Vol. 06, No. 3 (2024) 1391-1402, doi: 10.24874/PES.SI.25.03A.015

Proceedings on Engineering Sciences

www.pesjournal.net

SENTIMENT ANALYSIS ON SPEECH SIGNALS: LEVERAGING MFCC-LSTM TECHNIQUE FOR ENHANCED EMOTIONAL UNDERSTANDING

Suman Lata[1]
Neha Kishore
Pardeep Sangwan

Received 10.03.2024.
Received in revised form 22.04.2024.
Accepted 24.05.2024.
UDC – 004.85

Keywords:

Sentiments, Long short-term memory network, Speech emotion recognition, Mel Frequency Cepstral Coefficients, Deep learning

A B S T R A C T

The analysis of emotions expressed in spoken language holds a pivotal role in human communication, artificial intelligence, and human-computer interaction. While emotion recognition in text has seen considerable advancements, recognizing emotional states in spoken speech presents distinct challenges and opportunities. This research introduces an innovative approach that harnesses Mel Frequency Cepstral Coefficients (MFCC) and Long Short-Term Memory (LSTM) networks to facilitate deep emotion recognition in spoken speech signals. This study explores the profound potential of the MFCC-LSTM framework, a combination of established audio feature extraction and deep learning. Mel Frequency Cepstral Coefficients offer a powerful representation of spectral features over time, while LSTM networks excel at modeling temporal dependencies. This system classifies emotional states such as sadness, angry, neutral, and happiness from the speaker's utterances. Several performance assessments were carried out on the suggested MFCC-LSTM model. There is significant improvement in the recognition rates when compared with other models that are currently available. The proposed hybrid model reached 96 % recognition success.

1. INTRODUCTION

Human sentiment analysis assumes a significant part in endowing artifact with characteristics of human being and is stimulating an ever increasing number of considerations because of its potential application to machine human interaction. Human robots have been utilized in families or shops in household and retail settings which are fit for perceiving the important human feelings and adjusting their behavior according to the mood of their interlocutors. Inspite of the successful practical application, sentiment analysis confronting huge challenges because of its internal property of tangibility. To tackle this issue researchers utilize a variety of electronic devices to obtain sentiments in external signal form. With the advancement in hardware approaches, it is turning out to be not difficult to collect signals expressing human sentiments such as spectrograms, facial video sequences, audio signals, speech signals, acoustic waves and many more (Daneshfar et al., 2020). Sentiment recognition systems aims to directly interact with

[1] Corresponding author: Suman Lata
Email: suman@msit.in

computers through voice as opposed to conventional devices which interpret verbal content and facilitate humans to react. Business strategies- movie review, stock marketing, product feedback, call center conversations, utilization of sentiment patterns from speech in healthcare departments are some applications utilizing sentiment recognition system. Nevertheless there are a lot of challenges to HCI systems that require proper addressing, especially when these systems are transitioning from testing laboratories to real world applications. Therefore efforts are needed to properly address these issues and improve sentiment recognition by machines (Medhat et al., 2014).

Human emotions are difficult to predict. Although facial expression recognition is the method for identifying emotions but as people become older, they learn to manage their facial expressions. Few emotions such as boredom, dislike, or disgust, go unnoticed. Research is being done on speech based emotions recognition techniques to get around this problem. The Speech Emotion Recognition [SER] distinguishes between emotions on the basis of paralinguistic. Pitch, rhythm, pause, intonation, stress are some prosodic features which are extracted at preprocessing stage (Sandhya et al., 2020). Spectral characteristics analyze the frequency component from spoken signals. Classification of extracted features is the next step. The performance of SER system majorly rely on how exactly the features are extracted, that's why feature extraction is said to be the core part of SER systems. When compared with other algorithms, cepstral exhibits the best performance for pitch detection. Better accuracy for sentiment recognition can be obtained with fusion of cepstral features with MFCC and formants. High efficiency can be obtained by using MFCC with HMM by optimizing the acoustic parameters, number of states of HMM for each emotion and transitional probabilities between the states. By using HMM tool kit average accuracy of 78% is achieved authors (Vyas et al., 2015). Acoustic features have also been modeled by GMM on frame level. It has been noticed that using GMM at frame level is an expedient method for sentiment classification. Daniel (Neiberg et al., 2006) gained 80 % accuracy by this combination. Mostly classifier are good at neutral class but recognition at negative level (frustration) is not good because of its low frequency which results in poor training. As compared to traditional classifiers like HMM, KNN &GMM, SVM is the most efficient method with low complexity and high accuracy. 84% accuracy was gained for all emotions.(Dahake et al., 2016). Authors also focus on the points that for joy, sadness, highest accuracy is achieved by linear, quadratic function & for SER systems polynomials give worst results. Hybrid techniques will also give promising results in sentiment recognition. The MFCC were best suited for prosodic feature pitch. GMM is utilized as classifiers for training the module and in testing phase feed forward back propagation neural network is opted. Accuracy rates of 91 % is achieved by using this hybrid model (Devi et al., 2014). Moreover supervised learning back propagation neural network algorithms are also used for quantitative modeling and processing of data. Extracted features are trained with ANN for pattern recognition. It has been observed that the probability of confusion between anger and surprise is very high because of approximation of near formants and pitch acoustic features. These back-propagation algorithms prove to be an efficient method for emotion recognition and recognition accuracy of 83.5 % is obtained (Gilke et al., 2012).

1.1 Research gaps

Sentiment analysis of spoken speech signals is a challenging and evolving research area, and various acoustic features like MFCC (Mel-frequency cepstral coefficients), PLP (Perceptual Linear Prediction), and LPC (Linear Predictive Coding) have been used in this context. Identifying research gaps in this field can help drive further advancements. Here are some research gaps and areas that warrant attention:

1. Emotion Recognition: While sentiment analysis typically focuses on positive/negative sentiment, recognizing specific emotions (e.g., anger, happiness, sadness) from speech is an area that requires more attention. These emotions can provide more nuanced insights(Taherdoost & Madanchian., 2023).
2. Deep Learning Architectures: While deep learning models have been successful in various NLP tasks, their application in acoustic sentiment analysis is still relatively unexplored. Researchers should explore the potential of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for analyzing spoken speech signals.
3. Data Imbalance and Bias: Sentiment analysis models may suffer from data imbalance and bias, leading to skewed results. Researchers should focus on mitigating these issues to ensure fair and accurate sentiment analysis.
4. Long-Duration Speech: Most research focuses on short utterances. However, longer speeches or conversations may exhibit changes in sentiment. Models that can analyze sentiment over extended durations are needed.

1.2 Motivation and Novelty

The use of MFCC (Mel Frequency Cepstral Coefficients) and LSTM (Long Short-Term Memory) for sentiment analysis in spoken speech signals offer several novel aspects and motivations:

1. Novelty in Data Type: Sentiment analysis traditionally focuses on text data, such as reviews, tweets, or comments. Analyzing

sentiment in spoken speech signals represents a shift toward a more natural and direct mode of communication. This novelty arises from the need to adapt existing techniques to the specific challenges and characteristics of audio data.

2. Temporal Dynamics: Spoken language contains rich temporal dynamics that reflect emotional cues. MFCC captures spectral features over time, and LSTM is adept at modeling sequential data. Combining these two elements allows for the exploration of how sentiment evolves over the duration of spoken expressions, which is novel in comparison to static text-based sentiment analysis (Sangwan et al., 2023).
3. Real-time Applications: The use of MFCC-LSTM for spoken sentiment analysis is highly relevant in real-time and interactive applications. This includes customer service chatbots, voice assistants, and virtual therapists, where understanding the users emotional state during the conversation is crucial for providing appropriate responses.
4. Human-Machine Interaction: Sentiment analysis in spoken speech signals facilitates the development of emotionally intelligent human-machine interfaces. This enables machines to better understand and respond to human emotions, creating a more natural and engaging interaction. The novelty and motivation for using MFCC-LSTM for sentiment analysis in spoken speech signals lie in its potential to capture the emotional nuances of human communication, making it valuable in a wide range of applications and research areas.

1.3 Major contributions

The primary contributions of this work focuses on recognizing specific sentiments for example- happy, sad, anger, fear, disgust, neutral, pleasant surprise from speech by using deep learning models. The technique used in this work is MFCC-LSTM. MFCC provides a representation of the spectral features over time, and LSTM networks excel at modeling sequential data. This combination allows the model to understand how sentiment evolves throughout an utterance.

1.4 Paper organization

A review of state of arts on sentiment analysis and recognition with traditional techniques is given in section II. Preliminaries for the proposed work is explained in section III. Section IV covers the methodology used for sentiment analysis. Further, section V provides the experimental results. Finally, postscript conclusion is presented in section VI.

2. RELATED WORK

The voice signal has attracted researchers over the years for many applications such as sentiment perception, security, mental health diagnosis, business feedback, robotics, HCI, and others. Additionally, the major factors driving the increased prominence of research on sentiment analysis from the auditory signals of humans are increased due to a) productiveness of different digital signal processing techniques b) availability of computing systems c) acoustic variations in voice signals that are inherently present in distinct emotional situations. Primary emotions are classified into anger, disgust, fear, happy, neutral, and sad. By adding these basic emotions new ones will be produced such as love (trust + joy), delight (surprise +joy), guilt (fear + joy) and many more. These are called secondary emotions. Numerous studies in the area of SER have been carried out over the decades with general pipeline of dataset, preprocessing, feature extraction, sentiment classification. The extensive literature on sentiment analysis suggest that low level features and high level features can provide a more comprehensive picture of an individual's emotions (Cui et al., 2023). SER systems can be shaped by examining well-crafted features that effectively uncover each sentiment from voice signals. The varying energy level in the pitch, intensity of facial expressions, and rhythm of audio signals require instantaneous features and dynamic features for sentiment recognition. The instantaneous features are extracted from the data at single point for e.g speech rhythm, loudness, pitch while dynamic features are extracted over a period of time e.g change in intensity of facial expression and rhythm of speech over time(Deswal et al., 2019).

Speech signal consists of some basic characteristics such as- frequency, wavelength, amplitude, timber, duration. Some other factors which effect the voice signal is medium, shape of source, distance between source and listener. From the view point of physical interpretation features are categorized into voice source features, short term features, spectrotemporal features, prosodic features and high level features. High level features include pronunciation, personal lexicon, accent which depends on birth palace, language background, parental influence etc. Short term spectral and voice source features are related to spectrum, glottal pulse features and based on physiological factors like size of vocal folds, length and dimensions of the vocal tract. Spectrotemporal and prosodic features depend upon pitch, energy, rhythm & duration. In fact these prosodic and temporal features are responsible to classify various sentiment categories in both acoustic and perceptual terms(Chan et al., 2023). The above said features are extracted in frequency domain with the help of transforms and gained attention of researchers for many applications, because of their ability of presenting vocal cord variation features (Gupta et al., 2021).These features are analyzed by power spectrum where sound is

considered as a function of frequency. From these spectrums a set of features are derived using mel scale called MFCC. Logarithmic scale is preferred over the linear scale to represent Mel-frequency cepstral coefficients. Authors proposed basic preprocessing and feature extraction methods to analyze the basic voice characteristics e.g framing, segmentation, window length ,and overlap window type, frequency scaling , frequency ranges, filter banks , ceptstral features, signal magnitude, phase manipulation etc. It is analyzed that results can be improved by working on these basic preprocessing step (Kacur et al., 2021). Filters used in preprocessing stage plays an important role in removing noise from received audio samples. Researchers worked on vocal tract signals by using filter banks (Gammatone, mel, Bark). In next step features are extracted using pitch, energy, zero crossing rate, discrete wavelet transform (DWT) and MFCC algorithms (Koduru et al., 2020). The global feature algorithm is used during the feature selection step to eliminate redundant information and various machine learning classification algorithms are used to identify the emotions from the extracted features. Even if they are only a small portion of more complex scenario, each of these basic steps must be of utmost importance to get better results in SER systems. Next step is dimensionality reduction. Probabilistic principal component analysis (PPCA), principal component analysis (PCA), factor analysis methods can be utilized for this purpose. A new dimensionality reduction approach, i.e quantum behaved particle swarm optimization (QPSO) is proposed by Fatemeh. When training and testing data sets are different then features follow different distribution and will lead to reduction in recognition rate (Daneshfar et al., 2020). A novel transfer PCA algorithm is proposed by Peng. Different features are classified with various machine learning and deep learning approaches such as LDA, K-nearest neighbor, Navie bayes, RF, SVM, gradient boosting machine learning & many more (Song et al., 2015). Authors worked on LDA classifier and achieve a promising result for native English speakers (Krishnan et al., 2021). Researchers worked on MFCC and HMM to increase the recognition rate of SER systems by adding energy and speech speed to MFCC features (Hong et al., 2016). The primary contributions of this work is the ability to capture the temporal context of spoken speech. MFCC provides a representation of the spectral features over time, and LSTM networks excel at modeling sequential data. This combination allows the model to understand how sentiment evolves throughout an utterance.

3. PRELIMINARIES

3.1 Feature extraction

MFCC is the most commonly employed feature extraction method in sentiment recognition and other applications where it is important to represent the spectral characteristics of sound. MFCC are calculated using a mel scale from the power spectrum of a sound signal. The logarithmic mel scale is intended to be more perceptually relevant than the linear frequency scale. Robustness to noise, discriminative power, computational efficiency, Interpretability are some advantages of MFCC which makes it a good choice for sentiment recognition and other applications where accuracy and efficiency are more important (Sangwan et al., 2021). The speech signal is divided into short time frames. The Fourier transform is applied to each time frame to obtain the frequency spectrum. The frequency spectrum is converted to the Mel scale, which is a logarithmic scale that is more closely aligned with the human perception of pitch. The Mel spectrum is then filtered to retain only the most important frequencies. The filtered Mel spectrum is then passed through a Discrete Cosine Transform (DCT), which converts it into a set of coefficients. The DCT coefficients are the MFCCs. They are a compact representation of the frequency spectrum of the speech signal that is well-suited for use in speech recognition and other audio processing applications. Formula for finding out is- MFCC = DCT (log (Mel filter bank energies)).

3.2 Classifiers

LSTM is a type of RNN classifiers which are well suited for natural language processing applications. LSTM is utilized in sentiment analysis where long term dependencies between words in sequence are required. The vanishing gradient problem occurred in RNN is properly addressed by LSTM through the use of gating mechanism which controls how information is memorized in the network memory. Gating mechanism consists of three gates. a) The forget gate determines which information from the previous state should be forgotten. b) The input gate determines which new information should be added to the state. c) The output gate determines which information from the state should be output. The network is trained on a dataset that has been labelled with its sentiment. The network learns to identify the patterns in speech that are associated with each sentiment (Deshwal et al., 2023). Once the network is trained, it can be used to classify new speech segment. The network takes the speech utterance as input and outputs the probability of various sentiments. This probability can then be used to make a decision about the sentiment from the speech utterance.

4. METHODOLOGY

4.1 Dataset

In order to provide a realistic comparison, Toronto Emotional Speech Set (TESS) is used for training and testing the proposed sentiment analysis system from speech signals. The data has been taken from open source Tspace repository kaggle. The data set comprises of voice utterances of two actresses speaking 200 target words in native English. The utterances are captured

with seven different emotions namely disgust, anger, fear, happy, neutral, pleasant surprise, sad. The frame size of recorded samples is 3 sec. and sampling rate is 22KHz.For analysis 400 recordings of each sentiment class are taken to develop a sentiment analysis system database((Deshwal et al., 2020). It should be emphasized that the original recordings are of great quality and were made in a setting with minimal background noise, so no additional pre-processing is necessary. Figure 1. Illustrate the samples of audio files of different emotions for TESS dataset (Hazra et al., 2022).

The proposed framework utilizes MFCC approach for feature extraction and LSTM algorithm for classification of emotions. The work is carried out using python framework. Librosa and tensorflow libraries are used for converting raw data i.e. sound utterances into vector form and utilizing deep learning algorithms respectively. The proposed LSTM architecture has one input layer, two dense layer followed by softmax output layer. The following sections go over the essential elements of the suggested framework.

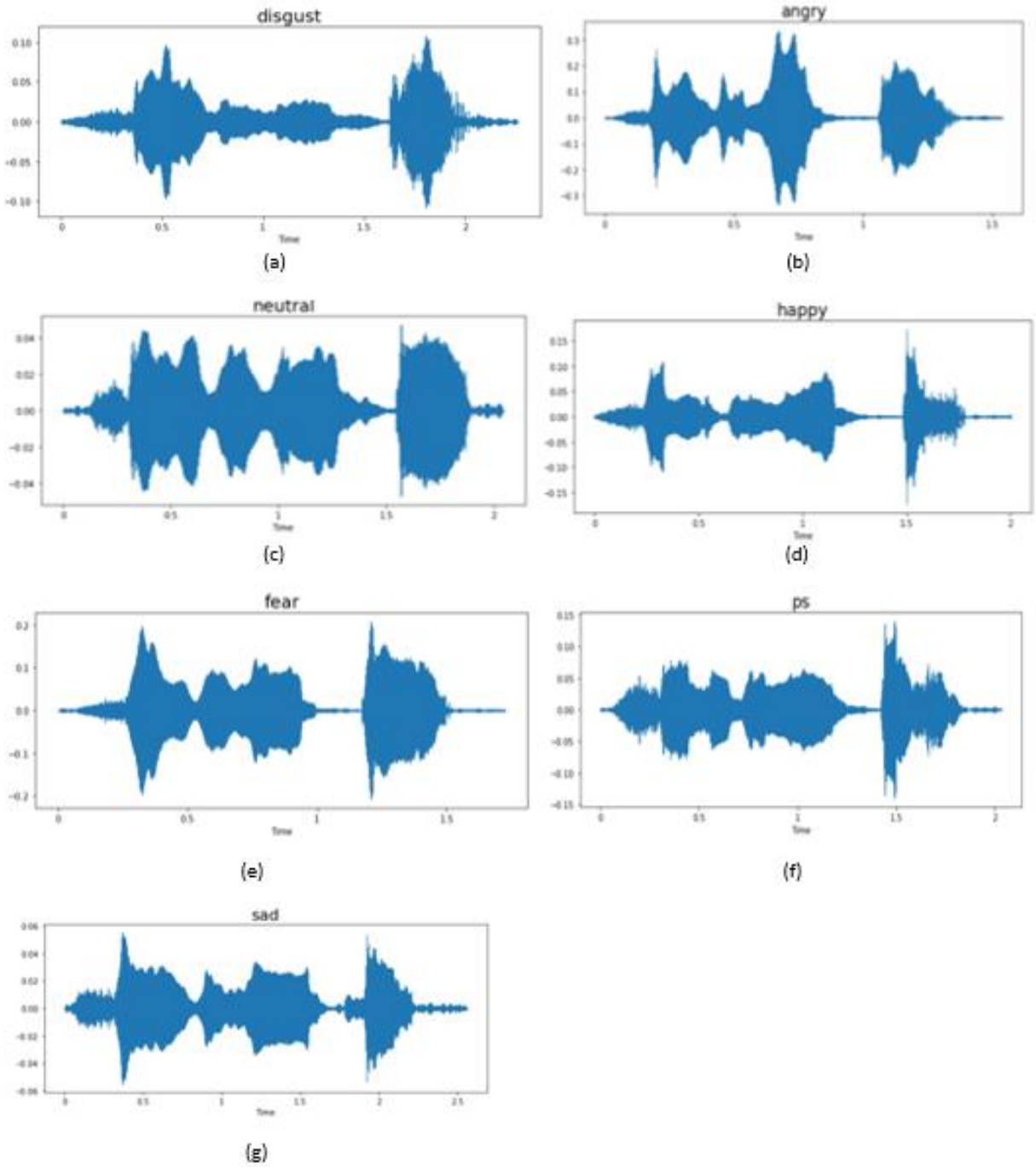

Figure 1. Audio Signals for different emotions: (a) Disgust, (b) Angry, (c) Neutral, (d) Happy, (e) Fear, (f) Surprise, (g) Sad

4.2 MFCC Generation

We have taken Toronto Emotional Speech Set (TESS) for training and testing which is publically available on Kaggle and since the data is recorded in noiseless environment so cleaning process is not required. This is the irony that the real world is analog and we have to deal with digital data, so a conversion of data is required for analyzing the sound signals. Because of random nature of speech signal there is a requirement of a technique which will analyze and capture the dynamics of speech signal over the time. Mel frequency cepstral coefficients technique fulfils this requirement. Measuring the average energy of signals is a valuable approach for excavating key characteristics from speech signal. By analyzing the average energy i.e signal power over time, one can identify patterns and changes that correspond to different emotions within the audio. Despite its non-stationary nature, the speech signal exhibits some quasi-stationary properties over short periods of time. This makes it possible to use stationary models to analyze the speech signal, which is important for speech processing. So the signal is divided into short overlapping frames, typically 20-30 milliseconds in length, to capture the temporal characteristics of speech. Speech sounds are produced by the vibration of the vocal cords, which generate a complex waveform that includes both low-frequency and high-frequency components. The energy of the speech signal decreases with increasing frequency. So to capture perceptual characteristics Pre-emphasis filter is utilized which helps to compensate for the natural roll-off of the human ear's sensitivity at lower frequencies. By amplifying the higher frequencies, pre-emphasis makes it easier to extract the relevant information from the signal. The impulse response of this pre-emphasis filter is given by the following equation.

$$H(z) = 1 - 0.97z^{-1} \quad (1)$$

The signal is abruptly truncated at the frame boundaries, so to minimizing the effect of Spectral leakage & smoothening of the signal at the edges of the frame, a window function is applied. The signal is multiplied by Hamming window function to avoid discontinuities between the frames. In our work 40 features are extracted within 3 second duration by taking 0.5 sec offset using librosa library and sampling rate is 22050 Hz. To represent the signal as a collection of frequencies and their corresponding amplitudes, it is converted from time domain to frequency domain with the help of FFT is used. The output of FFT is passed through mel filter bank which is a collection of band-pass filters that are used to capture the perceptual characteristics of speech. The filter outputs are collected to form the Mel energy spectrum which is calculated by following equation.

$$Mel(f) = 2595 \times \log_{10}(1 + f/100) \quad (2)$$

Where Mel (f) is the mel-frequency and f is frequency in Hz. The output of the Mel filter bank is a set of Mel energy coefficients, which represent the energy of the signal. These coefficients are then transformed into the cepstral domain using the Discrete Cosine Transform (DCT). The first few cepstral coefficients, typically 12-20, are retained as the MFCC feature matrix which is shown in equation 3 below.

$$M(r \times s) = \begin{bmatrix} m11 & m12 & m12 & \ldots\ldots. & m1s \\ m21 & m22 & m23 & \ldots\ldots. & m2s \\ m31 & m32 & m33 & \ldots\ldots. & m3s \\ . & . & . & . & . \\ . & . & . & . & . \\ mr1 & mr2 & mr3 & \ldots\ldots. & mrs \end{bmatrix} \quad (3)$$

Where M is the MFCC feature matrix, s represents the number of MFCC coefficients extracted per frame and r represents number of frames in audio signal. Let V1, V2, …….Vs be the s number of arrays developed by considering columns of matrix M. we use this above mentioned method for feature extraction in the proposed model. Figure 2 shows the MFCC array of the proposed work.

```
extract_mfcc(df['speech'][0])

array([-3.96986298e+02,  7.74405365e+01, -1.95927925e+01, -2.16666927e+01,
       -2.11275530e+00,  1.00753660e+01, -2.03667088e+01, -6.09244871e+00,
       -7.21227980e+00, -5.57366371e-01, -1.83255494e+00,  2.02103406e-01,
        7.27553129e-01,  1.31773472e+00,  2.88634229e+00,  2.85579419e+00,
       -4.71292019e+00, -4.43651009e+00, -1.62115526e+00, -1.02398405e+01,
       -7.55126143e+00, -1.79687607e+00, -7.03765059e+00,  9.43658543e+00,
        8.35586166e+00,  2.17123604e+01,  1.92169876e+01,  2.03489323e+01,
        1.34133692e+01,  8.33917236e+00,  3.94721806e-01,  5.11131763e+00,
        9.56873989e+00,  5.45486546e+00,  2.50996089e+00, -1.82390428e+00,
        4.86895704e+00,  9.31392193e+00,  2.08915091e+00, -1.90649259e+00],
      dtype=float32)
```

Figure 2. Extracted features in array form

In addition to average energy, other parameters related to the entire signal duration, such as mean value, standard deviation, maximum, and minimum, also play a crucial role in emotion classification (Badr et al., 2021). These parameters capture information about the pitch, timbre, and dynamic range of the audio signal, which are all important cues for conveying emotions.

4.3 LSTM

Long short term memory model is a powerful tool applied on long range sequential data. LSTMs can learn dependencies that are far apart in the sequence, unlike traditional RNNs. Flexibility and high accuracy are some key features of LSTM model which make it convenient to be used in various applications such as image captioning, time series forecasting and NLP etc (Ishfaque et al., 2022). The main motive of using this LSTM model is to overcome the evanescent gradient problem that plagues traditional RNNs. Below mentioned steps are the breakdown of the LSTM working model. Input: LSTM model receives an input vector (MFCC features).LSTM layer: This layer act as a cartridge belt with three stations – first one receives the current input data, next one applies weight

matrices, activation functions, and gates to transform the data and extract relevant information, and the last one generates the output for the current time step and updates the hidden state for the next step. Memory cell: It is the central component of LSTM layer responsible for storing and retrieving information over time. Gates: LSTM layer regulates the flow of information with the help of three gates- input gate, forget gate, output gate. This gating mechanism will decide how much amount of information can enter, exit and pass through the network. Input gate determine what new information to be stored in the cell state. Forget gate decide what information is to be discarded from previous cell state. Output gate selects what information from the updated cell state should be used as the output of the LSTM unit at the current time step. Cell state update: LSTM layer create some hidden layers which retains the data from previous states and combine it with current inputs to update the memory cell.

Output generation: This layer finally decides what information from updated memory cell act as output of current LSTM unit. Output is further sent to activation function to categorize the data. Activation functions: To deal with complex relationships and patterns in real world data activation functions such as tanh, sigmoid, ReLU is utilized. In this work Rectified linear unit is used to outputs the input value directly if positive and 0 otherwise. Back Propagation Through time (BPTT): Errors are calculated from actual and predicted values and propagated back through the network to improve the vanishing gradient problem of long sequences of data. Sequential Processing: The above said process continues till a satisfactory prediction of data is achieved. The computation mechanism is detailed in figure 3. Previous short term memory STM_{t-1} and current event vector CE_t are combined and multiplies with weight matrix W_n with bias which is passed to function tan h and finally give learn matrix N_t . For removing irrelevant information we introduced ignore factor i_t. To calculate i_t we combine STM_{t-1} and CE_t, take product with weight matrix, pass it through sigmoid function with same bias. This final learn gate output is fed to forgot gate with previous STM_{t-1} and CE_t to form a forget factor f_t and decide which information to forget and which to remember. Outcome of learn gate and forgot gate are combined to form remember gate output which will act as LTM for next cell. Use gate will add previous LTM and previous STM to create new STM which is the final output of current event and act as STM for the next cell.

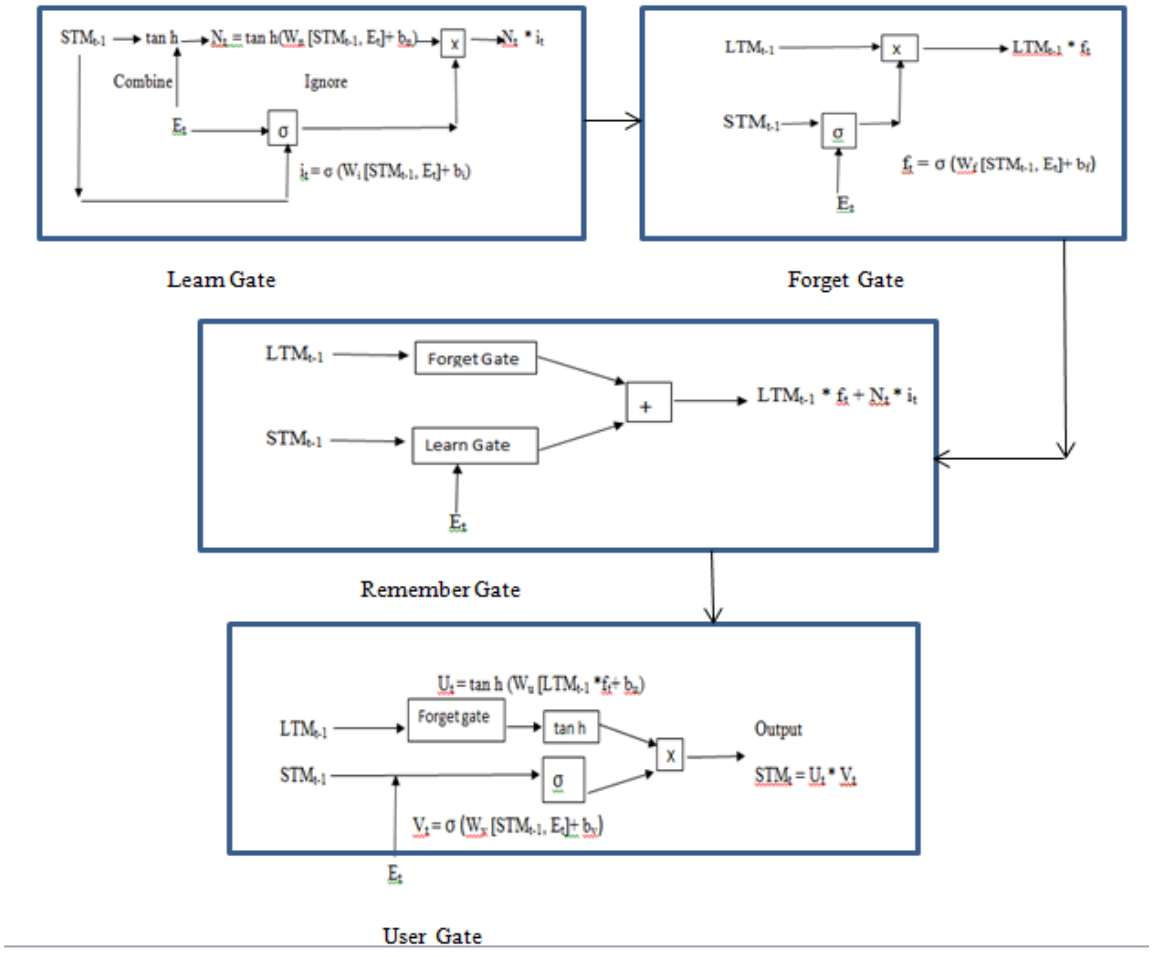

Figure 3. Basic gates used in LSTM Mechanism

4.4 Assessment Parameters

It is clear from the preliminary analysis covered in the above section that multiple assessment parameters are used when evaluating a deep learning model. To get better understanding of the model's weaknesses and strengths. Some basic parameters are used for the proposed model which are enlist in table 1.

Table 1. Basic assessment parameters.

Assessment of performance measures	**Formulae**
Precision	$P = \frac{TP}{TP + FP}$
Recall	$R = \frac{TP}{TP + FN}$
F-1 Score	$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$
Accuracy	$A = \frac{TP + TN}{TP + FN + TN + FP}$

4.5 Proposed Architecture

To preserve the sequence of data present in audio files we employ LSTM model. The temporal and spectral information which is present in audio files of data set will be converted into mel-frequency cepstral coefficients by applying Fourier transform. Input speech signal is divided into short windows of 3 seconds to obtain the mel scale frequency spectrum. Discrete cosine transform is applied on this spectrum to obtain MFCC vector. This array will act on input layer model. Two dense layers are used. For the purpose of classifying the output of LSTM layer into a set of predefined categories dense layers are used. Two dense layers are used to enhance the performance of proposed model. To prevent over fitting of LSTM network a regularization technique called dropout is used. Over fitting occurs when the model is unable to generalize the new data. Certain percentage of forget gate, input gate, output gate are randomly dropped out to avoid over fitting. This model utilized two dropout layers. The proposed LSTM model is shown in figure 4.

4.6 Pseudocode

The MFCC LSTM algorithm works by first extracting MFCC features from the audio signal. These features are then fed into the LSTM network, which learns to extract long-term temporal dependencies from the sequence of MFCC features. The output of the LSTM network can then be used for emotion recognition. The pseudo code for the proposed model is as shown in table 2.

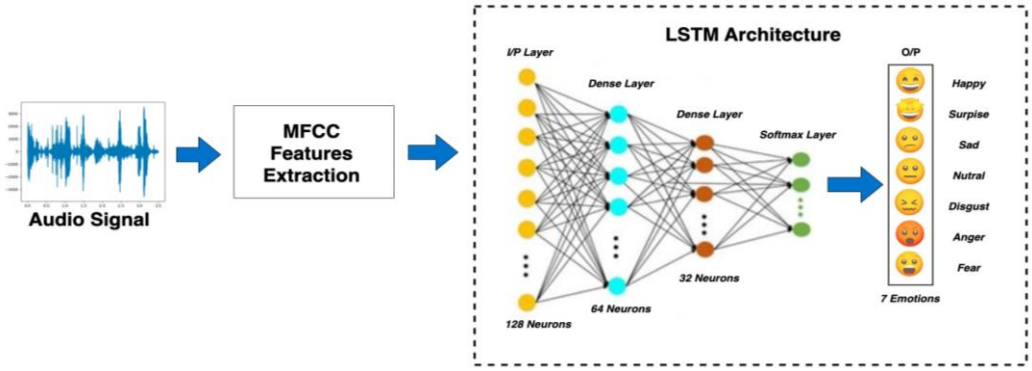

Figure 4. Proposed architecture for sentiment analysis

Table 2. Pseudo-Code for proposed model.

MFCC Feature extraction pseudo code
Input: audio signal , **Output :** MFCC ;
Function : MFCC parameters ;
Initialize parameters , Split into **frames** audio signal;
Apply **windowing** to frames ;
Use **Fast Fourier Transform** to get **spectrum** for all frames;
Find the matrix for a mel space filter bank and convert spectrum to **mel spectrum**;
Get **MFCC vector** for all frames by applying **discrete cosine transform** ;
Feed MFCC vector into **LSTM Network** ;
Initialize LSTM Cell : 4 layers,128 neurons, RELU;
Set hyper parameters: 5600 images,10 epochs, batch size-512 records;
Train LSTM Cell ;
Forward pass –input sequence, hidden state, cell state ;
Calculate loss;
Backward pass: backward pass loss ;
Update the **weights** ;
Save trained LSTM Cell – **Prediction on testing data**.

5. EXPERIMENTAL RESULTS

When evaluating an LSTM-based sentiment analysis model on spoken speech data from the Toronto dataset, the unique characteristics of audio data are considered as compared to text data. Load the spoken speech data from the Toronto dataset. Speech samples of two actresses in native English with seven primary emotions are taken. In the proposed work data set of four hundred recordings of each sentiment are taken which will further augmented to increase the accuracy in training phase. So in total eight hundred audio samples for each sentiment are used. 800*7=5600 images were extracted from the data set and act as input to the LSTM model. Convert the audio data into MFCCs (Mel-frequency cepstral coefficients). Converting audio-data into Mel-frequency cepstral coefficients (MFCCs) involves several steps. MFCCs are a representation of the short-term power spectrum of a sound signal and are commonly used in speech and audio processing tasks. Install the Python and the Librosa library. Load audio data using librosa library as this library makes it easy to work with audio data in Python. Pre-process the audio data as needed. Compute MFCCs. Use Librosa to compute the MFCCs of the audio data. MFCC is used for feature extraction at offset of 0.5 second. There are 40 MFCC coefficients. The shape of single data point is [40, 1]. In one batch there is 512 data points. One data point contain 40 coefficients. These MFCCs are used as input features for the proposed LSTM-based sentiment analysis model. Split the data into training and testing sets. Design an LSTM-based model that can handle sequential data. 1D convolutional layers are included before the LSTM layers to capture temporal patterns in the audio data. Labels are converted into numbers by encoders which are further converted into an array. Using tensor flow we apply it on LSTM architecture. Input layer of LSTM model contain 128 neurons. Next layer is dense layer where Relu activation function is applied. There is a dropout of 30% so 2nd layer contains 64 neurons. Next dense layer contain 32 neurons. We apply softmax function on last layer and classify 7 emotions. Values of different parameters used in suggested work is described in table 3. Then, a loss function for classification tasks, such as categorical cross-entropy is computed. Finally, the model is compiled with an appropriate optimizer and evaluation metric. Train the model on the training set. Monitor the training process by observing the loss and accuracy on the validation set. Then hyper-parameter tuning is done such as the number of LSTM units, learning rate, and batch size. Adjust the architecture based on the complexity and size of the spoken speech dataset. The model is evaluated on the test set using appropriate audio-specific metrics. Common metrics for audio classification include accuracy, precision, recall, F1 score, and confusion matrix.

Table 3. Various parameters used in proposed work.

Suggested Methods	Value
LSTM images	5600
No. of epochs	10 epoch
Batch size	512 records
Activation Function	RELU
Optimizer	ADAM
Dropout	30%
Loss Function	Categorical Cross Entropy
Activation output	Softmax

	precision	recall	f1-score	support
0	0.94	0.98	0.96	800
1	0.92	0.97	0.94	800
2	1.00	0.96	0.98	800
3	0.90	0.95	0.93	800
4	1.00	1.00	1.00	800
5	0.98	0.85	0.91	800
6	0.97	0.98	0.97	800
accuracy			0.96	5600
macro avg	0.96	0.96	0.96	5600
weighted avg	0.96	0.96	0.96	5600

Figure 5. Training performance of the proposed model

Accuracy is a good metric for evaluating a model's overall performance but it can mislead if the data set is imbalance. Precision and recall are two other assessment parameters that can be used to evaluate a models performance on imbalanced dataset. Precision and recall are often used to calculate F1 score, which is a single metric that combines both recall and precision. Figure 5 shows the values of accuracy, precision, recall and F1 score and figure 6 illustrate the accuracy graph of the model. Loss values are used to update models parameter during training. By minimizing the loss, the model can be trained to produce more accurate predictions. Figure 7 gives the losses detail of the model. For identifying the areas where the model performance can be improved, confusion matrix is used. Confusion matrix is used in addition to assessment parameters because they provide more detailed information about performance of model. It is helpful in calculating other metrics such as specificity and negative predictive values to increase the accuracy of the model. Figure 8 shows the confusion matrix for the proposed model.

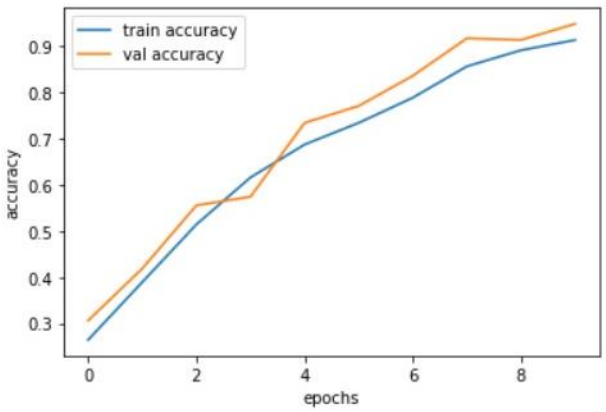

Figure 6. Accuracy of the proposed work

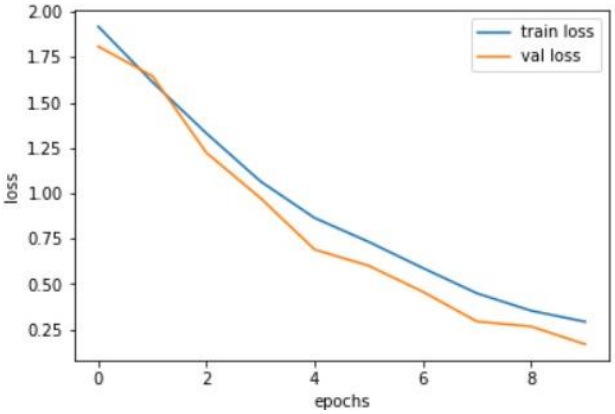

Figure 7. Loss value of the model

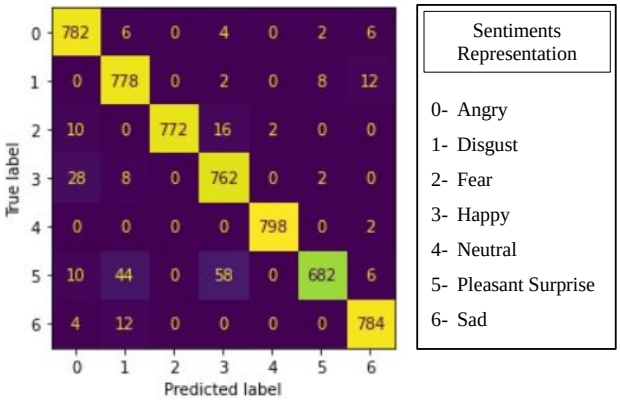

Figure 8. Confusion matrix for seven sentiments

Table 4. Comparison of proposed work model with previous models.

Reference	Model Used	No. of Emotions	Recognized Emotions	Accuracy
(Wang et al., 2020)	MFCC + LSTM +dual sequence LSTM	4	happy, neutral, anger, sad	73.3%
(Chen et al., 2022)	MFCC + dual channel LSTM	6	happy ,sad, anger, fear ,surprise, disgust	77.9 %
(Vyas et al., 2015)	MFCC + HMM	4	surprise ,sad ,fear, disgust	78%
(Gilke et al., 2012).	MFCC + ANN	5	happy anger, surprise, sad neutral	83.5 %
(Dahake et al., 2016)	MFCC+ SVM	6	anger, fear, joy, sadness, boredom, neutral	84 %
(Gupta et al., 2021).	MFCC+ LSTM	7	calm, happy, sad, angry, fearful, surprise, disgust	84.81 %
(Xie et al., 2019)	MFCC+ LSTM	6	happy ,sad, anger, fear ,surprise, disgust	89.6 %
(Sundarprasad, 2018)	MFCC + SVM	7	happy ,sad, anger, neutral, fear ,surprise, disgust	90 %
(Dan et al., 2013)	MFCC+ MEDC + Energy	4	joy, fury, sadness, neutral	91.30%
(Badr et al., 2021).	MFCC+ Conv LSTM	8	happy ,sad, anger, neutral, fear ,surprise, disgust, calm	91%
Proposed work	MFCC+ LSTM	7	happy ,sad, anger, fear , pleasant surprise, disgust, neutral	96

6. CONCLUSION

We are currently working on a deep learning system that can detect emotions in speech. To achieve this, we're utilizing a Long Short-Term Memory (LSTM) that can analyze speech signals and extract relevant features by using Mel Frequency Cepstral Coefficient (MFCC) to identify various emotional patterns. We believe that our system's success demonstrates the potential of deep learning techniques in analyzing complex speech signals and improving emotion recognition accuracy. Our work could have significant implications across numerous domains and applications. This paper proposes a hybrid model for SER by combining the MFCC and LSTM architectures. The strengths of both architectures are adapted to improve the recognition performance in the SER. Our hybrid model performed better at learning the long-term dependencies in speech signals by preserving the hidden state of input features using LSTM and the use of MFCC feature vectors. The proposed hybrid model is able to learn the temporal information from the frequency distributions in the MFCCs of each emotion in the dataset. The hybrid model's effectiveness is shown through the results obtained from the experiments in this research. The recognition rate of the proposed model is 96%. In the future, an improvement to the pre-processing method may be carried out, especially on the language-independent dataset. Data augmentation can be applied to overcome the problem of data shortage in training and testing datasets. Datasets from other languages can be added to improve the language-independent emotion recognition ability. The proposed hybrid model on cross-corpus speech emotion recognition should also be investigated.

References:

Badr, Y., Mukherjee, P., & Thumati, S. M. (2021). Speech Emotion Recognition using MFCC and Hybrid Neural Networks. In *IJCCI* (pp. 366-373). doi: 10.5220/0010707400003063 .

Chan, J. Y. L., Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2023). State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, *56*(1), 749-780.

Chen, W. (2022). A novel long short-term memory network model for multimodal music emotion analysis in affective computing. *Journal of Applied Science and Engineering*, *26*(3), 367-376. doi: 10.6180/jase.202303_26(3).0008. https://doi.org/10.6180/jase.202303_26(3).0008

Cui, J., Wang, Z., Ho, S. B., & Cambria, E. (2023). Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, *56*(8), 8469-8510.

Dahake, P. P., Shaw, K., & Malathi, P. (2016, September). Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)* (pp. 1080-1084). IEEE.

Dan, Z. M., & Monica, F. S. (2013, October). A study about MFCC relevance in emotion classification for SRoL database. In *2013 4th International Symposium on Electrical and Electronics Engineering (ISEEE)* (pp. 1-4). IEEE.doi: 10.1109/ISEEE.2013.6674323.

Daneshfar, F., Kabudian, S. J., & Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier. *Applied Acoustics*, *166*, 107360.

Deshwal, D., Sangwan, P., & Kumar, D. (2019). Feature extraction methods in language identification: a survey. *Wireless Personal Communications*, *107*, 2071-2103.

Deshwal, D., Sangwan, P., & Kumar, D. (2020, November). A Structured Approach towards Robust Database Collection for Language Identification. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-6). IEEE. doi: 10.1109/ACIT50332.2020.9299963.

Deshwal, D., Sangwan, P., Dahiya, N., Lilhore, U. K., Dalal, S., & Simaiya, S. (2023). COVID-19 Detection using Hybrid CNN-RNN Architecture with Transfer Learning from X-Rays. *Current medical imaging*. doi: 10.2174/1573405620666230817092337.

Devi, J. S., Srinivas, Y., & Nandyala, S. P. (2014). Automatic speech emotion and speaker recognition based on hybrid gmm and ffbnn. *International Journal on Computational Sciences & Applications (IJCSA)*, *4*(1), 35-42.doi: 10.5121/ijcsa.2014.4104.

Gilke, M., Kachare, P., Kothalikar, R., Rodrigues, V. P., & Pednekar, M. (2012). MFCC-based vocal emotion recognition using ANN. In International Conference on Electronics Engineering and Informatics (ICEEI 2012) IPCSIT (Vol. 49).

Gupta, M., & Chandra, S. (2021, August). Speech Emotion Recognition Using MFCC and Wide Residual Network. In 2021 *Thirteenth International Conference on Contemporary Computing* (IC3-2021) (pp. 320-327). doi: 10.1145/3474124.3474171.

Hazra, S. K., Ema, R. R., Galib, S. M., Kabir, S., & Adnan, N. (2022). Emotion recognition of human speech using deep learning method and MFCC features. *Radioelectronic and Computer Systems*, (4), 161-172. doi: 10.32620/reks.2022.4.13.

Hong, I. S., Ko, Y. J., Shin, H. S., & Kim, Y. J. (2016, July). Emotion recognition from Korean language using MFCC HMM and speech speed. In *The 12th International Conference on Multimedia Information Technology and Applications (MITA2016)* (pp. 12-15).

Ishfaque, M., Dai, Q., Haq, N. U., Jadoon, K., Shahzad, S. M., & Janjuhah, H. T. (2022). Use of recurrent neural network with long short-term memory for seepage prediction at Tarbela Dam, KP, Pakistan. *Energies*, *15*(9), 3123. doi: 10.3390/en15093123

Kacur, J., Puterka, B., Pavlovicova, J., & Oravec, M. (2021). On the speech properties and feature extraction methods in speech emotion recognition. *Sensors*, *21*(5), *1888*.doi: 10.1007/s10772-020-09672-4.

Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, 23(1), 45-55.

Krishnan, P. T., Joseph Raj, A. N., & Rajangam, V. (2021). Emotion classification from speech signal based on empirical mode decomposition and non-linear features: *Speech emotion recognition. Complex & Intelligent Systems*, *7*, 1919-1934. doi: 10.1007/s40747-021-00295-z.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

Neiberg, D., Elenius, K., & Laskowski, K.(2006).Emotion Recognition in Spontaneous Speech Using GMMs Department of Speech , Music and Hearing , KTH , Stockholm , Sweden Classifiers ReCALL, pp. 809–812, 2006. http://dx.doi.org/10.21437/Interspeech.2006-277

Sandhya, P., Spoorthy, V., Koolagudi, S. G., & Sobhana, N. V. (2020, December). Spectral features for emotional speaker recognition. In *2020 third international conference on advances in electronics, computers and communications (ICAECC)* (pp. 1-6). IEEE.

Sangwan, P., Deshwal, D., & Dahiya, N. (2021). Performance of a language identification system using hybrid features and ANN learning algorithms. *Applied Acoustics*, *175*, 107815. doi: 10.1016/j.apacoust.2020.107815.

Sangwan, P., Deshwal, D., Kumar, D., & Bhardwaj, S. (2023). Isolated word language identification system with hybrid features from a deep belief network. *International Journal of Communication Systems*, *36*(12), e4418.

Song, P., Zheng, W., Liu, J., Li, J., & Zhang, X. (2015). A novel speech emotion recognition method via transfer PCA and sparse coding. In *Biometric Recognition: 10th Chinese Conference*, CCBR 2015, Tianjin, China, November 13-15, 2015, Proceedings 10 (pp. 393-400). Springer International Publishing. doi: 10.1007/978-3-319-25417-3.

Sundarprasad, N. (2018). Speech emotion detection using machine learning techniques. https://doi.org/10.31979/etd.a5c2-v7e2

Taherdoost, H., & Madanchian, M. (2023). Artificial intelligence and sentiment analysis: A review in competitive research. *Computers*, *12*(2), 37.

Vyas, G., Dutta, M. K., Riha, K., & Prinosil, J. (2015, October). An automatic emotion recognizer using MFCCs and Hidden Markov Models. In *2015 7th International congress on ultra modern telecommunications and control systems and workshops* (ICUMT) (pp. 320-324). IEEE.

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., & Tarokh, V. (2020, May). Speech emotion recognition with dual-sequence LSTM architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6474-6478). IEEE.

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(11), 1675-1685. https://doi.org/10.1109/TASLP.2019.2925934

Suman Lata
Maharaja Agrasen University,
H.P,
India
suman@msit.in
ORCID 0000-0002-5339-2584

Neha Kishore
Maharaja Agrasen University,
H.P,
India
nehakishore.garg@gmail.com
ORCID 0000-0002-6591-3084

Pardeep Sangwan
Maharaja Surajmal Institute of Technology,
New Delhi,
India
sangwanpardeep@msit.in
ORCID 0000-0002-7301-6607