelings would be generated through this technique.

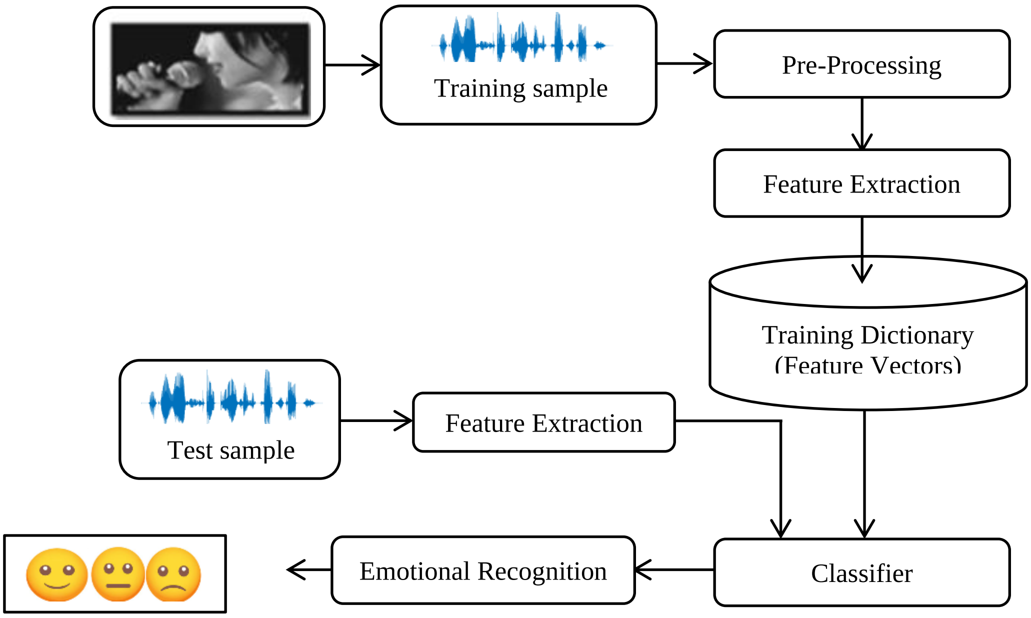

Fig. 3.3. The proposed Block Diagram for Automatic Speech Emotion Detection

Now that an attention system has been established in position, the attention map indicated by the following equation is computed using the attention weight vector v ∈ Q E.

$$B = d^F A \qquad \text{(vi)}$$

Where, F represents the number of frames and $\acute{F}$ indicates the number of features per frame, A $[a_1, a_2, a_{3,\dots,} a_z]$ is an input matrix of shape F × $\acute{F}$.

When performing weighted pooling with the LSTM output, the computed map could be normalized to unity and utilized as a weighting factor. The following equation may be applied to do it:

$$W = A\ \partial^F \qquad \text{(vii)}$$

Where, ∂ represents the SoftMax activation function.

The speech sequence's attention weight is initially determined, and it is then normalized using the SoftMax activation function, which is provided as follows:

$$\partial_{ab} = \frac{X(x_{ab})}{\sum_{y=1}^{Fn} X(x_{ay})} \qquad \text{(viii)}$$

Where, $\partial_{a,\ b}$ represented the weight assigned to the bth question-based response during the ath video, Fn is the number of questions, and $x_{a,1}, x_{a,2}, \dots, x_{a,5}$ is the series of similarities.

$$x_{a,b} = \frac{M_a . M_{a,b}}{\|M_a\|\ \|M_{a,b}\|} \qquad \text{(ix)}$$

Where x_{ab} alignment method assigns a score based on the rate at which the output at position a and the inputs surrounding location b match

The significance of annotation is shown by the probabilities ∂_{ab} and its corresponding energy, x_{ab}. This method will be applied throughout the annotation sequence as it is inherited from the decoder's attention function.

Proposed a modified variant of the relation-aware self-attention framework known as [25], it adds pairwise relationships between input items. The proposed Block Diagram for Automatic Speech Emotion Detection is shown in Figure 3.3. Here, the input sequences are described as fully connected, stated, and labelled graphs, which are then utilized to extract features. For the speech emotion recognition system to execute parallelization in the classification process, self-attention for all sequences, heads, and positions must be obtained through this process. When a relative position is taken into factoring, the representation for each position varies. By doing this, the computational complexity is reduced.

$$Gx_{a,b} = \frac{p_a D^E (p_b D^z)^F + p_a D^E ({x_{ab}}^z)^F}{\sqrt{y_l}} \qquad \text{(x)}$$

In this case, the input sequence p = $(p_1, \dots, p_z)$ contains n items where p_a ∈ H y_p. It then calculates the resulting sequence of an identical dimension, l = $(l_1, \dots, l_z)$, where l_a ∈ H y_l.

The parameter matrices are D^E, D^z, and $D^j \in H^{y_p \times\, y_l}$. Each layer and attention head has an individual collection of parameter matrices. "x" represents the edge input character. Currently, researchers have revised the SoftMax function for the suggested solution, Equation. (xi) by enhancing Equations (viii) and (x).

$$G\partial_{ab} = \frac{X(Gx_{a,b})}{\sum_{b=1}^{zh} X(Gx_{a,b})} \qquad \text{(xi)}$$

Thus, Equation (v) is altered to create Equation (xi). Here the updated value of G∂_{ab} affects the ability to simulate synthetic language samples, if it is included with the attention weights. Although in previous techniques, samples were selected based on chance, in the proposed framework will employ the attention weights to sample.

$$GC_{SIEOS} = G\partial_{ab} * C_{rep} + (1 - G\partial_{ab}) * C_{NN} \qquad \text{(xii)}$$

Equations (vi) and (vii) have been enhanced as follows:

$$\acute{G} = D^F GC_{SIEOS} \qquad \text{(xiii)}$$

$$\acute{H} = GC_{SIEOS}\ G{\partial_{ab}}^F$$

Here, the attention map is changed to the above equation which may be used in a variety of ways when the SIEOS method is executed to analyze the sample input data. This attention map's primary goal is to locate the appropriate information at the appropriate position.

4. Experimental Results and Analysis

The implementation process is conducted on a 2.7 GHz Intel Core i5 processor with 8 GB of 1867 MHz DDR3 memory. “Python 3.7.0 is the programming language, while PyCharm 2019.2 is a program utilized to implement this research. For this implementation, the Python modules numPy, tensorflow 1.3.0, sklearn, wave, cPickle, os, and python_speech_features” were used. “Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Interactive Emotional Dyadic Motion Capture (IEMOCAP)” databases were used for testing the relation-aware self-attention process that was the suggested approach in this paper. The prior state of the art had proposed an attention-based system. Since the number of samples in each of the chosen datasets varies, 10-fold cross-validation in which the training data is split into ten equal-sized definitions at random is performed. The IEMOCAP database, with a training batch loss of 1.5 percent and an accuracy of 90%, is used to train this algorithm. The study examines the following circumstances in the first part: a medium-sized sample from a limited number of recordings, a large dataset with a high number of samples, and a small dataset with a few samples.

Table 4.1. Dataset statistics applied to evaluate the proposed approach

Dataset Name	**Samples per Category**	**Emotional Samples**
IEMOCAP	50	350
RAVDESS	150	1000

90% of the data from each dataset was applied for training, and 10% was implemented as validation data. Table 4.1 shows the training and validation outcomes. Tables 4.2, 4.3, and 4.4 provide the findings of the second phase of the study, which involves testing the self-attention-based CNN with the LSTM network framework for each of the two datasets. The metric function from the Python TensorFlow library calculates the accuracy and processing time for each of the two datasets. The average of the processing times and accuracy is determined using the mean technique provided by the NumPy module in Python. Figures 4.1 and 4.2 demonstrate the findings for various datasets. The predict function from the Python python-speech-features module is used to calculate the classification accuracy for each dataset review. A functional interface to the WAV sound system is provided by the wave function. When features are extracted from the labeled training data, a convolutional neural network framework is autonomously generated. To achieve the fully connected layer to perform the “ASER implying SoftMax classifier in the recognition phase, the recovered features from the feature maps are provided into the max pooling layer”. Evaluation of the validation data is immediately performed with the trained model.

Table 4.2. Performance of the proposed method and the existing method for various datasets in terms of accuracy and processing speed

Dataset Name	**Phase**	**Existing System**		**Proposed System**	
		Accuracy (%)	**Processing Speed**	**Accuracy (%)**	**Processing Speed**
IEMOCAP	Training	80.53	318	87.41	350
	Validation	79.24	301	82.07	312
RAVDESS	Training	84.26	257	83.16	284
	Validation	81.32	294	80.25	250

Figure 4.1 demonstrates the accuracy results of our suggested solution as well as the advanced method for the IEMOCAP dataset during the training phase. While both solutions attain comparable accuracy throughout the training phase, the suggested approach reaches its maximum accuracy in a shorter number of epochs. This shows that, when the model is being trained, the suggested solution minimizes the time it takes to manage an enhanced dataset.

Table 4.3. Processing speed and accuracy Findings from IEMOCAP speech sample experiments using the existing and proposed system

Emotions	**Existing Accuracy (%)**	**Proposed System**	
		Accuracy (%)	**Processing Speed**
Happy	90	94.26	30.49
Sad	89.45	80.41	29.07
Anger	91.27	85.37	25.43

Figure 4.1 represents both the accuracy performance of the state-of-the-art method and our proposed strategy for the IEMOCAP dataset during the training phase. Comparing the proposed approach to the state-of-the-art solution, the accuracy has significantly improved in fewer epochs. Table 4.2 shows the processing durations and accuracy for two differentiated datasets according to the training and validation phases. Let’s compare the existing method with our proposed solution using the data analysis and graphical representations produced from the collected results. Following are the tables and figures that show the comparison. One of two methods is used to achieve the results. The result for each dataset is provided in the form of accuracy and processing periods. The accuracy is measured using a percentage of correctly categorized samples versus the total labeled samples whereas the processing speed is defined in epochs necessary for the algorithm to reach its goal. Twenty percent of the samples in two datasets were subjected to this test. The overall average processing time is determined by averaging the processing durations of the training and validation stages' results, and the overall average accuracy is determined by averaging the accuracy results of the training and validation stages.

Table 4.4. Accuracy and processing speed Results of both the existing approach and the suggested approach for speech samples from RAVDESS datasets

Emotions	**Existing Accuracy (%)**	**Proposed System**	
		Accuracy (%)	**Processing Speed**
Happy	92.15	95.43	30.57
Sad	90.27	91.26	28.41
Anger	94.68	93.18	23.75

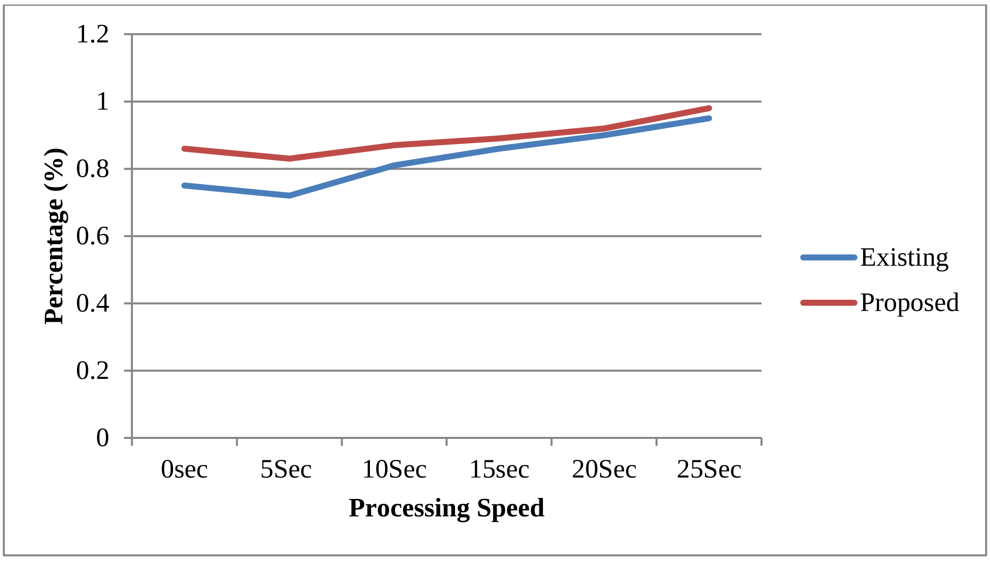

Fig. 4.1. The performance of the suggested and current solutions in terms of classification accuracy on the IEMOCAP dataset during the validation phase.

The metrics that are computed to assess the state of the art and the suggested solution are accuracy and processing times. These outcomes came from the convolutional neural network's classification phase. Tools like "Python with TensorFlow, cPickle, and the wave library" were used for this. Two stages comprise the analysis of the data: first, the model is trained and validated for every solution, and then it is reviewed for every one of the three datasets. The suggested method has reduced processing times and increased classification accuracy by preventing gradient saturation through Equation (x) and preventing over-fitting of the network by Equation (xiii), which reduces processing times. The following studies were performed to evaluate the proposed technique: one of which involved the removal of noise and the resulting extraction of features; the following one involved training the proposed CNN and BiLSTM network with the system for self-attention using IEMOCAP datasets; and the final component involved testing the model using two datasets.

In addition, the CNN and BiLSTM network with a self-attention feature achieved an accuracy of 94.26% for the IEMOCAP datasets and 95.43% for the RAVDESS datasets. Equation (xiv) is applied to compute the accuracy.

$$\text{Accuracy (Acc)} = \frac{Tp+TN}{Samples} \quad \text{(xiv)}$$

Where,

TP denotes the true positive → the quantity of accurately identified positive samples is known as True Positives.

TN represents the True Negatives → The number of negative samples that are accurately identified.

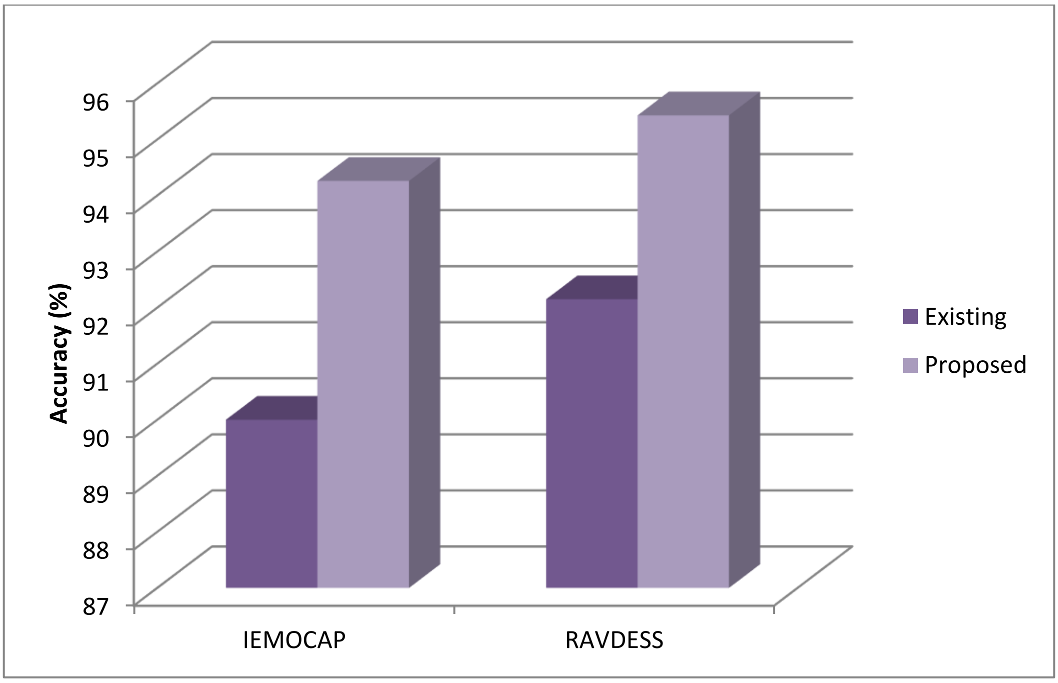

Fig. 4.2. Comparison of two datasets' classification accuracy using the existing and proposed system

The self-attention process is the primary characteristic of this approach, and it features an oversampling strategy to train the classifier if it can identify the various emotion classes. Achieving better classification accuracy has always been the main goal of the proposed approach [26]. A comparison of the two datasets' classification accuracy using the existing and proposed system is shown in Figure 4.2. The proposed approach in this study has overcome the existing system's limitation, with an average accuracy of 90.25% against 89.27%. Many modifications, including the oversampling method, data segmentation, noise reduction, and an upgraded self-attention process to improve seq2seq parallelization, were necessary to achieve this accuracy. Because of this, the proposed approach performs better even when the processing durations don't much increase.

5. Conclusion

The performance of an attention-based framework was evaluated in this work that concentrates on a speech emotion recognition framework. Several approaches were also examined for speech emotion recognition. The proposed Bi-LSTM network with a relation-aware attention system depends on CNN and is focused on the existing solution and its limitations. In this instance, relation-aware self-attention-based BiLSTM was utilized to characterize the long-term temporal development of emotional characteristics for all voice replies, whereas CNN was applied to simulate short-term temporal data. To improve the recognition accuracy, we suggested an equation that maintains the feature maps through the application of the oversampling methodology to customize the attention weights. In addition to solving the black box problems with deep learning approaches, the oversampling method is used to solve over-fitting problems.

To improve performance accuracy in the recognition of mood disorders, the proposed approach performed training. Subsequently compared the recommended approach's findings with those of the attention-based CNN-BiLSTM frameworks, everyone found that the proposed approach works significantly better while requiring almost identical processing speeds. With an average accuracy of 90.25% against 89.27%, the research's proposed approach has overcome the existing solution's limitations without significantly altering processing timeframes. Future research will examine several human characteristics that impact the ability to recognize emotions, including gender, age, and personality qualities.

References

[1] Li, S., Li, J., Liu, Q., & Gong, Z. (2022, June). Adversarial speech generation and natural speech recovery for speech content protection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7291-7297).

[2] Langari, S., Marvi, H., & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*, *20*, 100424.

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

[4] Mittal, S., Agarwal, S., & Nigam, M. J. (2018, November). Real-time multiple face recognition: A deep learning approach. In *Proceedings of the 2018 International Conference on Digital Medicine and Image Processing* (pp. 70-76).

[5] Huang, K. Y., Wu, C. H., Hong, Q. B., Su, M. H., & Chen, Y. H. (2019, May). Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5866-5870). IEEE.

[6] Mokgonyane, T. B., Sefara, T. J., Modipa, T. I., Mogale, M. M., Manamela, M. J., & Manamela, P. J. (2019, January). Automatic speaker recognition system based on machine learning algorithms. In *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)* (pp. 141-146). IEEE.

[7] Sokolov, A., & Savchenko, A. V. (2019, January). Voice command recognition in intelligent systems using deep neural networks. In *2019 IEEE 17th world symposium on applied machine intelligence and informatics (SAMI)* (pp. 113-116). IEEE.

[8] Kwon, M., & Choi, H. J. (2019, February). Automatic speech recognition dataset augmentation with pre-trained model and script. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 1-3). IEEE.

[9] Singh, A. P., Nath, R., & Kumar, S. (2018, November). A survey: Speech recognition approaches and techniques. In *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (pp. 1-4). IEEE.

[10] Byun, K., Song, E., Kim, J., Kim, J. M., & Kang, H. G. (2019, June). Excitation-by-SampleRNN Model for Text-to-Speech. In *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)* (pp. 1-4). IEEE.

[11] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., ... & Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[13] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4960-4964). IEEE.

[14] Hussain, M., Abishek, S., Ashwanth, K. P., Bharanidharan, C., & Girish, S. (2021, May). Feature Specific Hybrid Framework on composition of Deep learning architecture for speech emotion recognition. In *Journal of Physics: Conference Series* (Vol. 1916, No. 1, p. 012094). IOP Publishing.

[15] Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440-1444.

[16] Peng, Z., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2018, July). Auditory-inspired end-to-end speech emotion recognition using 3D convolutional recurrent neural networks based on spectral-temporal representation. In *2018 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.

[17] Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2020). Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, *8*, 16560-16572.

[18] Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., & Li, C. (2018). Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition.

[19] Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019). Exploring deep spectrum representations via attention-based recurrent and convolutional

neural networks for speech emotion recognition. *IEEE Access*, *7*, 97515-97525.

[20] Li, R., Wu, Z., Jia, J., Zhao, S., & Meng, H. (2019, May). Dilated residual network with multi-head self-attention for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6675-6679). IEEE.

[21] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control*, *47*, 312-323.

[22] Lotfian, R., & Busso, C. (2019). Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals. *IEEE Transactions on Affective Computing*, *12*(4), 870-882.

[23] Li, Y., Zhao, T., & Kawahara, T. (2019, September). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Interspeech* (pp. 2803-2807).

[24] Zayene, B., Jlassi, C., & Arous, N. (2020, September). 3D convolutional recurrent global neural network for speech emotion recognition. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* (pp. 1-5). IEEE.

[25] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

[26] Muruganandam, S., Joshi, R., Suresh, P., Balakrishna, N., Kishore, K. H., & Manikanthan, S. V. (2023). A deep learning based feed forward artificial neural network to predict the K-barriers for intrusion detection using a wireless sensor network. *Measurement: Sensors*, *25*, 100613.