Computers, Materials & Continua
DOI: 10.32604/cmc.2023.031177
Article

Tech Science Press

The Efficacy of Deep Learning-Based Mixed Model for Speech Emotion Recognition

Mohammad Amaz Uddin[1], Mohammad Salah Uddin Chowdury[1], Mayeen Uddin Khandaker[2,*], Nissren Tamam[3] and Abdelmoneim Sulieman[4]

[1]Department of Computer Science and Engineering, BGC Trust University Bangladesh, Chittagong, 4381, Bangladesh
[2]Centre for Applied Physics and Radiation Technologies, School of Engineering and Technology, Sunway University, Bandar Sunway, Selangor, 47500, Malaysia
[3]Department of Physics, College of Sciences, Princess Nourah bint Abdulrahman University, P.O Box 84428, Riyadh, 11671, Saudi Arabia
[4]Department of Radiology and Medical Imaging, Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia
*Corresponding Author: Mayeen Uddin Khandaker. Email: mayeenk@sunway.edu.my
Received: 12 April 2022; Accepted: 23 May 2022

Abstract: Human speech indirectly represents the mental state or emotion of others. The use of Artificial Intelligence (AI)-based techniques may bring revolution in this modern era by recognizing emotion from speech. In this study, we introduced a robust method for emotion recognition from human speech using a well-performed preprocessing technique together with the deep learning-based mixed model consisting of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). About 2800 audio files were extracted from the Toronto emotional speech set (TESS) database for this study. A high pass and Savitzky Golay Filter have been used to obtain noise-free as well as smooth audio data. A total of seven types of emotions; Angry, Disgust, Fear, Happy, Neutral, Pleasant-surprise, and Sad were used in this study. Energy, Fundamental frequency, and Mel Frequency Cepstral Coefficient (MFCC) have been used to extract the emotion features, and these features resulted in 97.5% accuracy in the mixed LSTM + CNN model. This mixed model is found to be performed better than the usual state-of-the-art models in emotion recognition from speech. It also indicates that this mixed model could be effectively utilized in advanced research dealing with sound processing.

Keywords: Emotion recognition; Savitzky Golay; fundamental frequency; MFCC; neural networks

1 Introduction

Emotion can describe a person's present situation. It can be evaluated in different ways, like physiological signals, facial expressions, or speech. The experiment of emotion recognition from human speech plays an important role in various real-time Human-Computer Interaction (HCI)

applications such as call centers [1], games, robotics, patient treatment, medical studies, etc. In the field of robotics, if a robot can understand a person's emotion through speech, it can adapt suitable behavior to communicate with the speaker. Albeit, it is quite difficult to recognize emotions from different people because people have different verbal expressions and amplitudes. Moreover, there exist several categories of emotions like sadness, fear, happiness, disgust, surprise, anger, and neutral in every person. Those emotions are created by a combination of love, affection, excitement, sorrow, and many other things.

In speech emotion recognition, appropriate feature selection from speech signals is the main task to be executed. The outcome of a model is directly dependent on the feature selection. If the feature selection is not good enough or appropriate, then the outcomes of the experiment don't show good accuracy. Modulation Spectral Features (MSFs), Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstrum Coefficients (LPCCs), Fundament frequency (F0), Zero crossing rate, Energy, etc. features have been widely used for the recognition of speech emotion. Although, these features are used for speech emotion recognition purposes, along with these features different types of acoustic features and harmonic features are also used for effectively classifying the emotion.

A review of the literature revealed that many studies [1–17] are conducted in recognizing the emotion from a human speech by using various methods or models. In some studies, the adopted models were further developed [12,15] or improved via different machine learning techniques for automatic emotion recognition. Nevertheless, most of the reported methods show non-negligible limitations because of the lack of proper selection of robust features and advanced machine learning methods. As a result, different researchers have attempted to come up with a viable method for selecting the appropriate features and improving artificial intelligence (AI)-based classification systems. Deep learning has grown fast since the beginning of the twenty-first century and demonstrates superior performance in a wide range of fields [18]. In particular, the use of deep learning algorithms shows promising in solving recognition or detection challenges such as speech recognition, emotion recognition, face recognition, gesture recognition, and object detection [19]. Nowadays, in addition to the single deep learning model, the hybrid ensemble deep learning model [18,20] is being used in a growing number of research studies. In this regard, the researchers have used several models for emotion recognition, including Convolution Neural network, Deep Belief Network (DBN), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) [2,5,7,8]. Nonetheless, in the case of speech emotion recognition, the recognition accuracy of these models was reported to be quite low. Such poor accuracy was attributed to the worst feature selection as well as poor model design.

In this study, we have introduced a method consisting of multi-modal filtering processes and deep learning-based mixed model. The principal contribution of this research work is to develop a good pre-processing technique for preparing the signal to extract the features from it, and also introduced a mixed LSTM + CNN model which may work better compared to many other existing deep learning-based single models for emotion classification. Along with this, we have measured the error rate using mean square error (MSE) technique as well as the performance of our model using different parameters such as precision, recall, and f1-score.

The remaining portion of this paper is structured in the following sections. In Section 2, we highlighted the review of different relevant research works in this field, and Section 3 represents the dataset information. Our proposed system with numerous sub-sections is shown in Section 4, Section 5 contains the proposed LSTM + CNN model with training and testing procedures, and Section 6 contains the detailed documentation of the result analysis and comparison with literature. Finally, Section 7 presents the conclusion of this work with some future directions.

2 Literature Review

Recently, in the field of artificial intelligence, many systems have already been developed or suggested by numerous researchers for speech emotion recognition. These include different types of real-time methods and approaches. Yoon et al. [1] proposed a method for speech emotions that is applicable to call center systems. This method can recognize two kinds of emotions such as neutral and anger from the speech which are captured by a cellular phone in real-time. Harár et al. [2] proposed a method to recognize three kinds of emotions (angry, neutral, and sad) using a Deep Neural Network (DNN). They have used a dataset from the Berlin Database of Emotional Speech which consists of 800 sentences with a sampling rate of 48 kHz followed by downsampled to 16 kHz (mono). The model achieved 96.97% accuracy on classification. Lalitha et al. [3] used the Berlin emotional database which consists of 10 speakers (5 male and 5 female) with seven emotions classes: sadness, anger, disgust, boredom, fear, happiness, and neutral. In this work, all of the seven emotions are recognized using pitch and prosody features, and classified these emotions using the Support Vector Machine (SVM) which gives 81% accuracy to recognize such emotions. However, the accuracy of this work is not good enough because of the selected database was not enriched with enough data and also, they didn't apply any kind of noise removal technique. Seehapoch et al. [4] used three language databases (Berlin, Japan, and Thai) to recognize seven emotions (Angry, Bored, Disgust, Fear, Happy, Natural, and Sad) using speech features like Fundamental Frequency (F0), Linear Predictive Coding (LPC), Energy, Zero Crossing Rate (ZCR), and Mel Frequency Cepstral (MFCC). In that work, SVM was used as a classifier which gives 89.80%, 93.57%, and 98.00% accuracy for the three emotions databases, respectively. Kerkeni et al. [5] used the Recurrent Neural Network (RNN) classifier to recognize seven emotions from the Berlin and Spanish databases. They extracted the emotion features by using MFCC and modulation spectral features (MSFs), and these combined features give 90.05% accuracy for Spanish emotional databases using the RNN classifier and 82.41% accuracy for Berlin emotional databases using the Multivariate Linear Regression (MLR) classifier. The number of selected features was too low as well as they didn't adopt any data preprocessing techniques that impacted the experimental accuracy. Jiang et al. [6] described a deep neural architecture to extract the acoustic feature of emotions from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset and used the SVM as a classifier to classify the emotions.

In [7], a CNN architecture was proposed to improve the accuracy of speech sentiment distinction and decrease the computational complexity of the speech detection model. Rawat et al. [8] proposed a method to recognize the emotion from human voice by analyzing the human speech signal. They used a high pass filter before extracting the emotion feature to reduce the noise from the speech signal as well as used Neural Network as a classifier to classify the speech emotions. Wan et al. [9] proposed a Dynamic Time Warping (DTW)-based speech emotion recognition model to recognize the mandarin Chinese oral speech emotions. Chen et al. [10] used a three-level speech classification model to categorize six emotions from the speech whereas the features were picked from 288 candidates by implementing the Fisher rate. Bisio et al. [11] presented a system consisting of two subsystems including gender detection and emotion detection from audio signals using SVM. It has been reported that the system can recognize six kinds of emotions including anger, boredom, disgust, fear, happiness, sadness, and the neutral state. Shaqra et al. [12] described a hierarchical classification model to analyze the emotion task. They manufactured two types of hierarchical models using the multi-layer perception (MLP) neural network and recognized that two separate models produce better performance than one model. They obtained the highest accuracy of 74% after combining the three models. In [13], Sequential Minimal Optimization (SMO) and Random Forest (RF) models together with spectral features have been proposed to perform the emotion recognition task. Moreover, three different types of databases

were used to measure the robustness of the proposed model to accomplish the job. The outcomes of this task showed that both models produced almost the same performance.

In ref. [14], MFCC and MSF were used as the features and linear regression, and SVM, RF, decision tree (DT), and CNN were used for classification. An interactive convolutional neural network (ICNN) was implemented [15] using two parallel channels MFCC for emotion recognition, which showed 96.32% accuracy. In this study, they worked with the CASIA database and provided multimodal fusion in the field of AI for emotion recognition. Although, those studies performed well, however, a poor data preprocessing technique was adopted for smoothing the data as well as the removal of the distorted audio signal which could help them to increase the accuracy. Mustaqeen et al. [16] presented a lightweight deep learning-based technique for speech emotion recognition named self-attention module (SAM). The authors used a multi-layer perception (MLP) to extract global features and a special dilated convolutional neural network (CNN) was used to extract spatial features from the input tensor. This experiment was executed using IEMOCAP, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Berlin Emotional Speech Database (EMO-DB) speech emotion datasets, and respectively showed 78.01%, 80.00%, and 93.00% accuracy for those datasets. In [17], the authors proposed a lightweight dilated CNN architecture that implements the multi-learning trick (MLT) approach for speech emotion recognition systems.

Overall, this survey reveals the non-existence of any prior study using deep learning-based mixed LSTM + CNN models for recognition of emotion from speech, thus this study forms such an interest with the mixed model. In this work, to overcome the existing drawbacks available in the literature, we have introduced a well-performed audio data preprocessing technique that helped to smooth out the data and also to solve the distortion problem. We extracted a good number of features and introduced a mixed model which works comparatively better than the usual state-of-the-art techniques/single model.

3 Dataset

To detect emotion from audio speech, we used the Toronto Emotional Speech Set (TESS) dataset, which consists of 2800 audio files. The audio files, recorded in English by two actresses (ages 26 and 64), contain seven emotions including anger, disgust, fear, happiness, surprise, sadness, and neutral. Each emotion has 400 audio files where 200 audio file is from a young lady (age 26) and the rest of the audio is from an old lady (age 64). This database was selected for this experiment because it contains a greater amount of data than the other available databases. Fig. 1 demonstrates a schematic flowchart to easily understand the data collection process in this study.

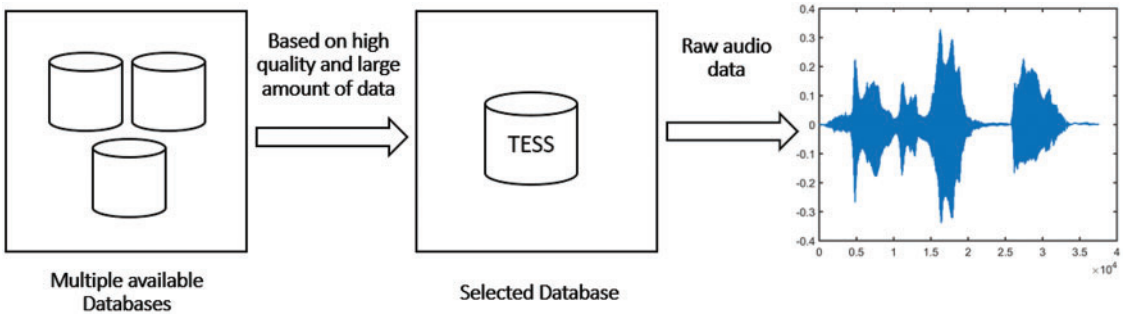

Figure 1: Data collection process

4 Proposed Method

The proposed method contains several filtering techniques and a feature extraction process. In pre-processing, to remove the noise from the raw audio data we have used a high pass filter. After removing the noise, the filtered data was mapped into a range using the linear interpolation method. At last, the Savitzky-Golay filter has been used to smooth the data. After pre-processing, the data was used to extract the features for emotion finding. MFCC, fundamental frequency, and energy are calculated from those pre-processed data. After completing the pre-processing and feature extraction, the extracted features are ready to be processed for the training using the mixed LSTM + CNN model. The proposed methodology is shown in Fig. 2

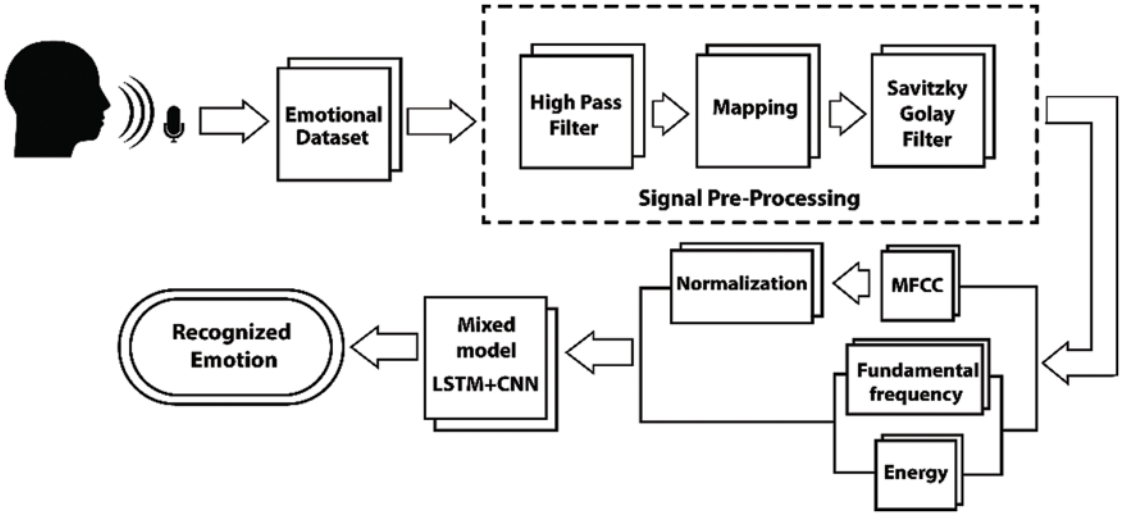

Figure 2: Architecture of the proposed method

4.1 Speech Data Preprocessing

4.1.1 High Pass Filter

It attenuates the signals that contain frequencies below a set point or the cut-off frequency. Mainly, it removes the low frequencies of the signal while allowing high frequencies to pass through. In this work, a high pass filter has been utilized to eliminate the undesirable noise close to the lower end of the audible range. To execute the high pass filter, a standardized passband recurrence of 0.05 has been utilized, which is indicated as a scaler in the period (0, 1).

4.1.2 Speech Data Mapping

In this study, we used a linear interpolation function to map values into a range. Generally, it makes new data points within the given range and provides interpolated 1D data. After filtering through the high pass filter, every audio data signal is mapped into 10000 values. The calculated mapped data using the interpolation function is shown in Fig. 3.

The linear interpolation function is shown in Eq. (1) for the two known data points (p1, p2) and (q1, q2).

$$q = q1 + (p - p1) \times (q2 - q1) / (p2 - p1) \tag{1}$$

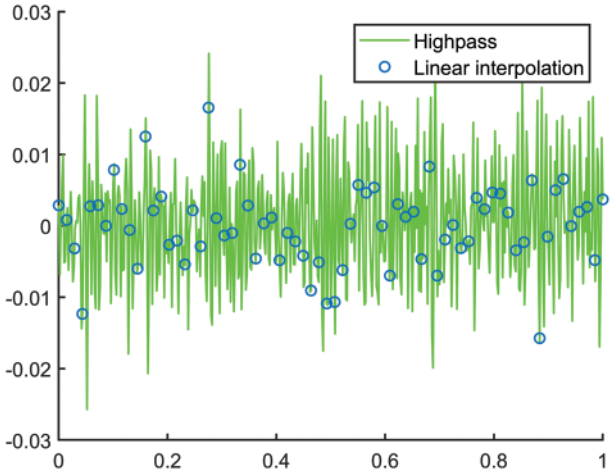

Figure 3: Speech data mapping using linear interpolation

4.1.3 Savitzky Golay Filter

It's a type of least square smoothing technique or digital filter for "smooth out" audio signal data with a wide frequency range. In general, it improves signal data precision without distorting or modifying the original signal. It also lowers the least-square error when fitting a polynomial to noisy signal data frames. For this type of filter, the frame size must be odd and the polynomial must be smaller than the frame size. To smooth the signal data in this investigation, we used a 2nd order polynomial filter with a frame size of 9. Fig. 4. depicts the speech data preprocessing.

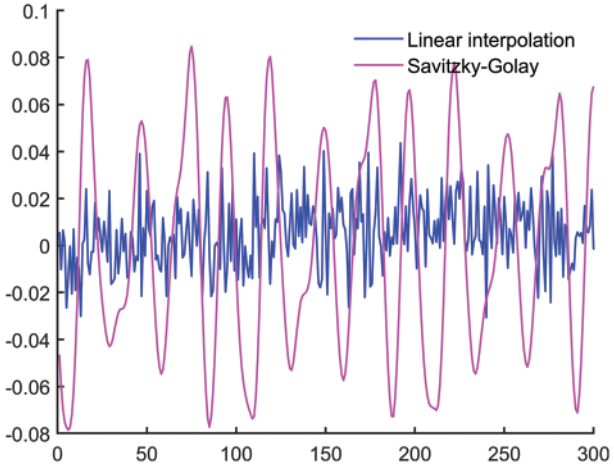

Figure 4: Speech data pre-processing

4.2 Feature Extraction

4.2.1 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is a feature of signal processing that is broadly used in automatic speech and speaker recognition [21]. It is determined by the properties of human hearing, which uses a nonlinear frequency

unit to mimic the human auditory system [4]. The MFCC feature extraction technique depends on many processes.

The first step in extracting the MFCC feature from the speech signal is frame blocking and windowing. The speech signal is converted to a short edge in this step to determine the coefficient or power spectrum. In this study, the audio sample rate is considered to be 24414 Hz. Every audio signal is split into frames of 30 ms. For the 24414 Hz audio signal, we obtained $24414 * 0.03 = 732$ frame length. Using an overlap length of 20 ms ($24414 * 0.02 = 488$ samples) the divided frames overlap the adjacent frame. After measuring the frame, Discrete Fourier Transform (DFT) is used in the next step to transform each windowed frame from the time domain to the frequency domain to obtain a frequency spectrum with Eq. (2).

$$A[f] = \sum_{t=0}^{L-1} a[t]\, e^{\frac{-j2\pi ft}{L}}, \mathrm{t} = 0, \ldots, \mathrm{L} - 1 \tag{2}$$

Here, the frequency domain is denoted by f, the length of the sequence to be changed is marked by L, and the time domain is denoted by t.

After using the DFT, the Mel filter bank has been determined in MFCC. It is usually executed in the time area and frequency domain. A Mel filter bank is typically implemented in the frequency domain. Moreover, MFCC also uses Band Edges to create half-overlapping triangle filters that modify the frequency data to replicate the nonlinear sound that humans hear. The filter bank's band edges are described as a non-negative expanding row vector in the range [0, sample rate/2] and measured in Hz.

The non-linear rectification utilized in MFCC is equivalent to the Discrete Cosine Transform (DCT), which is denoted by the Log. The acoustic variants that aren't important for speech recognition are reduced by log. DCT shows a limited group of information focused on several cosine functions influencing at different frequencies [22]. Generally, it is used to convert the log Mel spectrum into the time domain and provide a set of MFCC. It generates an N-by-M matrix of features, where N is the number of partitioned analysis frames of the speech signal and M is the number of coefficients returned per frame. In this study, MFCC produces 14 coefficients for each frame (a total of 35 frames available in each audio).

4.2.2 Z_Score Normalization

This technique is mainly used to modify the value of data in the dataset in a common range without distorting or losing its actual character or information. We have used z_score normalization in our experiment to handle or reduce the outlier issue with Eq. (3). Although it handles the outlier issue, it does not provide the normalized value with the same scale. Generally, it measures the distance of the data point based on the subtraction of a data value (p) from the mean (μ) value in terms of the standard deviation (σ) feature.

$$Z\_score = \frac{p - \mu}{\sigma} \tag{3}$$

4.2.3 Fundamental Frequency

The fundamental frequency is one of the most important key information sources for the automatic recognition of emotion [23]. It is measured in Hz and is defined as the average number of oscillations per second. One of the most significant aspects of voice recognition is the fundamental frequency, which reflects the signal's actual physical phenomena. To determine the fundamental frequency, we have used the autocorrelation method, which is a time-domain method. It is used when a

signal is compared to a time-delayed version of itself. The autocorrelation function is mathematically represented in Eq. (4).

$$A(k) = \frac{1}{T}\sum_{t=0}^{T-1} s(t)\, s(t+k)\,,\ k = [0,\ \mathrm{T}-1] \tag{4}$$

A(k) is an autocorrelation function at lag, where k addresses the lag. For all t values, the speech signal S(t) is defined. T is the size of a speech signal's window. According to Eq. (4), when the lag value is 0, it reflects the maximum value. After measuring the autocorrelation values from the speech audio data, the maximum, minimum, and average values of the fundamental frequency are calculated by using Eq. (5).

$$f_0 = 1/T \tag{5}$$

Here, f_0 represents the fundamental frequency.

4.2.4 Energy

Energy is the motion of energy through a substance in waves. The amplitude of the speech signal varies significantly over time. Short-time energy is a useful formulation that captures these amplitude changes. The most important aspect of this is that it establishes a foundation for distinguishing voiced from unvoiced speech. In this experiment, the speech energy has been calculated by using Eq. (6).

$$E = \sum_{n=1}^{n=N} X(n)^2 \tag{6}$$

Here, X (n) = Amplitude of speech signal, E = Speech energy, and N = Total number of amplitudes.

5 LSTM + CNN Mixed Model Architecture

5.1 Creating LSTM + CNN Model

We have utilized the LSTM + CNN mixed model with Python's Keras package to train the dataset. The model has an input shape of (534, 1) which is taken by the first LSTM layer with a filter size of 64. Again, we have used the LSTM layer with a filter size of 128. Multiple convolution layers with filter sizes of 64 and 32, and kernel sizes of 3, 2, and ReLu activation have been utilized in this mixed model. Along with the convolution layer, there are many batch normalization and max-pooling layers of pool sizes 2 and 3 that have been used to create the model. In this model, we have used dropout layers at 30% and 25% rates to avoid overfitting problems. After the flatten layer, the model is provided with an output emotion that has a dense layer of 7 units with Softmax activation. Finally, the Adam optimizer has been used to compile the model, and Mean Square Error is utilized as a loss calculator. In Fig. 5, the mixed model of LSTM + CNN is depicted, and the details are given in Tab. 1.

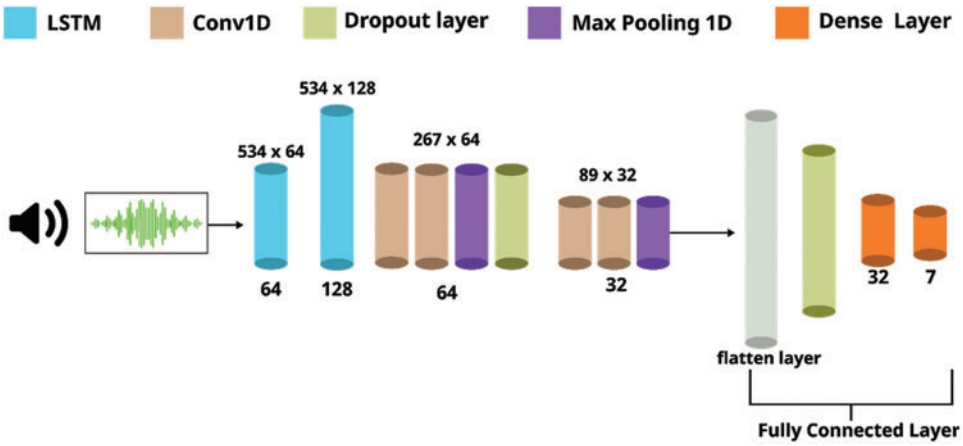

Figure 5: Proposed mixed LSTM + CNN model

Table 1: Detailed information on the LSTM + CNN model architecture

Layer (type)	Output shape	Parameter
lstm_1 (LSTM)	(None, 534, 64)	16896
lstm_2 (LSTM)	(None, 534, 128)	98816
conv1d_1 (Conv1D)	(None, 534, 64)	24640
conv1d_2 (Conv1D)	(None, 534, 64)	12352
max_pooling1d_1 (MaxPooling1D)	(None, 267, 64)	0
dropout_1 (Dropout)	(None, 267, 64)	0
conv1d_3 (Conv1D)	(None, 267, 32)	4128
conv1d_4 (Conv1D)	(None, 267, 32)	2080
max_pooling1d_2 (MaxPooling1D)	(None, 89, 32)	0
flatten_1 (Flatten)	(None, 2848)	0
dropout_2 (Dropout)	(None, 2848)	0
dense_1 (Dense)	(None, 32)	91168
dense_2 (Dense)	(None, 7)	91168

5.2 Model Validation and Performance

To train the dataset with a 70–30 train-test set, we have used 100 epochs with a batch size of 32. After 100 epochs the training accuracy is 99.79% for emotion with 0.0005 MSE loss. On the other hand, we get the validation accuracy for emotion is 97.50% with 0.0056 MSE loss. Fig. 6. shows a graph of accuracy and loss across 100 epochs.

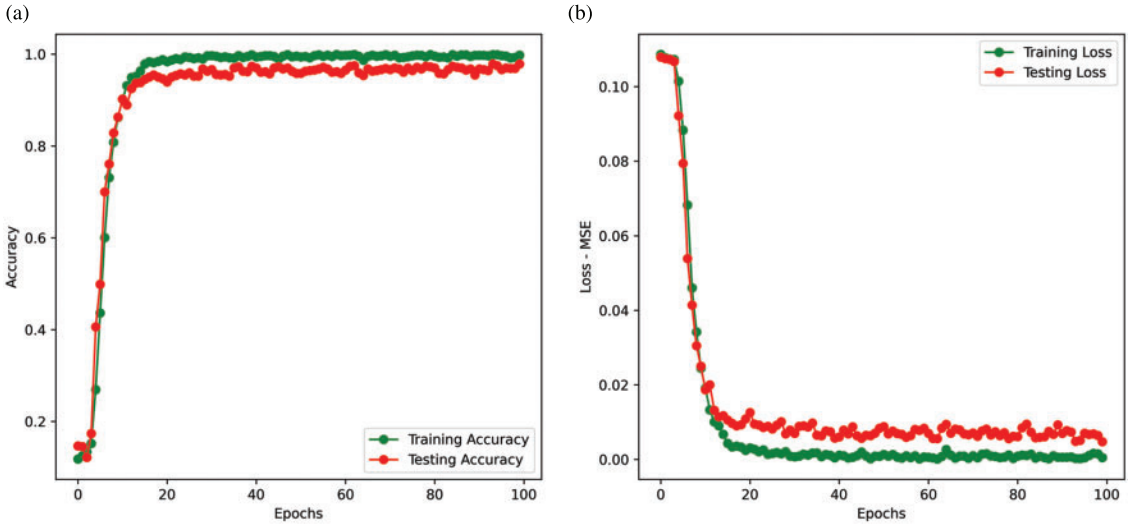

Figure 6: Training *vs.* testing (a) Accuracy and (b) Loss

6 Results and Discussion

The results have been analyzed by different measurements like confusion matrix, calculation precision, recall, and F1 score, and finally compared the proposed method and results with other similar kinds of literature. A total of 2800 audio files have been used in this experiment. Among them, 1960 audio files have been used as training datasets and 840 audio files as testing datasets. The training accuracy of the training dataset was 99.79%. Final test accuracy has been achieved at 97.5%. The confusion matrix for summarizing the performance of the model is given in Tab. 2.

Table 2: Confusion matrix of the emotion

	Angry	Disgust	Fear	Happy	Neutral	Pleasant surprise	Sad
Angry	117	1	0	1	0	2	2
Disgust	0	115	0	0	0	2	1
Fear	0	1	128	0	0	0	0
Happy	0	1	0	126	1	0	0
Neutral	0	0	1	0	114	0	0
Pleasant surprise	0	1	1	3	0	96	0
Sad	1	1	0	0	0	1	123

The precision, recall, and F1 score of the proposed LSTM + CNN model has been calculated, which shows excellent performance. To compute these values, first, it is needed to calculate the confusion matrix (shown in Tab. 3). True Positive refers to a class that is both positive and classed as (TP). True Negative refers to a class that is both negative and classed as (TN). False Positive occurs when a class is negative but labeled as positive (FP). Furthermore, when a class is positive but is

classified as negative, it is referred to as False Negative (FN). From these, we can conclude precision, recall, and F1-score shown in Tab. 3:

Table 3: Precision, recall, and F1 score for the testing set

Class	Precision	Recall	F1-score	Support
Angry	0.99	0.95	0.97	123
Disgust	0.96	0.97	0.97	118
Fear	0.98	0.99	0.99	129
Happy	0.97	0.98	0.98	128
Neutral	0.99	0.99	0.99	115
Pleasant surprise	0.95	0.95	0.95	101
Sad	0.98	0.98	0.98	126
Accuracy			0.97	840
Macro avg	0.97	0.97	0.97	840
Weighted avg	0.98	0.97	0.97	840

Precision: Precision is a measure of the ratio of positive prediction. It indicates how many total numbers of true positives are predicted.

$$Precision = true\ positive/\ (true\ positive + false\ positive)$$

Recall: Recall measures how many true positive cases are classified by the classifier from the whole data. Generally, it's the rate of TP and total positive observances in the real class.

$$Recall = (true\ positive/((true\ positive + false\ Negative)$$

F1-score: The F1-score is calculated by averaging precision and recall. Generally, it is the harmonic mean of precision and recall.

$$F1_{score} = 2\ \times\ \{(Precision\ \times\ Recall)\ /\ (Precision + Recall)\}$$

The proposed method's result displayed in Tab. 4 is compared to other existing approaches on speech emotion recognition systems. Our proposed mixed model shows better performance with an accuracy of 97.5%, however, other classifiers also showed a good performance producing an accuracy range of 78.20%−96.32%.

Table 4: Comparison of results with state-of-the-art methodologies

Year	Dataset	Algorithm	Features	Accuracy
2014 [3]	Berlin emotional database	SVM	Pitch and prosody features	81%
2018 [5]	Berlin Emotional Speech and Spanish Emotional Database	RNN, MLR	MFCC, MS	90.05%, 82.41%

(Continued)

Table 4: Continued

Year	Dataset	Algorithm	Features	Accuracy
2020 [14]	RAVDEES	CNN	MFCC, MS	78.20%
2020 [15]	CASIA	ICNN	MFCC	96.32%
This study	**TESS**	**LSTM + CNN**	**Proposed method**	**97.5%**

Particularly, in [3], researchers worked with the pitch and prosody features using the SVM model but they did not get high accuracy. Moreover, they worked with a small database, and also, they didn't use any kind of noise-removal technique in the pre-processing part. On the other hand, although Kerkeni et al. [5] and Christy et al. [14] worked with MFCC and MS as the features, their feature selection was not good enough as well as the number of features was low for achieving good accuracy. These limitations have been overcome in this study. We have introduced a good pre-processing technique to remove the noises and also smooth out the signal which help us to get good feature selections. To improve the traditional CNN model, an ICNN model was proposed for speech emotion recognition in [15]. This model showed good performance compared to other existing methods. However, our proposed mixed model gives better accuracy than the ICNN model.

7 Conclusion

This study presents an efficient system of emotion recognition using the deep learning-based mixed LSTM + CNN models. In this regard, necessary data on seven emotion classes (Angry, Disgust, Fear, Happy, Neutral, Pleasant-Surprise, and Sad) which consist of 2800 audio files were collected from the standard database 'Toronto Emotional Speech Set (TESS)'. Data were then pre-processed via several filtration steps like high pass filter, linear interpolation method, and Savitzky-Golay filter to obtain noise-free smooth data. MFCC is used to extract the emotion feature from these speeches, and it gives the feature vector which was trained and tested using the proposed mixed LSTM + CNN models. Our proposed mixed model provides an accuracy of 97.5% in emotion recognition, which indicates that the proposed mixed model performed better than the other state-of-the-art models available in the literature. It has been assumed that our adopted audio data preprocessing technique has performed better in smoothing the data as well as solving the distortion problem. In addition, our adopted good number of features was helpful for the mixed model to predict a better accuracy than the usual state-of-the-art techniques/single model.

In the future, we would like to implement more mixed models for this purpose as well as implement the proposed method into a real-time system. This research took a little bit more time in training and testing because the number of selected features for this recognition system is quite larger than other existing research works, but this showed good performance. We are now working on reducing computational time, which is a challenging task.

Acknowledgement: The authors express their gratitude to Princess Nourah bint Abdulrahman University Researchers Supporting Project (Grant No. PNURSP2022R12), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: The authors express their gratitude to Princess Nourah bint Abdulrahman University Researchers Supporting Project (Grant No. PNURSP2022R12), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. J. Yoon and K. S. Park, "A study of emotion recognition and its applications," *International Conference on Modeling Decisions for Artificial Intelligence*, vol. 4617, pp. 455–462, 2007.

[2] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," *International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 137–140, 2017. http://dx.doi.org/10.1109/SPIN.2017.8049931.

[3] S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition," *International Conference on Advances in Electronics Computers and Communications*, pp. 1–4, 2014. http://dx.doi.org/10.1109/ICAECC.2014.7002390.

[4] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," *International Conference on Knowledge and Smart Technology (KST)*, pp. 86–91, 2013. http://dx.doi.org/10.1109/KST.2013.6512793.

[5] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof and M. A. Mahjoub, "Speech emotion recognition: Methods and cases study," *International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 2, pp. 175–182, 2018.

[6] W. Jiang, Z. Wang, J. S. Jin, X. Han and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors*, vol. 19, no. 12, pp. 2730, 2019.

[7] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, pp. 183, 2019. https://doi.org/10.3390/s20010183.

[8] A. Rawat and P. K. Mishra, "Emotion recognition through speech using neural network," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 5, pp. 422–428, 2015.

[9] C. Wan and L. Liu, "Research of speech emotion recognition based on embedded system," *International Conference on Computer Science & Education*, pp. 1129–1133, 2010. https://doi.org/10.1109/ICCSE.2010.5593692.

[10] L. Chen, X. Mao, Y. Xue and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Processing*, vol. 22, no. 6, pp. 1154–1160, 2012.

[11] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Transactions on Emerging Topics in Computing*, vol. 1, no. 2, pp. 244–257, 2013.

[12] F. A. Shaqra, R. Duwairi and M. A. Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Procedia Computer Science*, vol. 151, pp. 37–44, 2019.

[13] A. R. Choudhury, A. Ghosh, R. Pandey and S. Barman, "Emotion recognition from speech signals using excitation source and spectral features," in *IEEE Applied Signal Processing Conf. (ASPCON)*, Kolkata, India, pp. 257–261, 2018. https://doi.org/10.1109/ASPCON.2018.8748626.

[14] A. Christy, S. Vaithyasubramanian, A. Jesudoss and M. D. Praveena, "Multimodal speech emotion recognition and classification using convolutional neural network techniques," *International Journal of Speech Technology*, vol. 23, no. 2, pp. 381–388, 2020.

[15] H. Cheng and X. Tang, "Speech emotion recognition based on interactive convolutional neural network," in *IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, Shanghai, China, pp. 163–167, 2020. https://doi.org/10.1109/ICICSP50920.2020.9232071.

[16] Mustaqeen and S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, pp. 107101, 2021. https://doi.org/10.1016/j.asoc.2021.107101.

[17] Mustaqeen and S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, pp. 114177, 2021. https://doi.org/10.1016/j.eswa.2020.114177.

[18] S. R. Zhou and B. Tan, "Electrocardiogram soft computing using hybrid deep learning CNN-ELM," *Applied Soft Computing*, vol. 86, pp. 105778, 2020. https://doi.org/10.1016/j.asoc.2019.105778.
[19] D. Zhang, J. Hu, F. Li, X. Ding, A. K. Sangaiah *et al.,* "Small object detection via precise region-based fully convolutional networks," *Computers, Materials and Continua*, vol. 69, no. 2, pp. 1503–1517, 2021.
[20] Mustaqeen, M. Ishaq and S. Kwon, "Short-term energy forecasting framework using an ensemble deep learning approach," *IEEE Access*, vol. 9, pp. 94262–94271, 2021. https://doi.org/10.1109/ACCESS.2021.3093053.
[21] Practical Cryptography. (n.d). Mel frequency cepstral coefficients (MFCC) tutorials. 2012. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.
[22] G. Vyas and B. Kumari, "Speaker recognition system based on MFCC and DCT," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 2, no. 5, pp. 145–148, 2013.
[23] C. Busso, M. U. R. T. A. Z. A. Bulut, S. U. N. G. B. O. K. Lee and S. S. Narayanan, "Fundamental frequency analysis for speech emotion processing," in *The Role of Prosody in Affective Speech*, Berlin, Germany: Peter Lang Publishing Group, pp. 309–337, 2009.