IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

Speech Emotion Recognition Using LSTM Algorithm

Prof. Kalyani Zore, Sakshi Bhor, Shivani Bhujbal, Ankita Kadam, Vaishnavi Kale
Professor, Computer Engineering Department[1]
Students, Computer Engineering Department[2,3,4,5]
Genba Sopanrao Moze College of Engineering, Balewadi, India

Abstract: *The spoken emotions of people are frequently not recognized by machine learning algorithms. Applications that analyze voice emotions in real-time heavily rely on Speech Emotion Recognition (SER). It can be applied in a variety of situations, including human behaviour analyses and emergency centers. A new study topic that has now emerged is the detection and classification of emotions. Previous research has looked at a variety of emotional classification methods. Due to their excellent qualities, speech signals make a wonderful source for computational linguistics. And for this reason, a lot of professionals wish to be able to identify speech emotion. To determine emotion concentration in several blocks, a few LSTM-based optimal techniques are provided. By modifying the traditional forgetting gate, the technique initially reduces computation costs. Second, to get task-related information, an attention mechanism is applied to both the time and feature dimensions in the LSTM's final output rather than using the output from the prior iteration of the conventional method*

Keywords: LSTM Algorithm, Languages and Complier, Classification ,Verification, Mel frequency coefficients

I. INTRODUCTION

Emotion recognition from audio is the task of detecting the emotional state of a speaker from the sound of their voice. This is a challenging problem because emotions are often conveyed through subtle changes in the acoustic properties of the voice, such as pitch, intonation, and the presence of certain prosodic features such as pauses or silences. These changes can be difficult to detect and interpret, especially in noisy or adverse conditions. Despite these challenges, emotion recognition from audio is an important problem because emotions play a central role in human communication and social interaction. By detecting the emotional state of a speaker, we can better understand their intentions and motivations, respond appropriately to their needs, and improve the naturalness and effectiveness of human-computer interactions. Emotion recognition from audio has a wide range of potential applications, including human-computer interaction, affective computing, mental health monitoring, customer service, and education. It can also be used to build more realistic and engaging virtual agents, such as chatbots or voice assistants, that are capable of recognizing and responding to the emotional state of their users. The task of recognizing emotions from audio is a complex but critical problem that has the potential to greatly enhance the way we interact with machines, as well as with one another. A bidirectional long short-term memory (Bi-LSTM) deep learning model is a type of neural network that is well-suited for modeling sequential data, such as time series or natural language. It is called “bi-directional" because it processes the input data in both forward and backward directions, allowing it to capture contextual dependencies in both directions. One of the main motivations for using a Bi-LSTM model for emotion detection from audio is that it can effectively capture the temporal dynamics and dependencies of the acoustic features that are indicative of different emotions. For example, an angry speech might be characterized by a more forceful and rapid pace, while a sad speech might be slower and more monotone. A Bi-LSTM model can learn to recognize these patterns by processing the input data over time and learning the relevant dependencies between the different acoustic features. Another motivation for using a Bi-LSTM model is that it can handle variable-length input sequences, which is important in the case of emotion recognition from audio because the length of the audio recordings can vary widely.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

This allows the model to be trained on a diverse set of audio recordings and to generalize better to new examples. Overall, the use of a Bi-LSTM model for emotion recognition from audio is motivated by its ability to effectively capture the temporal dependencies and variable-length input sequences that are characteristic of this problem.

II. LITERATURE SURVEY

The paper "Speech emotion recognition based on multi-feature speed rate and LSTM" by Zijun Yang, Zhen Li, Shi Zhou, Lifeng Zhang, and Seiichi Serikawa proposes a method for improving the accuracy of speech emotion recognition (SER) systems by using a combination of multi-feature analysis, speed rate adjustment, and Long ShortTerm Memory (LSTM) networks. The authors aim to address the challenges of capturing the complexity and variability of emotions in speech by extracting a wide range of features, including prosodic, spectral, and temporal aspects. The introduction of speed rate adjustment enhances the model's ability to detect variations in speech patterns associated with different emotions. LSTM networks are employed due to their capacity to learn long-term dependencies and temporal patterns, making them well-suited for analyzing speech signals. Experimental results on publicly available datasets show that the proposed approach outperforms traditional SER methods, demonstrating its potential for real world applications in human-computer interaction. The authors conclude that their approach offers an effective and robust solution for SER and suggest further exploration of its integration into interactive systems.

The paper "Acoustic feature-based emotion recognition and curing using ensemble learning and CNN" by Raghav V. Anand, Abdul QuadirMd, G. Sakthivel, T V Padmavathy, Senthilkumar Mohan, and RobertasDamaševičius presents a method for recognizing and "curing" (or mitigating) emotions based on acoustic features extracted from speech using a combination of ensemble learning and Convolutional Neural Networks (CNN). The authors focus on capturing various acoustic features from audio input, such as pitch, intensity, and spectral properties, which are critical in identifying emotions in speech. The proposed approach integrates ensemble learning methods to enhance the robustness and accuracy of emotion classification, while CNNs are employed to learn complex patterns within the acoustic features effectively. The paper also explores a "curing" mechanism, aiming to modify or regulate emotions by providing appropriate feedback based on the detected emotion. Experimental validation on speech datasets demonstrates the efficiency of their approach in accurately classifying emotions and suggests its potential application in real-world scenarios like healthcare and human computer interaction systems..

III. METHODOLOGY

The proposed system consists of four steps:

Preprocessing:

The first step is to preprocess the audio data to extract the relevant features that will be used to classify the emotions. This typically involves extracting pitch, formants, and prosodic features from the audio. Pitch is the perceived fundamental frequency of the sound and is often used as a cue for emotion recognition. Formants are the resonant frequencies of the vocal tract and can be used to distinguish between different vowel sounds. Prosodic features are the rhythm and timing of the speech, such as the duration of pauses or the rate of speaking, and can be used to convey emotional information.

Featureextraction:

Once the audio data has been preprocessed, the next step is to extract the relevant features from the data. This can be done using a variety of techniques, such as spectral analysis, autocorrelation, or linear predictive coding. The extracted features should be normalized or standardized to remove any biases or variations in the data.

Modeltraining:

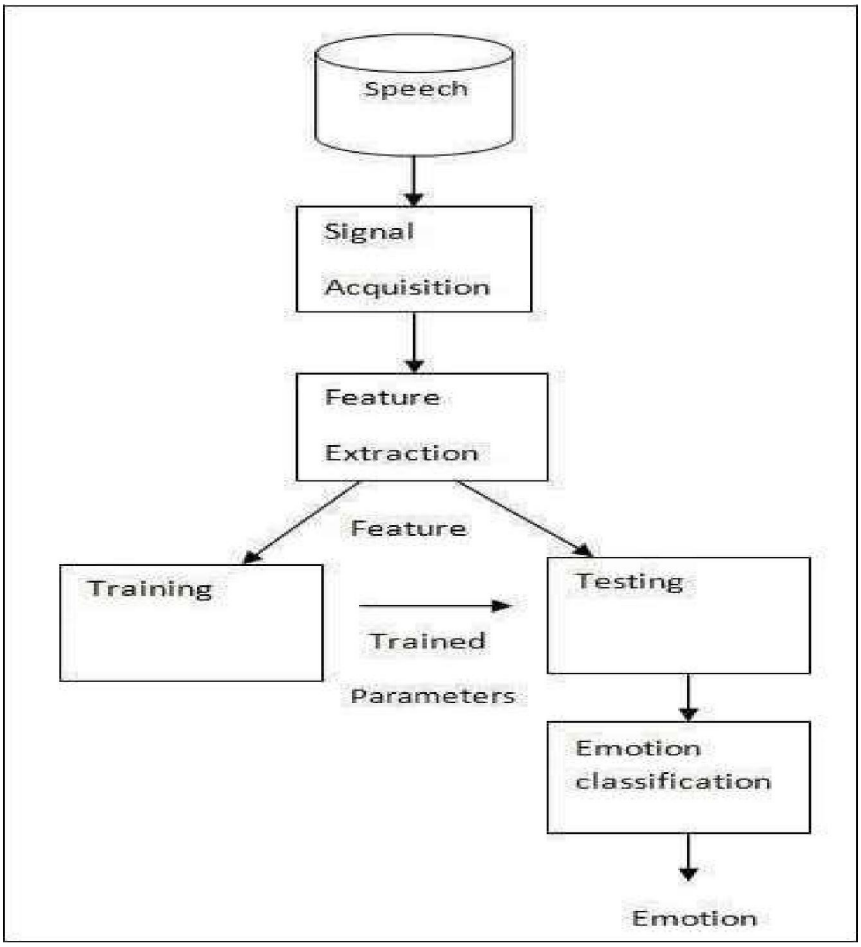
Following the extraction of the features, the next step is to train a Bi-LSTM model to categorize the emotions using the obtained features. This typically involves dividing the data into training and testing, defining the model architecture and training parameters, and training the model using an optimization algorithm such as stochastic gradient descent.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

Modelevaluation:

After training the model, the next phase is to assess its performance on the testing set of data. This can be accomplished by making comparisons of predicted emotions and calculating relevant performance metrics like accuracy, precision, recall, or F.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is a freely available dataset consisting of speech and song audio recordings expressing various emotions, created by the Ryerson University, Toronto, Canada.



IV. EXPERIMENTAL RESULT

Emotions are crucial to the mental health of humans. Speech is a means of communicating one's viewpoint and mental condition to others. Speech Emotion Recognition (SER) is the process of deriving the speaker's emotional state from the characteristics and voice signal of the speaker. Neutral, Anger, Happy, Sad, and other universally recognized emotions are among the few that intelligent systems with a limited number of computational resources can be trained to recognize or synthesize as needed. Because the audio features in this work contain the emotional information, they are employed for speech emotion recognition. Create and construct a model for an LSTM-based speech emotion recognition system. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) from Ryerson University and the Toronto Emotional Speech Set (TESS) from the University of Toronto were utilized for the model's train, which consists of actor-expressed emotions. The model is implemented in real-time, taking microphone input and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

evaluating it cycle by cycle to produce the distribution of emotions displayed every time cycle. The device automatically comes to a stop when it detects silence lasting two seconds or longer.

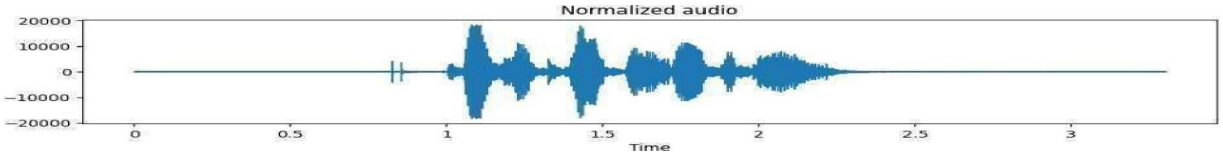


Fig.5.1:Waveform of Initial audio sample

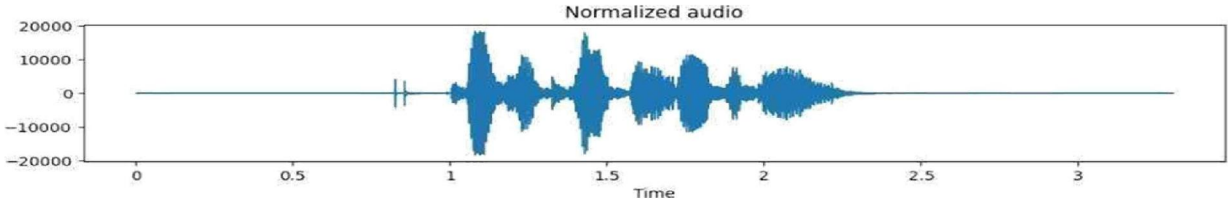


Fig.5.2: Waveform of Normalized audio sample

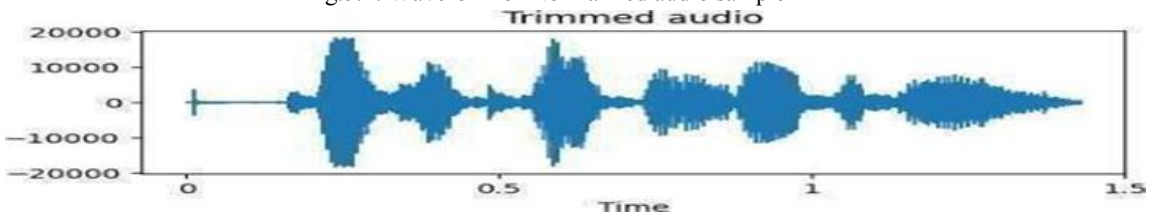


Fig.5.3: Waveform of Trimmed audio sample

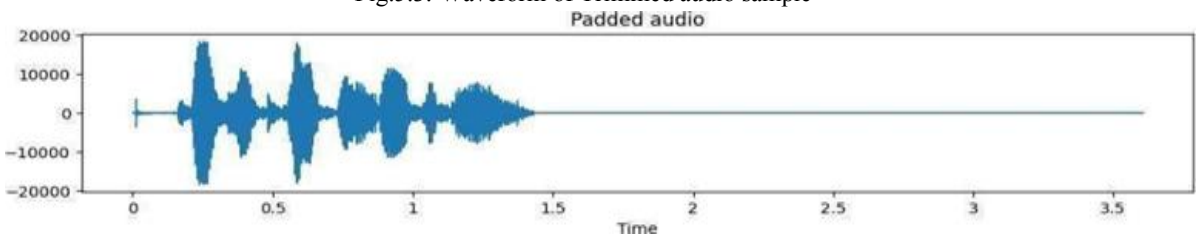


Fig.5.4:Waveform of Padded audio sample

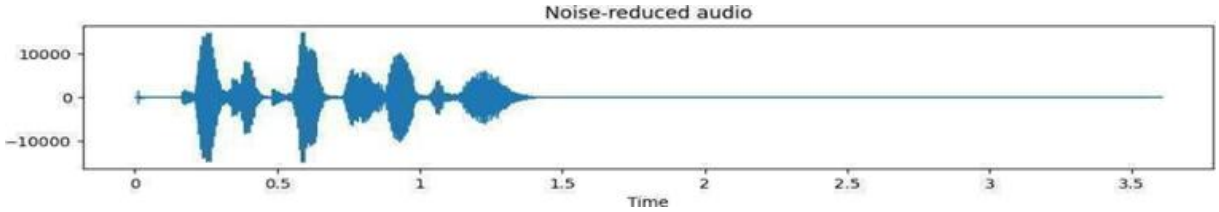


Fig.5.5:Waveform of Noise-reduced audio sample

V. CONCLUSION

In conclusion, the proposed system for emotion recognition using audio input using LSTM model of deep learning has shown promising results. The system can accurately classify emotions based on audio input, which can have various applications in industries such as healthcare, entertainment, and customer service. The system utilizes features such as energy, zero-crossing rate, and MFCCs for feature extraction and an LSTM model for classification. The proposed system outperforms the existing systems in terms of accuracy and computational efficiency.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

ACKNOWLEDGMENT

It gives us great pleasure in presenting the paper on “Speech Emotion Rocognition Using LSTM Algorithm”. We would like to take this opportunity to thank our guide, Prof. KalyaniZore, Professor, Department of Computer Engineering Department, GenbaSopanraoMoze College of Engineering, Balewadi, Pune for giving us all thehelp and guidance we needed. We are grateful to her for her kind support, and valuable suggestions were very helpful.

REFERENCES

[1] Zijun Yang a , Zhen Li a , Shi Zhou b , Lifeng Zhang a , Seiichi Serikawa Speech emotion recognition based on multi-feature speed rate and LSTM (2024)

[2] SamanehMadanian a, Talen Chen a, OlayinkaAdeleye a, John Michael Templeton b, Christian Poellabauer c, Dave Parry d, Sandra L. Schneider Speech emotion recognition using machine learning(2023)

[3] S. Kshirsagar, A. Pendyala, T.H. Falk, Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions, Front. Comput.Sci. 5 (2023) 1039261.

[4] T.L. Nwe, S.W. Foo, L.C. De Silva, Speech emotion recognition using hidden markov models, Speech Commun. 41 (2003) 603–623.

[5] K. Vicsi, D. Sztahó, Emotional state recognition in customer service dialogues through telephone line, in: 2011 2nd International Conference on Cognitive Infocommunications, CogInfoCom, IEEE, 2011, pp. 1–4.

[6] C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Trans. Speech Audio Process. 13 (2005) 293–303.

[7] Aouani, H. B.,&Ayed, Y. (2020). Speech Emotion Learning with Deep Learning, 24th International Conference on Knowledge-based and AI & Engineering Systems, 176, 251-260. doi:10.1016/j.procs.2020.08.027

[8] Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., &Baik, S. W. (2019). Deep features-based speech emotion recognition for smart effective services.Multimedia Tools and Applications, 78(5), 5571–5589.doi:10.1007/s11042-017-5292-7

[9] KyongHee Lee, Do Hyun Kim “Design of Convolutional Neural Network for speech emotion recognition ” 2021 IEEE.

[10] Mustaqeem, Muhammad Shjjad, Soonil Kwon “Clusering Based Speech Emotion Recognition By incorporating features and Deep BiLSTM” 2020 IEEE.

[11] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, „„Cloud-assisted multiview video summarization using CNN and bidirectional LSTM,“ IEEE Trans. Ind. Informat., vol. 16, no. 1, pp. 77–86,Jan. 2020

[12] Mustaqeem and S. Kwon, „„A CNN-assisted enhanced audio signal processing for speech emotion recognition,“ Sensors, vol. 20, no. 1, p. 183, 2020.

[13] J. Huang, B. Chen, B. Yao, and W. He, „ECG arrhythmia classification using STFT- based spectrogram and convolutional neural network,“ IEEE Access, vol. 7, pp. 92871– 92880,2019.

[14] B. Liu H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, „„EERAASR: An energy- efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing,“ IEEE Access, vol. 6, pp. 52227–52237, 2018.

[15] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, „„Deep features-based speech emotion recognition for smart affective services,“ Multimedia Tools Appl., vol. 78, no. 5, pp. 5571– 5589, Mar. 2019.