International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Speech Emotion Recognition Using Deep Learning

Ms. Apoorva Sharma[1], Mr. Himanshu Nawani[2], Mrs. Shalini Verma[3]

[1,2]Assistant Professor Computer Science, University of Petroleum Dehradun
[3]Assistant Professor Computer Science, Doon Institute of Engineering and Technology, Rishikesh

Abstract
It is frequently simple to identify emotions in human-to-human encounters since they may be seen in speech, body language, and facial expressions. Nevertheless, recognising human emotion in human-comFputer interactions (HCI) can be difficult. Speech emotion recognition (SER), with the aim of just identifying emotions through vocal intonation, has arisen as a way to enhance this connection. In this paper, a SER system based on deep learning methodologies is proposed. Here, RAVDESS dataset was utilised to assess the suggested system. To select the most appropriate vocal features that represent speech emotions for this MFCC is used. LSTM, CNN, and a hybrid model that combines CNN andLSTM are three different deep learning models we used to construct our SER system. By examining these various strategies, We were able to establish the most effective model for properly recognising emotional states from speech data in real-time situations. Overall, thiswork indicates the usefulness of the suggested deep learning approach.

Keywords: SER; HCI; emotions; CNN; LSTM; deep learning; MFCC; RAVDESS.

1. Introduction

Emotions influence our thinking and behaviour. Our everyday emotions can motivate us to act and influence thechoices we make in our life. To improve the effectiveness of human-machine interaction, machines must comprehend human emotions. As a result, a system for identifying voice emotions that may be employed in a variety of applications must be created. With everyday attempts to interpret spoken signals, voice recognition is the most rapidly increasing scientific discipline. This leads to the expanding research area of Speech EmotionRecognition (SER)[2], where multipleaccomplishments can lead to advancements in a range of domains such as translation systems and machine-to- human interaction employed in voice synthesis. SER is becoming increasinglysignificant in a variety of applications. At the moment SER is an advance topic of artificial intelligence (AI) and medical science. The research is frequently used in human-machine interaction, teaching,entertainment, and security domains, among others. A voice emotion processingand recognition system is often divided into three parts: speech signal acquisition, feature extraction, and recognition of emotions. SER is a problem that aims to extract emotions from audio signals. According to several studies, many systems are easy to operate due to advancement in emotion recognition.

Emotion recognition is a difficult problem because emotion changes depending on theenvironment.The challenge of understanding emotions from audio signals

is known as SER, and it is critical in the evolution of HCI. Emotion recognition from voice signals is an important yet difficult task for HCI. Many methods for extracting emotions from signals are used in the SER survey, including approaches such as voice analysis and classification. Deep learning algorithms were later suggested as an alternative to SER.

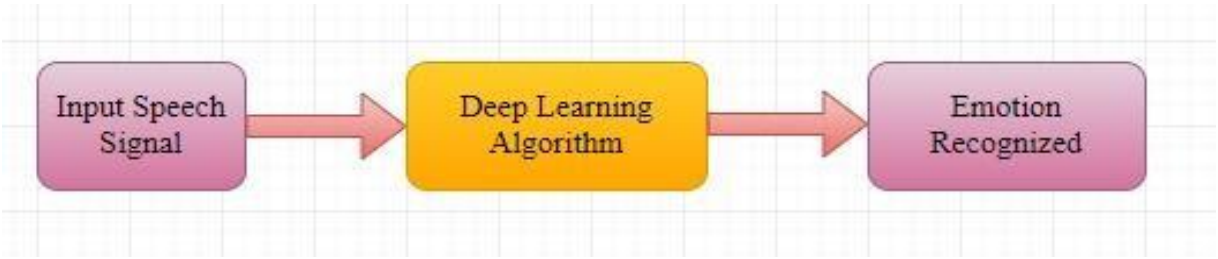

Fig 1. Deep Learning Flow Mechanism

SER is becoming increasingly significant in a variety of applications. SER is a growing field that straddles artificial intelligence and medical science. Teaching,HCI, entertainment, and security are all applications of the research. The processes used in SER are signal acquisition, feature extraction, and emotion recognition The most important method for voice recognition is based on neural network. Itis the process of convert an audio signal taken by a microphone or phone into a character string. Several models have previously been presented to improve the SER forecast accuracy.

SVM is one type of classifier that employs mathematics to predict emotion by determining the properties of an audio stream [3]. In the SER field, this practise has proven to be effective. SER is used ina variety of fields, including call centres and business process outsourcing, to determine customer satisfaction with a product using emotion.

Basu.et al. [2] published a brief examination of the significance of emotiondatasets and functions, noise reduction, and the importance of various type algorithms, including SVM and HMM, in 2020. The study's strength is the identification of several factors associated with voice emotion popularity; nevertheless, its weakness is the leaking ofextra modern techniques' evaluation, which tells us recurrent and convolution neural networks are deep methods.

Akcay et al. [1] survey dataset, capability, categorization, and emotion models. Furthermore, the examination is inquiring about device studying tactics for category development. Except for preliminary data from their unique studies, no equivalentoutcomes from chance approaches were published. This study contributes to other published surveys by providing a thoroughexamination of significant datasets and brief mastery of specific tactics in SER. The reason for no longer specialising in various previous tactics is recent advancements in neural networks, whichhave significantly deep understanding. That is, to the best of our understanding, the major analysis in SER that specialises in deep gains.

Humans have a natural ability to use all of their senses to get the most understanding of the message they have received. Emotion detection is a natural human talent, but it is a difficult task for machines.In its most basic form, speech processing operates on an audio signal. It is necessary for a variety of audio-based applications such as SER, speech noise removal, and music categorization. An SER system is made up of several components, including feature selection and extraction, classification, acoustic modelling, and language based modelling.

SER is utilised in numerous applications. The detection of anger can be used to improve the quality of voice portals or contact centres [3]. It enables the services given to be tailored to the emotional condition of the clients.

2. Literature Survey

In [1] Girija Deshmukh et al. suggested a system for acquiring audio samples of Short-Term Energy (STE), Pitch, and MFCC coefficients in the emotions of frustration, happiness, and melancholy. Natural speech was captured using open source North American English as expressiveness and feedback. As a consequence, just three emotions were identified: rage, happiness, and melancholy.They also recognised the speaker's distinctive qualities, such as sound, loudness, and pitch. The Ryerson Audio- Visual Database of Emotional Speech and Song (RAVDESS) dataset is manually separated into train and test sets. The multi-class Support vector machine (SVM) accepts feature vectors as input and generates a model for each emotion.

In [2] Peng Shi et.al created a discrete model and a continuous model of speech emotionrecognition; several aspects are analysed tocreate a better description of emotions. Deep Belief Networks (DBNs) beat Artificial Neural Networks (ANNs) and support vector machines (SVMs) by about 5%. The results show that the Deep Belief Networks-derived features considerably superior to the original feature. DBN-SVM outperformed DBN- DNN because SVM classifies better in small sizes. DBN transforms empty features into deep abstract qualities, resulting in better categorisation.

In [3] J. Umamaheswari et al. described pre-processing with K-Nearest Neighbour (KNN) and Pattern Recognition Neural Network (PRNN) algorithms, while feature extraction was explained with a descended structure encompassing Grey Level Co-occurrence Matrix (GLCM) and Mel Frequency Cepstral Coefficient (MFCC). The findings were compared in terms of precision rate, accuracy, and f-Measure with standard algorithms such as Hidden Markov Model (HMM) and Gaussian mixture models (GMMs), which were found to produce better results than the benchmark techniques. The emotional waves generate a pattern, which KNN laterrecognises. The KNN approach finds the most likely pattern that is closest to thesignal. There are six basic classes in the Speech database: neutral, fear etc

M.S. Likitha et al. discovered in [4] that recognition entails evaluating the verbal communication wave in order to define the required feeling, based on training of its features such as sound, format, and phoneme. A large variety of voice signal algorithms were established on the side of functionality withdrawal and examination. Communication kinesics' acousticprecision is a quality. The act of removing a little amount of information from a voicesignal that is then used to reflect each speaker is known as feature withdrawal. There are other extraction procedures available, but the most popular is coefficient extraction (MFCC). The sound that conveys the spoken signal is the feature. An activity that extracts a small amount of information from spokenutterance to be used later to act for each speaker is called feature extraction.

According to Zhang Lin et al. in [5] SER innovation is used to monitor the driver's abnormal emotions and employs particular word recognition innovation to select parking guidance in emergency scenarios. There are three types of extracted speech features: prosodic, spectral, and quality. Two often used characteristic factors in voice recognition (LPCC) are the Mel- cepstral coefficient (MFCC) and the Linear Predictive Cepstral Model. To extract the characteristics, SVM is utilised.The terror in the driver's words is recognised as an emergency circumstance by this method.

Asaf Varol et al. defined sound in [6] as a pressure wave formed by the vibration of components present in a molecule. Experiments with the speechsignal spectrogram and Artificial NeuralNetworks produce more effective outcomes (ANNs). SER accomplishes SER using the EMO-DB dataset using role extraction methodologies

In addition, they have investigated the rising scope of SERs in disciplines like as signal processing and pattern recognition in these contemporary developments. Furthermore,the author claims that different

machinelearning algorithms should be used to various sorts of datasets with various typesof tests in order to achieve higher success rates.

Abhijit Mohanta et al. [7] used emotional speech signal metrics including loudness, recognising voiced region, and excitation energy to analyse emotions like angry, frightened, glad, and neutral. These characteristics are referred to as sub-segmental characteristics. These feelings were examined using factors such asinstantaneous fundamental frequency (F0) using Zero Frequency Filtering (ZFF),signal strength, formant frequencies, anddominating frequencies. The study focuseson the features of the development of four different emotion states rather than the classification of emotional states depicted by the actor. Several signal processing techniques, including ZFF and STE, were used to determine instantaneous F0 and the zero-crossing rate (ZCR).

In [8], Edward Jones et al. found speech emotion recognition to be an intriguing component of Human Computer Interaction. (HCI). The primarymethodologies for SER must be feature extraction and feature classification. Both linear and non-linear classifiers can be used for feature classification. Common linear classifiers (BN) include Support Vector Machines (SVMs) and Bayesian Networks. These types of classifiers arebeneficial for SER since speech signals areconsidered changeable. Deep learning approaches offer extra benefits for SER when compared to prior methods. Deep learning algorithms do not require human feature extraction and tweaking and are capable of recognising complex structures within.

Michael Neumann et al. published their findings in [9], illustrating how learningabout unlabeled voice entities might help with Speech Emotion Recognition (SER). They used t-distributed neighbour embeddings (t-SNE) to analyse visualisations of various representations. In contrast, there are no divisible clustersin the 2D projections. Some plots are rejected because they require too much capacity. To train the auto encoder, a massive dataset is used. They have included auto encoder-generated representations, which has resulted in continual improvements in the SERmodel's identification accuracy. The research also enables us to identify andevaluate alternative auto encoders, as well as investigate procreative adversarial networks for representation learning.

3. Proposed Methodology

It contains the following steps: Data collection (3.1), data preparation (3.2), deep learning feature models (3.5), learning and testing of the constructed models, and finally classification. In our suggested SER technique, we use three models: LSTM, CNN, and CNN+LSTM. Furthermore, MFCC are employed as an input to improve the performance of our proposed models. The suggested speech emotion recognition system's overall design is depicted in Figure 1. In this paper,we compare deep learning algorithms and choose the best one based on accuracy and loss. Then create a hybrid model to improve overall system performance.

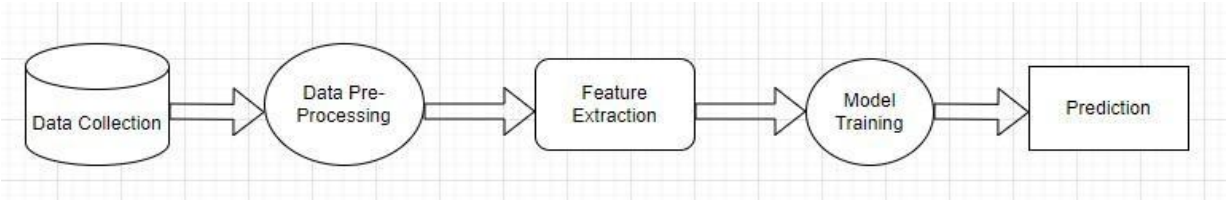

Fig2: Data Flow Diagram of SER Model

3.1 Data Collection

The most commonly used datasets for SERwere selected after getting a briefcomprehension of adequate domainexpertise. These significant datasets are available for free on Kaggle. This study examines the performance of the modelbuilt with these datasets and aids in thedevelopment of intriguing conclusions.

Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS):

The RAVDESS (Livingstone and Russo; 2018) has 1440 files and 24 professional actors, 12 female and 12 male, who read two lexically-related lines each. It is a dataset comprising 24 actors acting out 'happy, "sad,' 'angry,' 'fearful,' 'disgusted,' and 'neutral' emotions. RAVDESS has a lotof sample variations, and each emotion is played in two distinct intensities, with botha normal and singing voice. This is an important aspect of RAVDESS, and only afew datasets can claim to have it. Furthermore, RAVDESS is one of the few datasets with a North America English accent. The expressions for spoken emotions include: calm, happy, sad, angry,afraid, astonished, disgusted, and neutral. Each expression has two emotionalintensity levels (normal and strong), as well as a neutral expression. Every one of the 1440 files has a distinct file name. Figure 3 depicts the various categories of emotions found in the RAVDESS dataset.

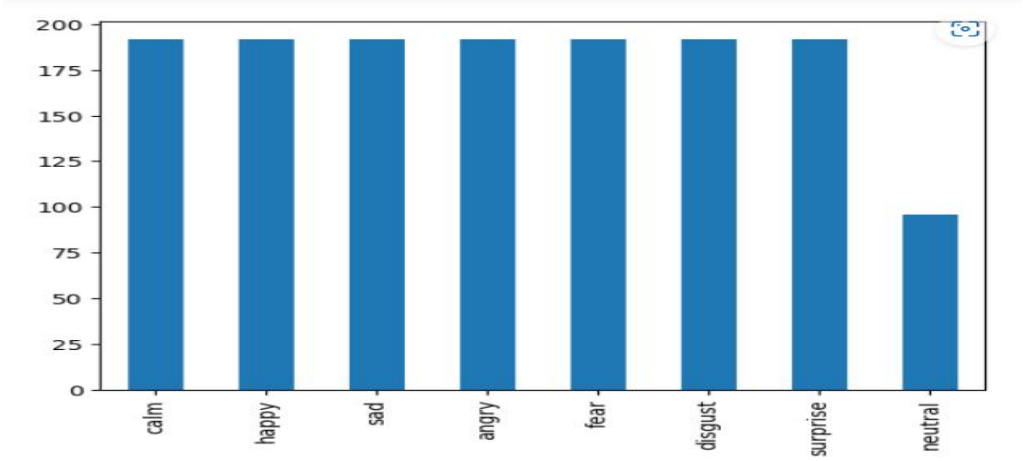

Figure3. RAVDESS Emotions Categories

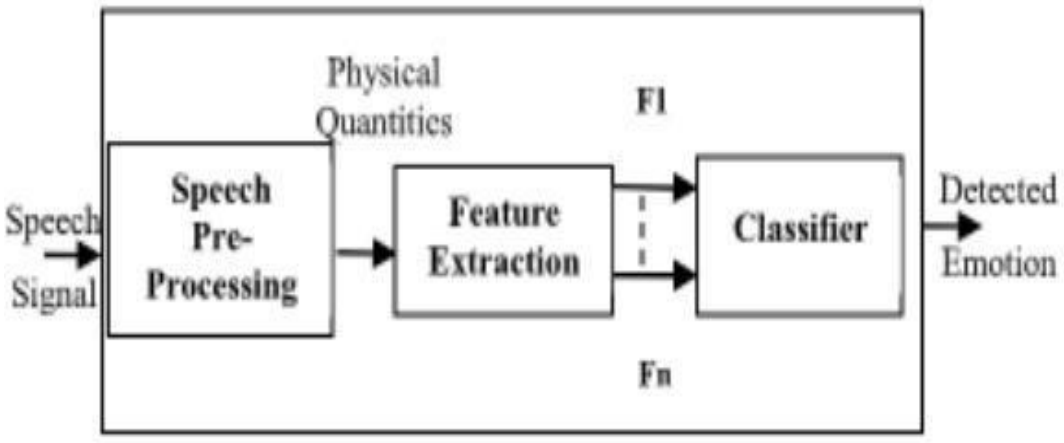

Figure4. SER System

3.2 Data Pre-Processing

Following the selection of the datasets, the first step was to identify and analyse the audio files. Each

dataset had its own name strategy. The emotion label was derived from the file names and then used to categorise the files. The records were then assessed using wave plots that exhibited randomly selected audio samples. Because there isn't much noise in this dataset, noise removal isn't required.

Data loading: Datasets are made up of data samples (.wav). This format cannotbe used as input to the proposed machine learning models. So, to begin, we import the data samples and convert them from an audio file type to a time series representation. We utilise librosa, a Python Programme for analysing and processingaudio signals, and the librosa.load()function to load audio files and convertthem to time series representations.

3.2.1 Data labelling: Data annotation,sometimes referred to as data enrichment, is crucial for our supervised approach to SER since it improves the accuracy and efficiency of the proposed machine learning models. Each sample in the datasets is labelled. For example, for each sample describing the emotion "happy," we assign a number (1), and so on.

3.2.2 Data Augmentation: We used two data augmentation techniques to improve the richness of the datasets used in thisstudy: noise addition and spectrogram shift.

3.2.3 Data splitting: During this final stepof data preparation, each dataset is divided into 75% for training/validation and 25% for testing.

3.3 Feature Extraction

The purpose of feature extraction is to keep as much information as possible while reducing the dimensionality of the input data.

The process of extracting valuable featuresfrom human speech in order to extract human emotion is referred to as featureextraction. The MFCC approach extractscrucial facts and features from subsets of speech data to analyze speech. In this study, we used MFCC to determine theemotional state of voice signals.

The Mel Frequency Cepstral Coefficients (MFCC) are without a doubt the most often utilised speech feature, since they are the most popular and flexible due to their accurate estimation of speech characteristics and efficient calculationmethodology [37].Speech signals are represented using the MFCC technique by turning their short-term power spectrum into a linear cosine transform of thelogarithmic power spectrum on a nonlinear Mel frequency scale.

3.4 Model Training

To train, we will utilise the fit() function of our model with the following parameters: training data (train X), target data (train Y),validation data, and the number of epochs. For validation data, we will use the test set provided in our dataset, which has been separated into X test and Y test. The number of epochs represents how many times the model will iterate over the data. To a degree, the more epochs we run, the better the model becomes. After that, the model will no longer improve with each epoch. For our model, we'll set the number of epochs at 100.

3.5 Model Training

To train, we will utilise the fit() function of our model with the following parameters: training data (train X), target data (train Y),validation data, and the number of epochs. For validation data, we will use the test set provided in our dataset, which has been separated into X test and Y test. The number of epochs represents how many times the model will iterate over the data. To a degree, the more epochs we run, the

better the model becomes. After that, the model will no longer improve with each epoch. For our model, we'll set the number of epochs at 100.

3.5.1 Existing Models

Some of the classification algorithm usedto classify the speech signals according tothe emotions are stated as follows:-

3.5.1.1 LSTM

Long Short Term Memory (LSTM)networks are a type of Recurrent Neural Network capable of learning order dependence. The previous step's output is used as input in the current phase of the RNN. The LSTM was designed by Hochreiter and Schmidhuber. It addressed the problem of RNN long-term reliance, which occurs when the RNN is unable to predict words stored in long-term memory but can make more accurate predictionsbased on current input. RNN does not function well as the gap length increases. By default, the LSTM may preserve information for a long period. It is employed in the processing of time series data, prediction, and categorization.

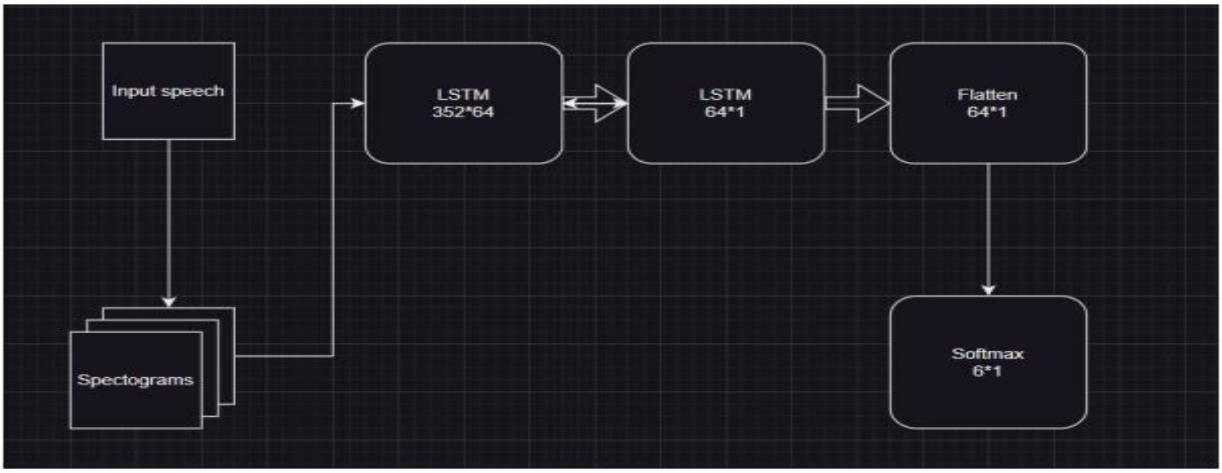

Fig5. Architecture of LSTM

3.5.1.2 Convolution Neural Network(CNN)

Another sort of neural network is CNN, which can find important information inboth time series and visual data. As a result, it's extremely useful for picture- related tasks including image identification,object classification, and pattern recognition. A CNN uses linear algebra methods such as matrix multiplication to discover patterns inside a picture. CNNsare also capable of classifying audio and signal data. A CNN's architecture is analogous to the connectivity pattern of the human brain. Just like the brain consists of billions of neurons, CNNs also have neurons arranged in a specific way. In fact,a CNN's neurons are arranged like the brain's frontal lobe, the area responsiblefor processing visual stimuli. This arrangement ensures that the entire visual field is covered, thus avoiding thepiecemeal image processing problem of traditional neural networks, which must be fed images in reduced-resolution pieces.Compared to the older networks, a CNN delivers better performance with image inputs, and also with speech or audio signal inputs.

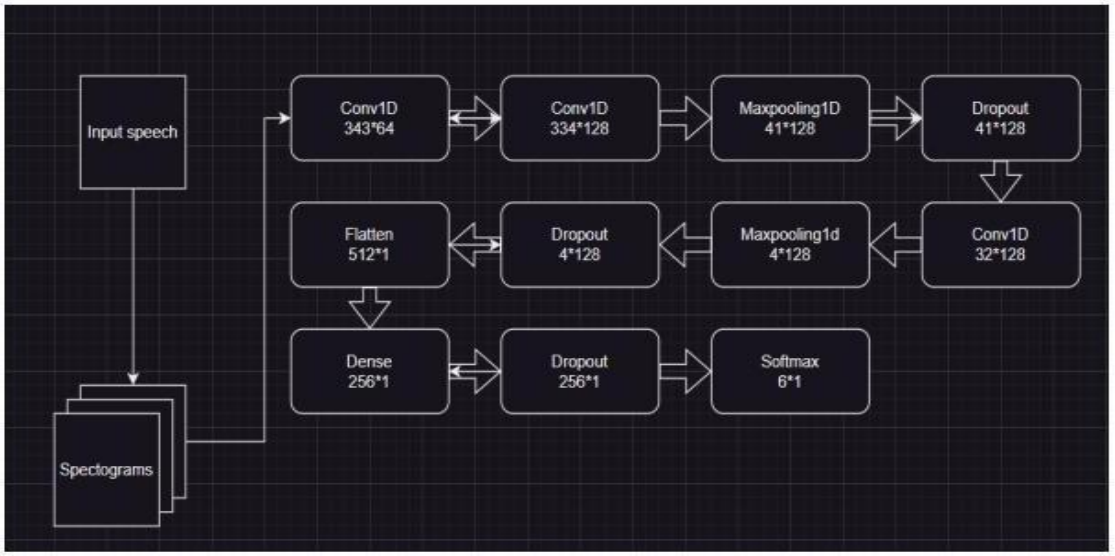

Fig6. Architecture of CNN

3.5.2 Proposed Model

In proposed model develop or build a hybrid CNN+LSTM model which give better accuracy then the existing modelslike CNN , LSTM , MLP.

3.5.2.1 Hybrid model (CNN+LSTM)

Speech is a complex signal having two major structural components: the textualsequence aspect, which refers to the numerical values of the signal, and the temporal aspect, which explains how material is dispersed across time. We create a model that combines two strong techniques, a Convolution Neural Network (CNN) and a Long Short-Term Memory (LSTM) network, for both characteristics and accomplish accurate speech analysis.

The feature vector from the LSTM layersis then flattened and transferred into the classification layer. The classification layer is a fully connected layer that returns the forecast of all activity classes. In the classification layer, a Softmax activation function is used to quantify the probability distribution of the activity classes and squashes all outputs to a scale from 0 to 1, so producing an applicable probability distribution.

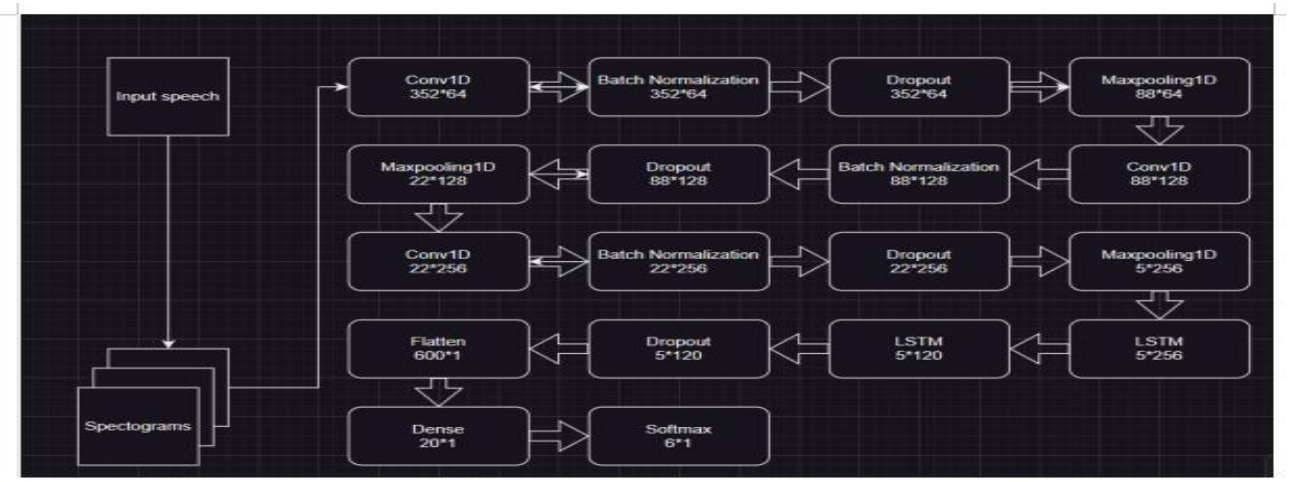

Fig 7. Architecture of CNN+LSTM

1.1 Evaluation Metrics

We must evaluate the performance of our deep learning model after it has been built and trained. Precision, recall, accuracy, and F1-score are four distinct criteria used to classify SER evaluation measures [9]. We utilise these measures to assess oursuggested model. Precision is defined as:

$$Precision = \frac{TP}{(TP + FP)} \dots (i)$$

And the recall is defined as :

$$Recall = \frac{TP}{(TP + FN)} \dots (ii)$$

Where TP is the number of true positivesin the dataset, FN is the number of false negatives, and FP is the number of false positives.

The F1-score is calculated by taking the harmonic mean of these two indicators (recall and precision), both of which must be high [9]. It's defined as follows:

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

The ratio of true outcomes (both true positive and true negative) to the total number of cases studied is defined as the accuracy[10]. It is defined as follows:

$$Accuracy = \frac{TP + TN}{Total\ population}$$

2. Experimental Results andDiscussion

In this paper, classification metrics such asthe confusion matrix, accuracy plots, loss plots, and precision, recall, and F1 score are used. The goal of SER research is to build strong and ready systems for recognising emotions.

On this study, weprovided a thorough examination of SER structures. It uses speech databases to deliver educational records. Following the pre-processing of the vocal sign, feature extraction is carried out. To improve HCI, the SER system requires more secure algorithms, as well as an emphasis on establishing classification approaches and a set of attributes. We are using a dataset of1440 files to achieve this criteria, with 60 trials per actor x 24 actors = 1440. It contain expressions like calm, happiness, sadness, anger, fear, surprise, and disgust.

2.1 Model Summary

CNN+LSTM

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 352, 64)	2944
batch_normalization (BatchNormalization)	(None, 352, 64)	256
dropout (Dropout)	(None, 352, 64)	0
max_pooling1d (MaxPooling1D)	(None, 88, 64)	0
conv1d_1 (Conv1D)	(None, 88, 128)	24704
batch_normalization_1 (BatchNormalization)	(None, 88, 128)	512
dropout_1 (Dropout)	(None, 88, 128)	0
max_pooling1d_1 (MaxPooling1D)	(None, 22, 128)	0
conv1d_2 (Conv1D)	(None, 22, 256)	98560
batch_normalization_2 (BatchNormalization)	(None, 22, 256)	1024
dropout_2 (Dropout)	(None, 22, 256)	0
max_pooling1d_2 (MaxPooling1D)	(None, 5, 256)	0
lstm (LSTM)	(None, 5, 256)	525312
lstm_1 (LSTM)	(None, 5, 120)	180960
dropout_3 (Dropout)	(None, 5, 120)	0
flatten (Flatten)	(None, 600)	0
dense (Dense)	(None, 20)	12020
dense_1 (Dense)	(None, 6)	126

Total params: 846,418
Trainable params: 845,522
Non-trainable params: 896

LSTM

Model: "sequential_3"

Layer (type)	Output Shape	Param #
conv1d_6 (Conv1D)	(None, 343, 64)	9664
conv1d_7 (Conv1D)	(None, 334, 128)	82048
max_pooling1d_5 (MaxPooling1D)	(None, 41, 128)	0
dropout_7 (Dropout)	(None, 41, 128)	0
conv1d_8 (Conv1D)	(None, 32, 128)	163968
max_pooling1d_6 (MaxPooling1D)	(None, 4, 128)	0
dropout_8 (Dropout)	(None, 4, 128)	0
flatten_3 (Flatten)	(None, 512)	0
dense_5 (Dense)	(None, 256)	131328
dropout_9 (Dropout)	(None, 256)	0
dense_6 (Dense)	(None, 6)	1542

Total params: 388,550
Trainable params: 388,550
Non-trainable params: 0

Model Created Successfully!

Fig 8. Comparison between three deep learning model

2.2 Comparison Chart

Model	Epochs	Accuracy	Validation Accuracy	Loss	Precision	Recall	F1-Score
1. CNN	100	0.4163	0.4144	1.5843	0.2671	0.9099	0.4130
2. LSTM	100	0.7478	0.5135	1.6639	0.5568	0.4414	0.4924
3. CNN+LSTM	100	0.9833	0.6306	2.3256	0.6296	0.6126	0.6210

Fig 9. Comparison chart of various deep learning models

2.3 Confusion Matrix

CNN+LSTM	LSTM	CNN
[[9 2 0 0 0 0] [0 13 1 0 0 0] [7 6 11 0 0 1] [2 7 0 8 1 2] [1 5 1 0 12 0] [1 0 2 1 1 17]]	[[5 0 2 0 2 2] [1 6 1 1 4 1] [2 2 10 2 4 5] [1 2 1 10 3 3] [0 0 2 1 15 1] [1 1 7 1 1 11]]	[[8 0 3 0 0 0] [3 8 0 0 2 1] [5 5 12 0 0 3] [0 13 1 3 1 2] [1 12 2 0 2 2] [3 2 4 0 0 13]]

Fig 10. Confusion Matrix of various deep learning model.

3. Conclusion

In this paper, we present a comprehensive study of a speech emotion recognition (SER) system that employs multiple acoustic features and three neural network models, namely LSTM, CNN, and CNN+LSTM architecture is designed to leverage the power of acoustic features,including MFCCs, to accurately classifyspeech emotions and diversity of thedataset samples.

The trials are carried out on the RAVDESS dataset, and the results show the usefulness of our suggested SER system. Specifically, the proposed model, a hybrid model CNN+LSTM, has the greatest accuracy rates of 98.33%. The existing model likewise performs well, with an accuracy of 74.78% for LSTM and 41.63% for CNN.Finally, the CNN+LSTM model exceeds the previous models with an accuracy of 98.99%. The findings show that the suggested SER system, which employs features and neural network models, can correctly classify speech emotions as well as other models.

References

1. Peng Shi, "*Speech Emotion RecognitionBased on Deep Belief Network*", Institute Of Electrical And Electronics Engineers,March 2020.

2. Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju and K. Hasitha *"Speech Based Human Emotion Recognition Using MFCC",* Institute Of Electrical And Electronics Engineers, March 2020.
3. Esther Ramdinmawii, Abhijit Mohanta, Vinay Kumar Mittal, *"Emotion recognition from speech signal",* Institute Of ElectricalAnd Electronics Engineers, Nov. 2021.
4. Michael Neumann, Ngoc Thang Vu, *"Improving Speech Emotion Recognition with Unsupervised RepresentationLearning on Unlabeled Speech",* Institute of Electrical and Electronics Engineers, May 2021.
5. Ak¸cay, M.B., O˘guz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods,supporting modalities, and classifiers. Speech Communication 116, 56–76 (2020).
6. Kaur, J., Kumar, A.: Speech emotion recognition using cnn, k-nn, mlp andrandom forest. In: Computer Networks and Inventive Communication Technologies: Proceedings of Third ICCNCT 2020, pp. 499–509 (2021), Springer.
7. Nam,Y.,Lee,C.: Cascaded convolutional neural network architecture for speech emotion recognition in noisyconditions. Sensors 21(13), 4399(2021).
8. Kwon,S.:lstm: Deep feature-based speech emotion recognition using the hierarchical convlstm network.Mathematics 8(12), 2133 (2020).
9. Alnuaim,Hatamleh: Human-computerinteraction for recognizing speech emotions using multi layer perceptron classifier, vol. 2022, Hindawi.
10. Aggarwal, A., Srivastava, N., Singh, D., Alnuaim: Two-way feature extraction for speech emotion recognition using deep learning. Sensors 22(6),2378 (2022).