2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) | 979-8-3503-1860-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCECE60827.2024.10548174

Speech Emotion Recognition Using LSTM

1st **Khushi J Shetty**
Department Of MCA,
NMAM Institute Of Technology,
NITTE (Deemed to be University),
Karnataka, INDIA
khushijshetty020@gmail.com

2nd **Spoorthi Shetty**
Department of MCA,
NMAM Institute Of Technology,
NITTE (Deemed to be University),
Karnataka, INDIA
sshetty.07@nitte.edu.in

3rd **Manish**
Department of MCA,
NMAM Institute Of Technology,
NITTE (Deemed to be University),
Karnataka, INDIA
iammsalian@gmail.com

4th **Mangala Shetty**
Department Of MCA,
NMAM Institute Of Technology,
Karnataka, INDIA
mangalapshetty@nitte.edu.in

Abstract—**Emotion recognition from speech is a crucial task in many applications, such as affective computing, psychological research, and human-computer interaction. LSTM networks shows promise in modeling sequential data, such as voice signals, because of its capacity to capture long-term dependencies. The main objective of this research work is to assign emotions to one of the four categories: fear, rage, sadness, and happiness. The samples are used to determine energy, pitch, MFCC and LPCC coefficients, speaker rate, and other important characteristics. Speech is a widely utilized signal for human-to-human communication, which implies that speech is also used for communication between people and the machines. This interactive system's aim is to improve the speech emotion recognition (SER) technology by using LSTM Model. Speech Emotion Recognition (SER) is critically important for applications that assess spoken emotions in real-time. Dataset used in this Research work are taken from kaggle which is named as Toronto emotional speech set (TESS). The dataset is arranged according to the emotion expressed by the two female actors, which is organized within a separate folder, And the audio clips are arranged with 200 targeted words. The audio file is saved in WAV format. Additionally, a system utilizing an LSTM algorithm and an MFCC features are presented in this paper's preliminary results.**

Keywords—- LSTM, MFCC, SER, WavePlot, Spectrogram

I. Introduction

The goal of developing field in human computer intelligence research is to enable computer's to learn from their experiences and make decisions about how to react in certain scenarios. As a result, there is now better communication between user's and the computer. It is possible to train the computer to recognize different traits in the speech samples and infer the underlying emotion. By presenting an efficient ensemble model that combines 1D-CNN, LSTM, and GRU architectures with data augmentation techniques, the research advances the state-of-the-art in SER[1]. The introduction to the complexities on how emotions are expressed in speech and the difficulties found in identifying emotions from audio data. The study advances the field of emotional speech identification by putting out and assessing a deep neural network-based method. It sheds light on the viability and efficiency of employing DNNs to extract emotional content from voice signals[2]. Overview of deep learning techniques and how they can help emotion identification systems more reliable and accurate. An overview of the paper's organization, including a description of its sections and how each one advances the field of speech emotion recognition utilizing LSTM models. Many methods and classifiers, such as KNN, Artificial Neural Networks, Hidden Markov Models, Support Vector Machines, and others, can be used to categorize human emotions based on training datasets. The first thing that must be done in order to detect emotions is feature extraction. By this novel strategy that blends Bi-directional LSTM architecture and Deep Belief Networks, the article advances speech emotion recognition. It sheds light on the viability and efficiency of utilizing these architectures to extract emotional content from voice signals[3]. For this project, MFCC has been utilized. For emotion recognition, As a classifier, the Support Vector Machine is used. Regression and classification are two uses for the SVM. In order to categorize the data, Utilizing the spectrogram characteristics and incorporating neural network techniques, such as convolutional neural networks (CNNs), the efforts on SER has achieved great progress. By examining the efficacy of CNNs for feature extraction from speech signals, the research advances the field of SER. It sheds light on the viability and efficiency of utilizing CNN-based techniques to extract emotional content from speech signals[4]. The data is classified using a linear or nonlinear separation surface in the input feature space of the dataset. The basic idea is to convert the original input set into a higher dimensional feature space by using a kernel function, and then optimize classification in the new feature space, Audio Data's are used by the Speech Emotion Recognition System. It uses a portion of the voice as input to identify the mood of the speaker by expressing different emotions which can be recognized, such as happiness, sadness, surprise, anger, etc. Previous studies in this field have looked at architecture design, training methods, and performance assessment of CNN-LSTM based models for face emotion recognition. Numerous studies have looked into

979-8-3503-1860-9/24/$31.00 ©2024 IEEE

various CNN architectures for feature extraction, including specially built CNNs for the job of face emotion recognition or pre-trained models like VGG and ResNet. Researchers have also looked into other approaches to include temporal information in the model, like capturing motion cues in films using optical flow-based representations or 3D convolutions. This project helps to determine the emotions of customers when they are on the call with a call center. Currently, the majority of work in this field classifies emotions into several groups by extracting discriminatory features. The main objective is to access the speech content and offer details regarding the voice, pitch, intensity, and spectral modulation in the supplied audio files. The system can only support English as the language for real-time voice data input. Hospitals could utilize speech emotion recognition as a feedback tool to classify patient remarks into relevant groups, such as feelings of anger or grief. SER will benefit applications in human-computer interaction, healthcare, and other fields as technology develops.

II. Literature Survey

The ensemble 1D-CNN-LSTM-GRU model with data augmentation was proposed by Ahmed et al. (2023)[1] for speech emotion recognition. Their research expands on the combination of Gated Recurrent Unit (GRU), long short-term memory (LSTM), and convolutional networks. Using the advantages of each architecture, this method efficiently extracts features and classifies emotions from speech data.

Van et al. (2022) investigated the use of deep neural networks for emotional voice identification. Their research explores the use of deep learning skills to accurately classify emotions from speech data[2]. Through the use of neural network topologies, they seek to extract complex patterns from voice signals that correspond to various emotional states. This study advances the field of emotion recognition technology and has potential implications in human-computer interaction and affective computing.

Senthilkumar et al. (2022) used Deep Belief Networks (DBN) in conjunction with bi-directional long short-term memory (LSTM) architecture to study voice emotion recognition[3]. Their method uses LSTM to take advantage of the temporal relationships in speech data and DBNs' feature learning power. They want to improve the resilience and accuracy of emotion categorization systems by combining these methods, which will further develop emotional computing and related domains. Convolutional neural networks (CNNs) were used in the 2021 IEEE 10th Global Conference on Consumer Electronics to study feature extraction for speech emotion recognition[4]. In an effort to improve the accuracy of emotion recognition, the study examined CNN-based techniques for extracting discriminative features from speech data. The field of emotional computing and allied fields are benefited from this study[5]. It is Published in 2023 in IEEE Access, the study extensively validates speech emotion recognition using Convolutional Neural Networks (CNNs) and deep learning techniques, contributing to advancements in affective computing. A paper demonstrating an emotion recognition system using a two-level ensemble of deep-convolutional neural network models was published in IEEE Access in 2023 [6]. This work advances the accuracy of emotion recognition in deep learning frameworks by using ensemble learning approaches.

A speech emotion recognition technique using recurrent neural networks with directional self-attention was presented by Li et al. in 2021. [7] Their work advances the field of emotional computing research by utilizing recurrent networks and attention mechanism to improve the accuracy of emotion classification.

Zhang et al. (2021) introduced a novel heterogeneous parallel convolution Bi-LSTM architecture for speech emotion recognition.[8] This research addresses the need for efficient feature extraction and sequential modeling, contributing to advancements in emotion recognition systems using deep learning techniques.

The research, is been presented by the authors[9], the paper focuses on speech emotion recognition using enhanced features and deep learning approaches. By combining augmented characteristics with deep learning models, this research improves the accuracy of emotion categorization and advances the field of emotional computing.

Arnold Sachith A Hans and Smitha Rao (2021) [10] proposed a CNN-LSTM-based deep neural network for facial emotion detection in videos. Their study contributes to improve the emotion recognition accuracy in video data by leveraging the capabilities of convolutional and recurrent neural networks.

Wani et al. (2021) carried out an extensive analysis of speech emotion detection systems[11]. Their research adds to our understanding of emotion recognition technology and suggests avenues for future development by shedding light on current methods, problems, and developments in that field.

Andayani et al. (2022)[12] introduced a hybrid LSTM-Transformer model for emotion recognition from speech audio files. This study amalgamates the strengths of LSTM and Transformer architectures to enhance emotion classification accuracy. Their research contributes to advancements in affective computing by leveraging deep learning techniques tailored by speech emotion recognition tasks.

Roy et al. (2021) used deep learning techniques to study the recognition of emotions in speech. Their research shows how to effectively recognize emotions from speech data using deep learning models. [13] This work advances the field of affective computing by providing new information on how well deep learning techniques perform when it comes to speech emotion recognition tasks.

Mahmoudi and Bouami (2023) [14] presented a study on deep neural network-based Arabic speech emotion recognition. Their study advances our knowledge of Arabic voice emotion recognition by examining the usefulness of deep learning techniques in this domain. Applications on affective computing geared at Arabic-speaking populations who will find value in this work.[15]A review of speech emotion recognition is continuously provided by De Lope and Graña (2023). The objective of their work is to provide an overview of the field's recent developments, difficulties, and new trends. This review helps researchers and practitioners gain a better grasp of the state-of-the-art in speech emotion recognition by providing insightful information.

III. Methodology

Speech Emotion Recognition uses the input voice samples to determine the speaker's emotions. This is accomplished by doing a thorough process and analyzing the voice signal input. The speech signal is first obtained, and then features which include voice patterns, emotional content, and coefficients are extracted using MFCC. These features are then sent into the classification system for additional analysis. LSTM architecture is the model design for SER. Tuning parameters and the training process. The way it is designed, the training model remains unchanged and the vanishing gradient issue is virtually eliminated. A more sophisticated form of recurrent neural network (RNN) architecture called long short term memory is intended to reflect chronological sequences and their long-range relationships more accurately than regular RNNs. Sequential neural networks with deep learning capabilities and information preservation are known as long short-term memory networks.

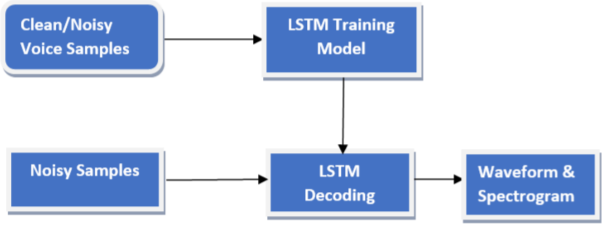
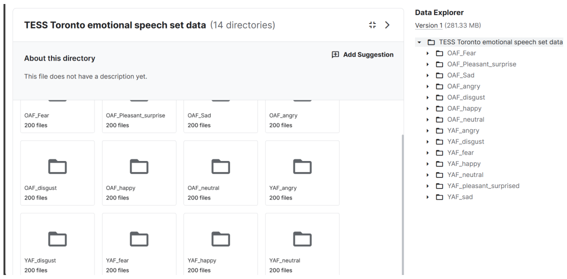


Fig. 1. **Architecture Of LSTM Model**

Initially the noisy voice samples are processed to the LSTM model by training the model. Then the decoding of noisy samples takes place where the recordings are classified based on the emotion and categorised into waveplots and spectrograms which are shown in Fig1.

A. *Data Collection*

Assemble a selection of speech recordings with the corresponding emotional labels applied. Ensure that the dataset encompasses a diverse range of emotions in addition to variations in the accents and environment of the speaker. The explanation of features, size, and origin of the dataset that were used for evaluation and training. A description of the emotional categories in the dataset and their distribution are showed in Fig2.



TESS Toronto emotional speech set data (14 directories)

About this directory

This file does not have a description yet.

OAF_Fear 200 files
OAF_Pleasant_surprise 200 files
OAF_Sad 200 files
OAF_angry 200 files
OAF_disgust 200 files
OAF_happy 200 files
OAF_neutral 200 files
YAF_angry 200 files
YAF_disgust 200 files
YAF_fear 200 files
YAF_happy 200 files
YAF_neutral 200 files

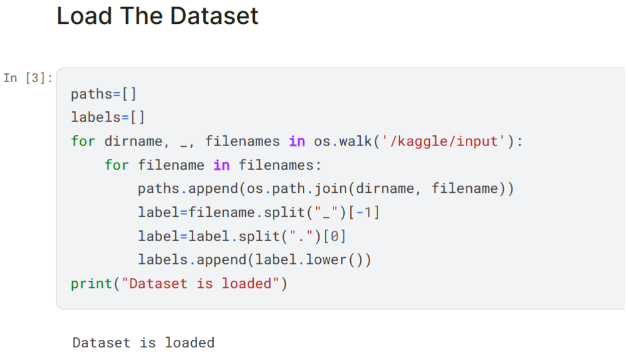
Data Explorer
Version 1 (281.33 MB)

TESS Toronto emotional speech set data
OAF_Fear
OAF_Pleasant_surprise
OAF_Sad
OAF_angry
OAF_disgust
OAF_happy
OAF_neutral
YAF_angry
YAF_disgust
YAF_fear
YAF_happy
YAF_neutral
YAF_pleasant_surprised
YAF_sad

Fig. 2. **Toronto emotional speech set (TESS)**

B. *Data Preprocessing*

Using spectrograms, MFCCs (Mel-Frequency Cepstral Co efficients), or other acoustic features, extract pertinent features from the voice data. Standardize and normalize the features to improve the training of the model is shown in Fig3. Segmentation, normalization, and audio cleaning are some of the data preprocessing techniques that are covered.



Load The Dataset

```
In [3]:
paths=[]
labels=[]
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        paths.append(os.path.join(dirname, filename))
        label=filename.split("_")[-1]
        label=label.split(".")[0]
        labels.append(label.lower())
print("Dataset is loaded")

Dataset is loaded
```

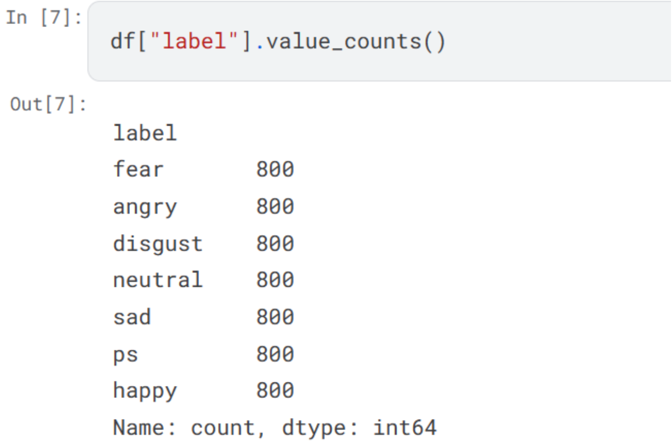
Fig. 3. **Preprocessing of the dataset**

C. *Data Spliting*

This step accomplishes training, validation, and test models out of the dataset. To prevent prejudice, make sure each set depicts a fair distribution of feelings as shown in Fig4. LSTM network efficiently divides the SER data, allowing model to train, validate, and assess the model for speech-based emotion identification applications.

D. *Model Architecture Design*

Initially the input layer is set up which is mainly used to Receive the preprocessed audio features, theese features are extracted to capture the temporal dependencies. LSTM Layers will Stack one or more LSTM layers to capture temporal dependencies. The number of units in each LSTM layer can be adjusted based on the complexity of the problem. Optionally the dropout layers are used To prevent overfitting. The last layer is output layer which is a fully connected layer with softmax activation to output probabilities for each emotion class.



```
In [7]:  df["label"].value_counts()

Out[7]:
         label
         fear       800
         angry      800
         disgust    800
         neutral    800
         sad        800
         ps         800
         happy      800
         Name: count, dtype: int64
```

Fig. 4. **Spliting data into different emotions**

E. Model Training

The trained LSTM model can be deployed for real-world applications, where it can classify the emotional content of speech signals in various contexts, such as customer service interactions, sentiment analysis in social media, or mental health monitoring. Continuous monitoring and retraining of the model ensure its effectiveness in capturing and recognizing nuanced emotional cues in speech.

F. Integration and Deployment

Include the trained LSTM model in the intended system or application. Install the SER system at the appropriate setting,taking platform compatibility and scalability into account.

IV. LSTM ALGORITHM

LSTM, is a type of Recurrent Neural Network (RNN) architecture that excels at processing sequence data. LSTMs are used in Speech Emotion Recognition (SER) to identify and interpret emotional patterns in spoken language. The intricate patterns and temporal relationships seen in voice data makes long short term memory networks as an excellent ally for context-aware and sophisticated emotion recognition. Comprehensive explanation of the LSTM architecture, encompasses the design and functionality of LSTM cells. The explanation of how LSTM layers update their internal states in response to processing input sequences over time steps. The function of hidden layers in acquiring abstract features from the input data and learning higher-level representations is explained. The explanation of how regularization approaches and contributes to the LSTM model's increased stability and robustness during training. Recurrent neural networks (RNNs) of the LSTM type are designed to extract long-term dependencies from sequential input. A novel technique to speech recognition is proposed, which blends attention-based long short-term memory (LSTM) recurrent neural networks with frame-level speech parameters. To sum up, long-term dependencies with sequential data can be learned by using LSTM, a specific kind of RNN that includes memory cells and gates. This design has shown to be successful in number of situations where temporal pattern recognition and context awareness are essential. LSTMs are ideally suited for applications like time series prediction, machine translation, and speech emotion identification because of their capacity to selectively retain and forget information over time. TESS(Toronto Emotional Speech Set) includes about 2,800 samples of 200 isolated words

Formula Used For LSTM Calculation:-

$$C_t = C_{t-1} \cdot f_t \tag{1}$$

The previous memory state is fully erased if the forget gate value is 0.

The prior memory state is fully transferred to the cell if the forget gate value is 1. (Remember f gate yields values between 0 and 1)

Now with current memory state Ct now calculate new memory state from input state and C layer.

$$C_t = C_{t-1} + (I_t \cdot \tilde{C}_t) \tag{2}$$

Ct = Current memory state at time step t. and it gets passed to next time step.

V. RESULT

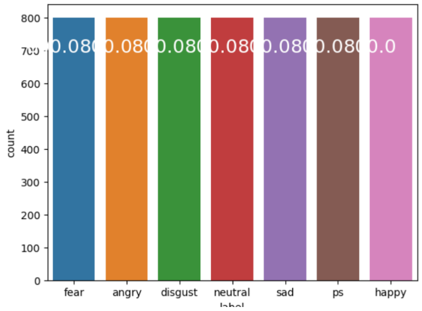


Fig. 5. **Exploratory Data Analysis**

It shows a great deal about the speech's dataset properties and the connection between speech features and emotions by performing exploratory data analysis which is shown in Fig 5. This knowledge helps to create emotion detection models which are more accurate.

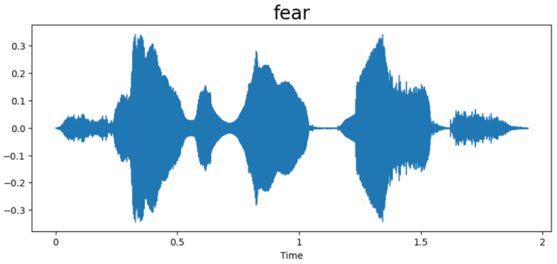


Fig. 6. **Sound Waveplot of emotion FEAR**

Fig 6 shows the Waveplots offering valuable insights into the temporal dynamics of speech signals and can aid in the analysis and interpretation of speech emotion recognition data.

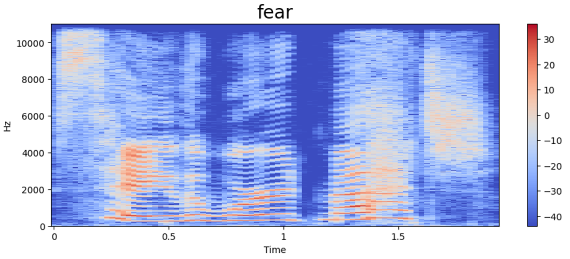


Fig. 7. **Spectrum View Of emotion FEAR**

Through the examination of spectrogram, scientists can learn more about the spectrum properties of voice signals linked to various emotional states. The more efficient speech emotion recognition algorithms can be developed with the help of these discoveries which is shown in Fig 7.

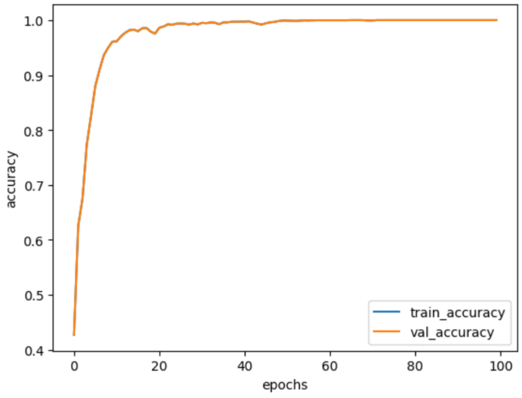


Fig. 8. **Accuracy Obtained**

By plotting accuracy, Fig 8 can visualize the performance of speech emotion recognition system and gain insights into its effectiveness in classifying emotions from speech data. This work has obtained 96% percentage of accuracy.

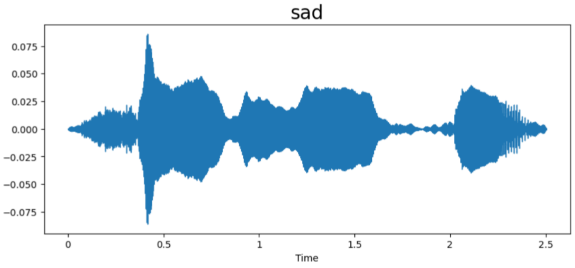


Fig. 9. **Waveplot Of Emotion ”SAD”**

Waveplots and spectrograms of an audio file from each class is plotted. A sample of each class's emotion speech on audio is shown in Fig 9. Darker hues are used in lower-pitched voices and vibrant hues are used by higher pitched Voices

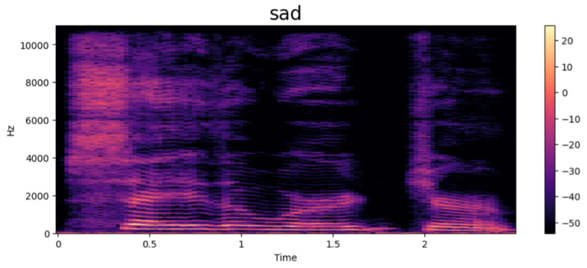


Fig. 10. **Spectrogram View Of Emotion ”SAD”**

The spectrogram in Fig 10 represents the intensity of different frequencies present in the speech signal over time. Darker regions indicate higher intensity at specific frequencies. The spectrogram obtained for emotion 'SAD' is shown in the above figure.

TABLE I
EMOTION CLASSIFICATION REPORT

Class	**Precision**	**Recall**	**F1-Score**	**Support**
angry	0.96	0.97	0.97	1484
disgust	0.97	0.95	0.96	1558
fear	0.96	0.97	0.96	1505
happy	0.96	0.95	0.96	1619
neutral	0.97	0.98	0.97	1558
sad	0.96	0.97	0.96	1478
surprise	0.98	0.97	0.97	528
Accuracy	0.96			
Macro Avg	0.96			
Weighted Avg	0.96			

Table 1 provides a comprehensive overview of the model's performance across different emotion classes and overall accuracy.

VI. CONCLUSION

Utilizing recurrent neural networks and long short-term memory is the study's primary goal to recognize an individual's emotion, which has been achieved. A dataset of two thousand eight hundred sample files that comprises feelings such as fear, disgust, surprise, calm, joyful, sad, and neutral are evaluated. Numerical numbers are frequently used as input for machine learning models, before using this data to extract features, we have to first turn it into arrays. The librosa package is been used to extract the MFCC feature, which is utilized by this model. This model is yielded with the overall accuracy of 97 percentage. The nature of emotion expression in speech and the difficulties found in extracting emotions from audio data, are revealed by the study. The LSTM model's performance is examined in relation to how it illuminates the fundamental processes of emotion processing and recognition in the human brain. Considerations for model deployment, monitoring, and maintenance are included in these useful advice for implementing LSTM-based SER systems in practical applications. Recommendations for incorporating emotion detection technologies into current platforms or systems to improve interaction results and user experiences by Accessing the usability and efficacy of LSTM-based SER systems in practical contexts which requires careful consideration of user input and validation research. Thank you for giving me the opportunity to support the research community's joint efforts in Speech Emotion Recognition. Explaining about how to make emotion recognition interfaces and apps more user-friendly and efficient by giving user needs, preferences, and feedback as a top priority. The vision statement delineates the enduring objectives and ambitions for LSTM-based SER technology, encompassing its capacity to transform emotional communication and human-computer interaction. A consideration on how emotion detection technology has revolutionized society and how practitioners and scholars might influence its future course, addressing issues including unresolved problems, user-centered design, flexibility, ongoing education, chances for collaboration, long-term goals, and requests for input and cooperation. All the subjects are combined to form a future-focused conclusion that highlights the significance and promises the LSTM-based Speech Emotion Recognition technology to influence emotional communication and human-computer interaction. The main motive of this paper is to recognise voice notes or pre recordings based on the human emotion or by the tone and pitch of the voice recordings.

VII. Future Scope

By using a digital signal processor, the model's real-time speaker detection mechanism is greatly enhanced. It is possible to build sound systems that adjust to ambient noise levels. This technique can also be utilized in voice-based chatbots and virtual assistants, as well as in complaining contact centers. Studying techniques are used to combine voice signals with other modalities, like text, physiological signs, and facial expressions, in order to improve the resilience and accuracy of emotion recognition. Research is been done on domain adaptation techniques that allows information to be transferred from labeled source domains to the targeted domains with fewer labeled data. Research into low-latency and real-time processing methods for LSTM based SER systems is used to support applications that need quick reaction times, like interactive games, real-time feedback systems, and emotion-aware assistive technologies. Investigation of hardware acceleration techniques, effective algorithms, and lightweight model architectures is used to maximize inference speed and minimize computing overhead. Exploration of community-driven initiatives, such as challenges, workshops, and competitions, are used to support Poster collaboration, comparison, and advancement in the field of research.

References

[1] An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition by Md. Rayhan Ahmed 1, Salekul Islam 2, A.K.M. Muzahidul Islam 3, Swakkhar Shatabda 15 May 2023

[2] Emotional Speech Recognition Using Deep Neural Networks by Loan Trinh Van 1ORCID,Thuy Dao Thi Le 2,Thanh Le Xuan 1,ORCID andEric Castelli 12 February 2022

[3] Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks Author links open overlay panelN. Senthilkumar a, S. Karpakam b, M. Gayathri Devi c, R. Balakumaresan d, P. Dhilipkumar in 20 April 2022.

[4] Analysis of Feature Extraction by Convolutional Neural Network for Speech Emotion Recognition 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE) Published: 2021

[5] Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks IEEE Access Published: 2023

[6] Emotion Recognition System Based on Two-Level Ensemble of Deep-Convolutional Neural Network Models IEEE Access Published: 2023

[7] Speech emotion recognition using recurrent neural networks with directional self-attention Dongdong Li a b, Jinlin Liu a, Zhuo Yang a, Linyu Sun a, Zhe Wang 1 July 2021

[8] A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition by Huiyun Zhang 1,2,Heming Huang 1,2, andHenry Han 22 October 2021

[9] Speech Emotion Recognition Using Deep Learning Techniques and Augmented Features 2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) Published: 2023

[10] A CNN-LSTM BASED DEEP NEURAL NETWORKS FOR FACIAL EMOTION DETECTION IN VIDEOS Arnold Sachith A Hans Smitha Rao 2021-06-30

[11] A Comprehensive Review of Speech Emotion Recognition Systems by Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, Eliathamby Ambikairajah in 22 March 2021.

[12] Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files by Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun, Caslon Chua published on 31 March 2022.

[13] Speech emotion recognition using deep learning Author links open overlay panelTanmoy Roy a, Marwala Tshilidzi a, Snehashish Chakraverty b published on 5 February 2021

[14] Arabic Speech Emotion Recognition Using Deep Neural Network by Omayma Mahmoudi and Mouncef Filali Bouami in the year 28 April 2023

[15] An ongoing review of speech emotion recognition by Javier de Lope a, Manuel Graña b published in 1 April 2023,
Theese Paper's are from google scholar