Kalpa Publications in Computing

Volume 19, 2024, Pages 172–180

Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

Speech Emotion Recognition Using LSTM and MFCC features

Arun K[1.2], Nandhini M [1.2] and Sasirekha R [1.2]

[1] Department of Computer Science
[2]Sathyabama Institute of Science and Technology
Chennai, India

[1]arunkari007@gmail.com, [2]sigapim6@gmail.com and [3]sairekha.r.cse@sathyabama.ac.uk

Abstract

Speech is utilized in human-machine connection and serves as a signal of human involvement. The Speech Emotion Recognition (SER) system is a novel form of this interactive system. Sufficient intelligence is provided by the SER to facilitate effective human-computer interaction. Based on the speaker's words, the SER system classifies emotions into groups such as "neutral," "calm," "happy," "sad," "angry," "fearful," "disgust," and "surprise." Languages and machine learning models suitable for SER are defined in this paper. Deep learning is used by this system to effectively classify and learn from multidimensional data. Primary results for a system using the LSTM algorithm and MFCC feature tools are also presented in this work. For the simplicity of user engagement, we have then implemented this model as a website through the usage of a third party.

Keywords: Machine Learning; Mel Frequency Cepstrum Coefficient; Long Short Term Memory; Speech Emotion Recognition.

1. Introduction

People often use words to communicate. These words encourage scientists to think about communication with machines. Many systems, such as smartphone services, speech-to-text, and voice-activated system commands, have been developed from this concept. However, the system is slow to communicate with people. By giving the machine some abilities, this can be made better. When a machine can recognize human emotions, it can understand people better. Speech Emotion Recognition (SER) helps machines recognize human emotions and respond accordingly.

Security, entertainment, and biomedicine are areas where SER is useful. Automatic pager services are a prime example of the need to understand customer strategies in order to provide appropriate service. Anger or frustration, referred to as "so-called" by customers, will be transferred to humans rather than to the automatic Pager. Online video and computer learning applications can manage or respond to the watcher's desire for an improved experience. SER can be used as a safety device. If the SER system predicts that the driver is mentally disabled, it will take control of the vehicle. The SER system can be used as an analytical tool for medical professionals.

Speech is challenging to understand because of variations in speakers, speech patterns, speech rate, and sentence structure. The fact that certain demands vary depending on the speaker, culture, and setting presents another difficulty. Scholars have presented a number of techniques utilizing deep learning and multilingualism. Analysis-useful speech characteristics include formants, force, intonation, linear frequency cepstral coefficients (LFCC), MFCC, TEO, etc.

The individual who speaks is what makes a speech sound good. These advantages are always personalized. This modification leads to issues with categorization. Large volumes of complicated data cannot be processed efficiently by standard categorization techniques. Deep learning algorithms are employed in SER systems as a result. Learning algorithms like CNN (Convolutional Neural Network), ANN (Artificial Neural Network), LSTM (Long Term Memory), and SVM have been introduced by numerous researchers for SER systems.

This study is divided into four sections: the second section reviews the literature on cognitive conversation theory; the third section outlines the approach; and the fourth section discusses how information about conversational curiosity is treated and controlled. describing the implementation of this model into a website and elucidating how to apply the method's results in the system.

2. Review of Existing Work

Speech emotion recognition plays a vital role in predicting the emotion for safety and precaution.

Authors [1] Build a model it contributes to the growing body of knowledge in the area of SER by employing deep learning methodology coupled with MFCC features. The main objective of this research is to design and tool a deep learning model that effectively leverages MFCC features for accurate emotion recognition in human speech. They aim to exceed the limitations of traditional methods by exploiting the temporal dependencies present in sequential speech data.

Authors [2] This methodology involves the utilization of RNNs, specifically personalized to the sequential nature of speech data. The incorporation of Local Attention mechanisms enables the model to selectively attend to specific temporal segments, providing a fine-grained analysis of the emotional content present in the speech signal. This work contributes to the field of ASER by introducing a novel approach that combines the strengths of RNNs and Local Attention mechanisms.

In [3], The authors suggested a technique for voice emotion recognition that makes use of LSTM and auto encoder bottleneck features. It is probable that the suggested model's effectiveness is assessed using common metrics for speech emotion identification. Accuracy, precision, recall, F1 score, and perhaps measures unique to the emotion identification domain are examples of common metrics.

In [4], The methodology put forth by the authors comprises multitask learning, tackling age/gender and speech emotion identification tasks at the same time. For both purposes, the DNN model might be built to share and learn common representations [5-11].

3. Proposed Methodology

3.1 Feature extraction

The feature extraction process in speech emotion recognition using LSTM begins with preprocessing the raw speech signals, which involves steps such as framing, windowing, and pre-emphasis to prepare the data for analysis. Once the speech signals are segmented into frames, various acoustic features are extracted from each frame to represent the underlying characteristics of the speech. These features typically include Mel Frequency Cepstral Coefficients (MFCCs), which capture the spectral content of the speech signal, as well as prosodic features such as pitch, energy, and duration. Additionally, other time-domain and frequency-domain features may be extracted to capture temporal and spectral variations in the speech signal. These extracted features serve as input to the LSTM neural network, which is capable of learning long-term dependencies in sequential data. The LSTM model processes the sequential feature representations over time, capturing temporal dynamics and patterns in the speech signals to classify the corresponding emotions accurately. By leveraging the capabilities of LSTM networks to model temporal dependencies effectively, the feature extraction process enhances the discriminative power of the speech emotion recognition system, enabling it to effectively classify emotions based on acoustic cues extracted from speech signals.

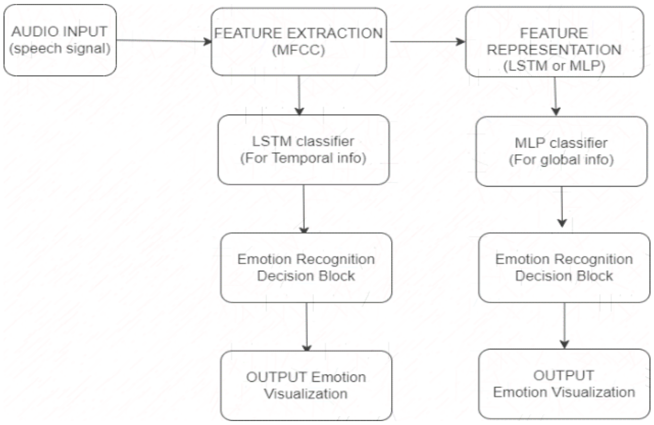


Figure.1: Flow diagram of MFCC Process

Figure 1 shows the flow diagram of the MFCC process. The first priority is the processing of speech, where first filters are used to obtain a softer spectrum. In the frame block technique, the audio signal is divided into many small, overlapping parts. The plan's frame size and connecting line spacing are both 20 milliseconds. The procedure needed to analyse the lengthy signal called windowing. This method gets rid of aliasing. A spectrum is created by applying the Fast Fourier Transform (FFT) to the time domain signal. The linear frequency scale is converted to the Mel frequency scale using the Mel frequency filter bank. The human ear's perception of sound frequency serves as the foundation for the Mel frequency scale. Because of its logarithmic scale, Mel Frequency is more sensitive to low frequencies than high ones. The Mel spectrum for the time domain is transformed using the Discrete Cosine Transform (DCT) in the cepstral process to yield the Mel frequency cepstral coefficients (MFCC).

3.2 Dataset Description

The Ryerson Emotional Speech and Song Audio-Visual Database (RAVDESS) is the database utilized in this system. Twelve female and twelve male actors' voices can be heard in the 7356 files that make up the collection. Speaking in the centre of North America, he uttered two words. Seven emotions are involved in speech: surprise, fear, rage, pride, happiness, sadness, and shame. Every expression has a naturally occurring, intensely emotional supplementary middle expression. Waveform (.wav) is the file format used for audio files, which are 16-bit, 48 kHz.

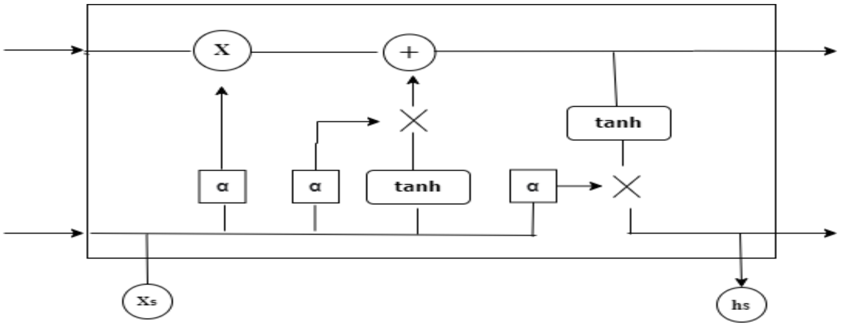
Emotion	Count
Neutral	96
Calm	192
Happy	192
Sad	192
Angry	192
Fearful	192
Disgust	192
Surprised	192
Total	1440

Table 1: Audio’s in each class of dataset.

The total number of audio files in every category of emotion—happy, sad, angry, afraid, neutral, disgusted, and surprised—is displayed in Table 1 above.

3.3 Machine Learning Model

This method's machine learning model is based on the LSTM architecture. An adaptation of the Artificial Recurrent Neural Network (RNN) architecture is Long Short-Term Memory (LSTM). Big data and enough training data are more helpful to the effectiveness of LSTM. The primary benefit of using RNN instead than ANN is its higher accuracy when dealing with sequential data. In the case of speech as a sign, this little part is subdivided into smaller sections, and it is up to the searcher to determine how each part depends on the part before it. Therefore, in this case, LSTM can offer higher performance.



$$i_t = \sigma\,(W_i h_{t-1} + U_i X_t + b_i) \qquad 1.1$$
$$f_i = \sigma\,(W_f h_{t-1} + U_f X_t + b_f) \qquad 1.2$$
$$o_t = \sigma\,(Wo\ h_{t-1} + U_o X_t + b_o) \qquad 1.3$$
$$c_t = \tanh\,(W h_{t-1} + U X_t + b) \qquad 1.4$$
$$c_t = ft \circ C_{t-1} + i_t \circ c_t \qquad 1.5$$
$$h_t = Ot \circ \tanh\,(c_t) \qquad 1.6$$
$$y_t = h_t \qquad 1.7$$

1.8

Figure.2: LSTM Architecture

The equation defining the LSTM model (Figure 2)and the LSTM unit's construction are depicted in the above figure. W is the link between the current layer and the preceding layer, and U is the weight matrix containing the hidden input layer. The internal memory, denoted as C, is the result of combining the new computation hidden state access timings with the prior time gate memory. Based on the candidate's prior and current secret statuses, the candidate's secret status is determined. The idea of machine learning models is explained in the sentence above.

The first line shows the layers; the second line shows the activation function; the third line shows the total amount of unit / cells per layer; and the fourth line shows the quadruple pay-out. Each and every residence has multiple. The version value of the published system is 0.5. The total number of samples used for training was ninety-nine thousand

3.4 Deployment

A web application has been developed using Streamlit, a Python framework designed for machine learning and data science teams. This interactive platform allows users to analyze emotions in audio data by leveraging TensorFlow and librosa libraries. The application supports WAV format inputs, enabling users to upload selected audio files for emotion prediction. The underlying Python 3.11 code seamlessly integrates Streamlit for the user interface, while TensorFlow and librosa handle audio processing and emotion prediction, making it an efficient tool for emotion analysis through audio input on the web.

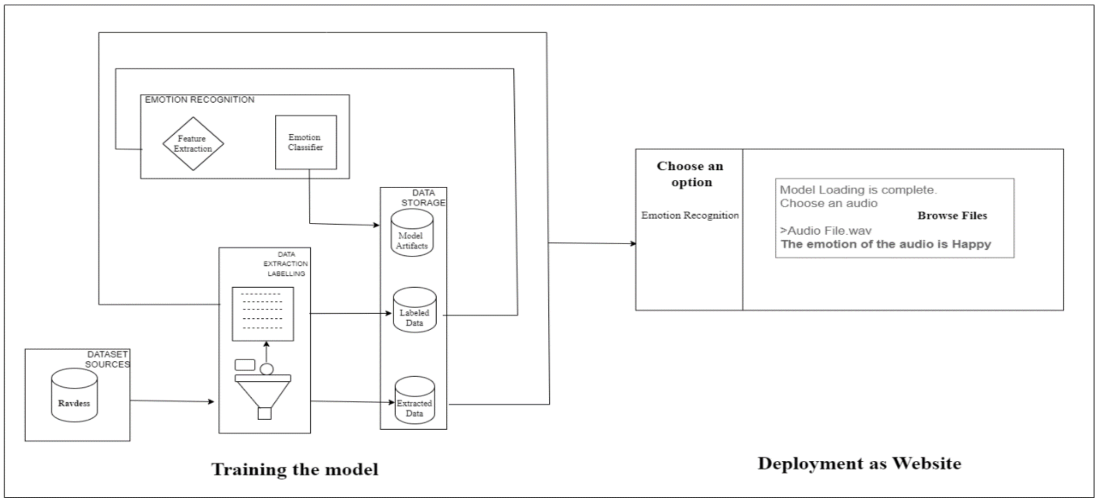


Figure.3: Deploying the model into a website

The above figure 3outlines the process of developing and training a model, followed by its deployment as a website to enhance user experience.

4. Preliminary Results

Python was used in the development of the MFCC feature extraction and machine learning models mentioned above. Coefficients of MFCC39 were eliminated during the extraction procedure. The reference model's real curve is depicted in Figure 3. The model received 100 training runs. Compared to published data, test data is less accurate. This curve demonstrates that, in fact, after the 30th cycle, it begins to stabilize. On test data, the average accuracy is 0.9722, or 97.22%.

Using the model, Figure 4 illustrates how inclusivity declines with time. It has been demonstrated that lengthening the training material's periods lowers attrition. The research indicates that the model may achieve local losses about 30 times, which will lessen the adverse effect on the pricing, if the loss increases after 30 times. In the model, the average loss is 0.1608. Consequently, there are chances to lower loss and boost productivity. Training and Validation Accuracy is shown in Figure 5.

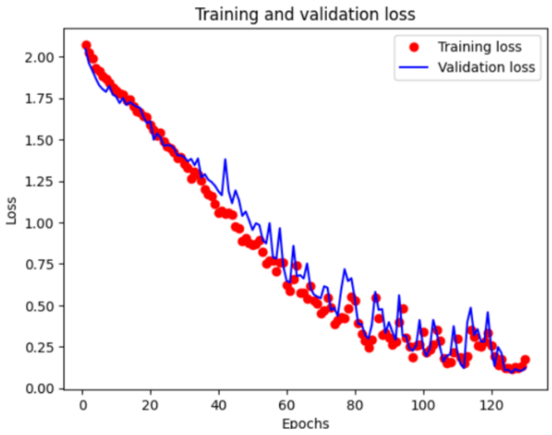


Figure.4: Training and Validation Loss

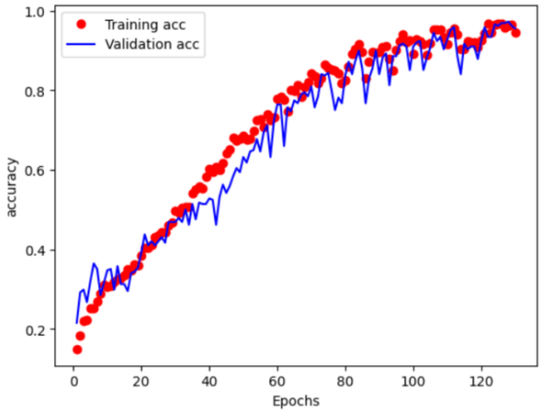
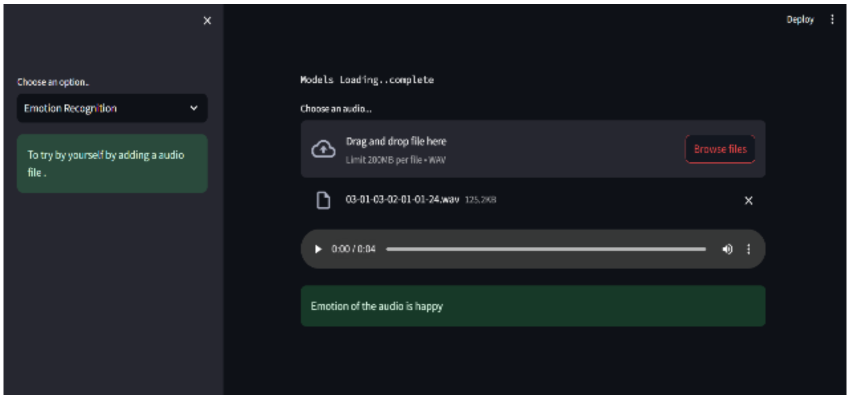


Figure.5: Training and Validation Accuracy



Choose an option..

Emotion Recognition

To try by yourself by adding a audio file .

```
Models Loading..complete
```

Choose an audio..

Emotion of the audio is happy

Figure.6: The Result predicted in a website

The website(Figure 6) uses an intuitive interface to display the predicted emotion classification results. This interactive platform makes use of advanced machine learning algorithms to forecast emotions accurately, giving users a quick and easy way to examine and understand the emotions included in audio data. The website's entire emotion classification experience is improved by the smooth integration of the WAV file format, which guarantees compatibility and convenience of usage.

5.Conclusion

This paper uses the LSTM model and MFCC features for speech recognition. Experiments have revealed that the widely used MFCC produces better outcomes for cognitive theory in SER. In this instance, a loss of 16.08% was noted which had to be fixed. The system's accuracy level of 97.22% indicated the potential of the reference model. To obtain a higher score, still, it is possible to refine the machine learning model and merge multiple characteristics. To provide greater user comfort, we have used online application to the deployed model to predict emotions.

References

[1] Emotion Recognition of Human Speech Using Deep Learning Method and MFCC Features by Sumon Kumar Hazra, Romana Rahman Ema, Syed Md. Galib, Shalauddin Kabir, Nasim Adnan. Radio Electronic and Computer Systems Vol.2022(4),pp. 161172,http://nti.khai.edu/ojs/index.php/reks/article/view/reks.2022.4.13/1959.

[2] Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention Seyedmahdad Mir Samadi, Emad Barsoum, Cha Zhang. IEEE conference on 05-09 March 2017.published on June 19, 2017.

[3] Speech Emotion Recognition Using Auto Encoder Bottleneck Features and LSTM Kun-Yi Huang, Chung-Hsien Wu, Tsung-Hsien Yang, Ming-Hsiang Su, and Jiahui chou. Publisher IEEE and published in**:** 2016 International Conference on Orange Technologies (ICOT).

[4] Learning Utterance-Level Representations for Speech Emotion and Age/Gender Recognition Using Deep Neural Networks Zhong-Qiu Wang and Ivan Tashev. Publisher IEEE Published in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[5] End - to - End speech emotion recognition using a deep convolutional recurrent Neural network - George Tri Georgis, Fabien Ringe Val, Raymond Brueckner, Erik Marche Michalis A. Nicolaou, Björn Schuller, Stefanos Zafeiriou. Published in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[6] Impact of autoencoder based compact representation on emotion detection from audio Nivedita Patel, Shireen Patel, Sapan H Mankad. Springer February 2022 Journal of Ambient Intelligence and Humanized Computing 13(3):1-19 DOI:10.1007/s12652-021-02979-3

[7] Efficient Multi- angle Audio-visual Speech Recognition using Parallel Wavegen Based scene classifier Shinnosuke Isobe, Satoshi Tamura, Yuuto Gotoh and Masaki Nose. DOI: 10.5220/0010846000003122 In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022).

[8] A Machine Learning Model for Automatic Emotion Detection from Speech. Nataliia Kholodna, Victoria Vysotska, Solomiia Albota DOI: 10.5220/0010846000003122 In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)

[9] Excitation Features of Speech for Speaker-Specific Emotion Detection. Sudarsana Reddy Kadiri Paavo Alku Published in IEEE on 24 March 2022.

[10] Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results**.**Misaki Sakurai Tetsuo Kosaka 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)**.**

[11] Multi-task Learning for Speech Emotion and Emotion Intensity Recognition Pengcheng Yue, Leyuan Qu, Shukai Zheng, Taihao Li. Published in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).