Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 21

Speech Emotion Recognition Using Deep Learning

Suraj Pradeepkumar[1], Ragul K[2], Raghul S[3] Mr.Stalin M, M.E,MBA, (Ph.D)[4],

[1,2,3] Students,and [4] Faculty
Dept. of Information Technology,
Jeppiaar Engineering College,
Chennai, India.
suraj12141509@gmail.com

Abstract: *Speech emotion recognition (SER) is a significant area of research with applications in human-computer interaction, healthcare, and entertainment. This paper presents a comprehensive review of recent advancements in SER, focusing on the use of deep learning techniques to improve recognition accuracy. The review covers various aspects of SER, including feature extraction, using methods such as Mel-frequency cepstral coefficients (MFCCs) to perform emotion modeling, and classification methods. Additionally, it discusses the challenges and future directions of SER research, such as dealing with imbalanced datasets and improving model generalization. The paper concludes with a discussion on the potential impact of SER on enhancing user experiences and enabling new applications, highlighting how SER can assist humans in their day-to-day tasks by providing emotional intelligence to machines.*
Keywords: *Speech Emotion Recognition, Deep Learning, Feature Extraction, Mel Frequency Cepstral Coefficients*

1. INTRODUCTION:

Speech emotion recognition (SER) is a modernistic approach to enhancing human-computer interaction. It involves the automated identification of emotions from speech, reducing the need for manual interpretation of vocal expressions. In SER, deep learning models process audio data to detect emotional states, interacting with users through existing interfaces without requiring system replacements. These models simulate human-like understanding by analyzing features such as tone, pitch, and rhythm in speech.

The implementation of SER is efficient, leveraging current technological infrastructures to integrate seamlessly with user interfaces. Adjustments to the model can be made dynamically, returning control to the user when needed, thus enhancing flexibility in operation. SER naturally improves user experience and operational efficiency while minimizing costs with minimal technological alterations. Models are designed to operate continuously, providing consistent emotional insights.

The deep learning models in SER adapt to changing data patterns, performing their tasks with high accuracy. SER is designed to handle frequent and repetitive emotion detection tasks, which is a crucial feature. As the volume of interaction data increases, the likelihood of human error in interpreting emotions rises, whereas deep learning models maintain reliable performance.

SER significantly impacts various domains, saving time and improving workflow efficiency in sectors like customer service and healthcare. Decision-making in SER is based on predefined algorithms and training, with no need for human intervention in routine tasks. The models adhere to specific rules and guidelines, ensuring consistent emotional analysis. Once trained, these models can be deployed across multiple systems, offering scalable and versatile emotion recognition solutions.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 21

2. THE DATASET

We take the audio data from 4 major datasets

• Crema (Crowd-sourced Emotional Multimodal Actors)

• Ravdess (Ryerson Audio-Visual Database of Emotional Speech and Song)

• Savee (Surrey Audio-Visual Expressed Emotion)

• Tess (Toronto Emotional Speech Set)

3. FEATURE EXTRACTION:

We extract features from the audio using a feature called MFCC or Mel Frequency Cepstral Coefficients. Mel-frequency cepstral coefficients (MFCCs) are a critical tool in the field of speech and audio signal processing, widely used for feature extraction in tasks such as speech recognition, speaker identification, and emotion recognition.

They are based on the Mel-frequency scale, which is a perceptual scale of pitches that are perceived by humans. MFCCs are particularly effective because they capture important aspects of the audio signal that are relevant for human perception while discarding irrelevant information, making them ideal for use in machine learning algorithms

The process of calculating MFCCs involves several key steps. First, the audio signal is divided into short, overlapping frames, typically around 20-40 milliseconds long, with a 50% overlap between consecutive frames. Each frame is then passed through a series of processing steps to extract the MFCCs:

1. Pre-emphasis: The signal is passed through a high-pass filter to emphasize high-frequency components, which helps to improve the signal-to-noise ratio.

2. Framing: The signal is divided into frames, each of which is multiplied by a window function (e.g., Hamming or Hanning window) to reduce spectral leakage

3. Fourier Transform: A Fourier transform is applied to each frame to convert the signal from the time domain to the frequency domain.

4. Mel-filterbank: The spectrum is passed through a series of triangular filters spaced evenly on the Mel-frequency scale. These filters are designed to mimic the frequency resolution of the human auditory system.

5. Logarithm: The logarithm of the filterbank energies is taken to account for the logarithmic nature of human perception of sound intensity.

6. Discrete Cosine Transform (DCT: Finally, a DCT is applied to the log filterbank energies to decorrelate the coefficients and extract the most relevant information.

The resulting MFCCs are a set of coefficients that represent the spectral characteristics of the audio signal in a compact and informative manner. Typically, the first 13 coefficients are used, as higher-order coefficients tend to contain less useful information. These coefficients can then be used as features for machine learning models to perform various tasks

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 21

Fig 1. MFCCs of Anger and Fear

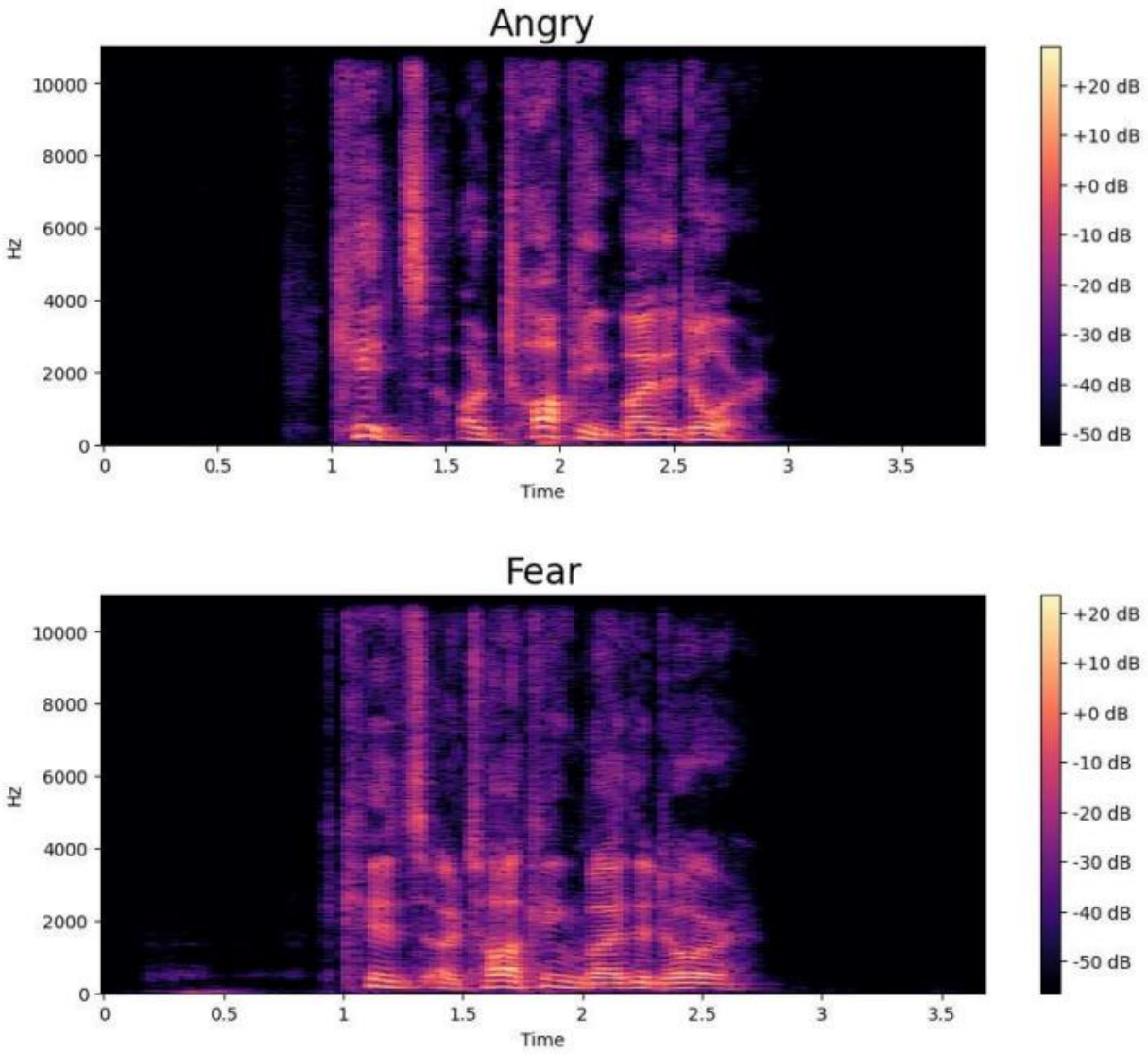
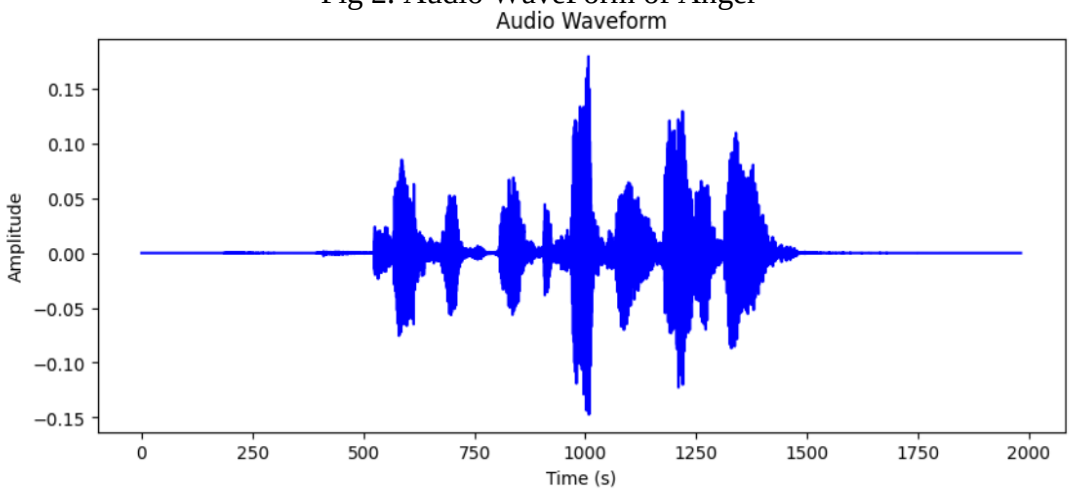


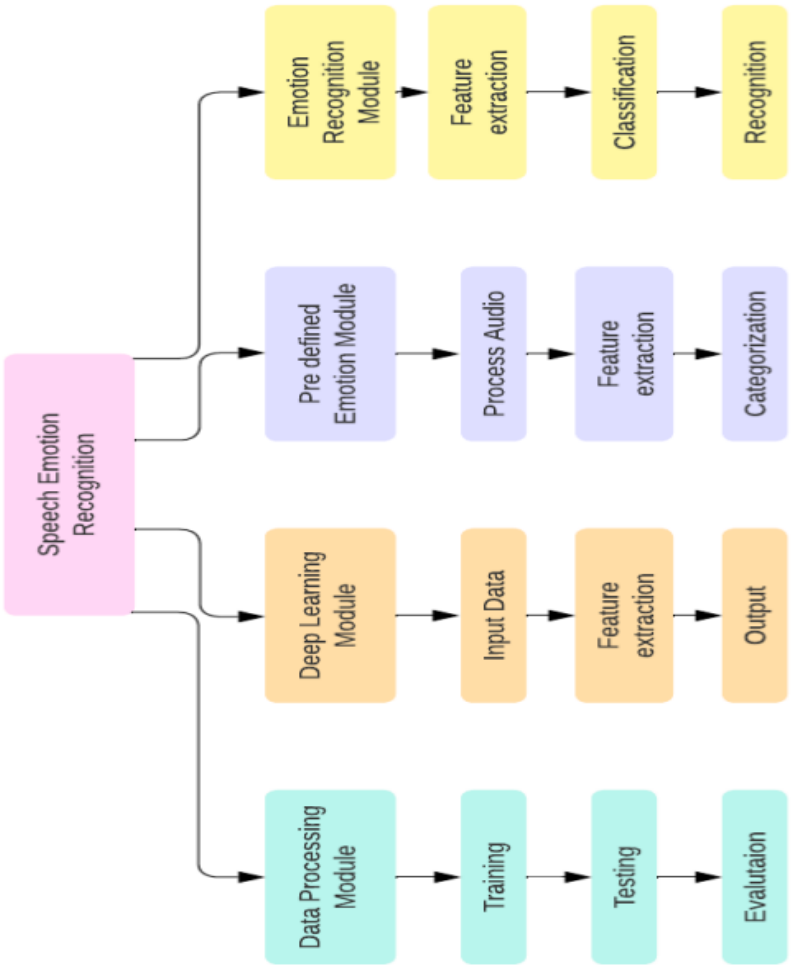
Fig 2. Audio WaveForm of Anger



Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 21

4. SYSTEM ARCHITECTURE:

Fig. 3. System Architecture



5. MODEL BUILDING:

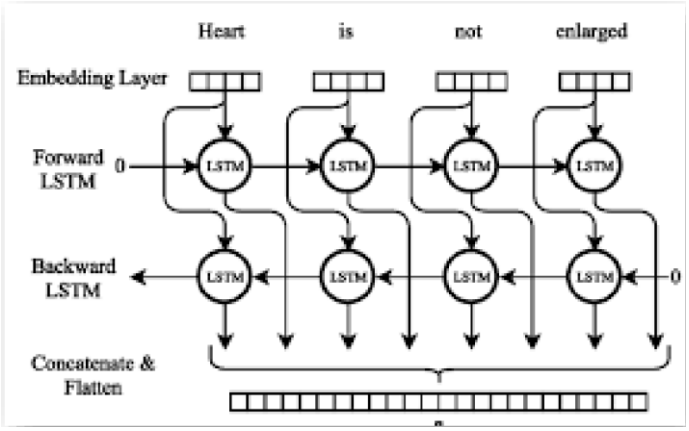
```
model = Sequential()
model.add(layers.Conv1D(256, 5, input_shape=(60, 1), activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.Conv1D(128, kernel_size=3, activation='relu'))
model.add(layers.MaxPooling1D(pool_size=2))
model.add(layers.BatchNormalization())
model.add(layers.Bidirectional(layers.LSTM(128, return_sequences=True)))
model.add(layers.Bidirectional(layers.LSTM(128, return_sequences=True)))
model.add(layers.GlobalMaxPooling1D())
model.add(layers.Dense(128, activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.Dropout(0.2))
model.add(layers.Dense(7, activation='softmax')) model.compile(loss='categorical_crossentropy',
optimizer='rmsprop', metrics=['accuracy'])
```

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 21

This code snippet defines a deep learning model using the Keras framework with a TensorFlow backend. The model is designed for sequential data processing, particularly for tasks like speech and audio signal classification, where the input data has a temporal sequence.

Overall, this model architecture is well-suited for sequential data processing tasks, with convolutional layers for feature extraction, LSTM layers for sequence modeling, and dense layers for classification. Batch normalization and dropout help improve the model's generalization and training stability.

Fig 4 Model Build



6. FUTURE ENHANCEMENTS:

We can further enhance this project by increasing the number of emotions it can detect. Furthermore, we can also get user data of auditory information with their emotions to widen the amount of data which we use for the project, which can be used to enhance the chances of getting the proper emotion hence improving the efficiency and making the project more accessible and faster than human response.

7. CONCLUSION:

Hence we can conclude that we successfully built a Speech emotion recognition model using deep learning which captures features from the given audio input and classifies the emotion portrayed accordingly. This project was then successfully deployed using streamlit and python.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 21

8. REFERENCES:

[1] Z. Huijuan, Y. Ning and W. Ruchuan, "Improved Cross Corpus Speech Emotion Recognition Using Deep Local Domain Adaptation," in Chinese Journal of Electronics, vol. 32, no. 3, pp. 640-646, May 2023, doi: 10.23919/cje.2021.00.196.
[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
[3] S. Kakuba, A. Poulose and D. S. Han, "Deep Learning Approaches for Bimodal Speech Emotion Recognition: Advancements, Challenges, and a Multi-Learning Model," in IEEE Access, vol. 11, pp. 113769-113789, 2023, doi: 10.1109/ACCESS.2023.3325037.
[4] Schuller B., Batliner A., Gollan C., Black A.W., Cohn J.F., Roller J., Picard R.W., EmoVoice 2000: A Sensitive Aural and Visual Emotion Recognition Corpus. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2001 (Vol. 1, pp. 1-4).
[5] Trigeorgis G., Ringeval K., Schuller B. A Deep Learning Framework for Multimodal Emotion Recognition in Speech and Text. In: Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction (ACII) 2017 (pp. 260-266).