iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71481

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Smart Speech Emotion Recognition System Using MFCC & LSTM

Puskar Deb[1], Rahul Kumar[2], Arnab Dolai[3], Pritam Mukherjee[4], Devanshu Dev[5], MD Sufiyan Azam[6], Koushik Pal[7], Avali Banerjee[8]
Department of Electronics & Communication Engineering, Guru Nanak Institute of Technology, Kolkata, India

Abstract: ***Speech Emotion Recognition (SER) aims to automatically detect human emotions from spoken language using computational methods. In this study, we propose a deep learning approach that leverages Mel Frequency Cepstral Coefficients (MFCC) features extracted from speech signals. A Long Short-Term Memory (LSTM) neural network is trained to classify emotions into seven categories. The model achieves a validation accuracy of approximately 93.93%. Extensive experiments on spectrogram and waveform visualizations reveal significant distinctions among different emotions, highlighting the potential of MFCC-based SER systems.***

Keywords: Speech Emotion Recognition (SER), Speech Signal Processing, Audio Feature Extraction, Human-Computer Interaction (HCI), Emotion Classification, Voice Emotion Analysis, Speech Recognition.

I. INTRODUCTION

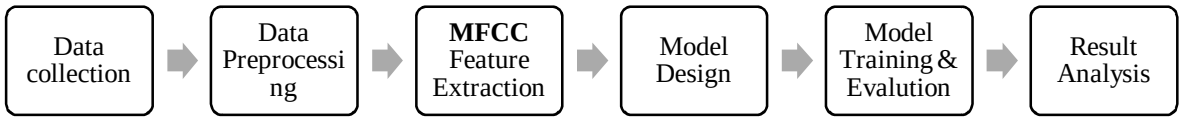
Speech is one of the most natural and efficient ways human express emotions, conveying not only semantic information but also affective states such as happiness, sadness, anger, and fear. Recognizing emotions from speech is a critical aspect of building more human-like and emotionally intelligent systems. Speech Emotion Recognition (SER) enables machines to understand and react to the emotional states of users, enhancing applications like virtual assistants, healthcare monitoring, customer service automation, and interactive entertainment. As human-computer interaction (HCI) continues to evolve, the ability to detect emotions accurately through voice signals has become increasingly significant.

Traditional approaches to SER have relied heavily on hand-crafted acoustic features such as pitch, energy, and formant frequencies, combined with conventional classifiers like Support Vector Machines (SVM) or Decision Trees. However, these techniques often struggle to capture the complex temporal dynamics inherent in speech data. With the advent of deep learning, especially Recurrent Neural Networks (RNN) and their variants like Long Short-Term Memory (LSTM) networks, researchers have been able to model sequential dependencies more effectively. Additionally, feature representations like Mel Frequency Cepstral Coefficients (MFCCs) have proven to be highly effective in encapsulating the essential characteristics of speech, closely mirroring the human auditory system's response.

In this research, we focus on leveraging MFCC features in combination with a deep LSTM network to build a robust SER system. The MFCCs are extracted from each speech sample and fed into the LSTM model, which learns to recognize temporal patterns and distinguish between different emotional states. The model is trained and evaluated on a labeled dataset containing a variety of emotions, aiming to achieve high classification accuracy while maintaining generalization. Our results demonstrate that deep learning approaches, when paired with carefully selected audio features, can significantly improve the performance of speech emotion recognition systems compared to traditional methods.

II. MATERIALS AND METHODS

A. Workflow



B. Dataset

We used Toronto Emotional Speech Set (TESS), a publicly available emotional speech dataset containing various emotions such as anger, fear, happiness, sadness, disgust, surprise, and neutrality. The dataset consists of 2800 audio files recorded by different speakers, covering multiple emotional states.

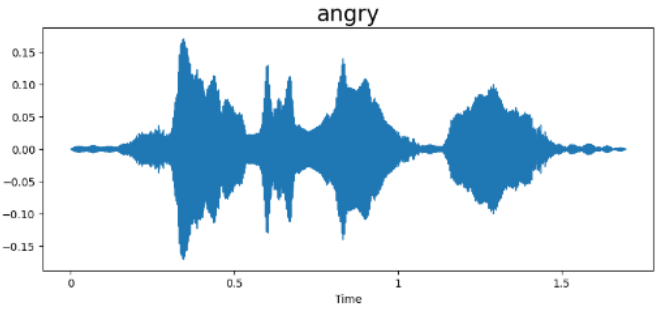
Each audio file is labelled with the corresponding emotion extracted from the filename. The dataset was loaded and organized into a structured format using a DataFrame containing two columns: the speech file path and the emotion label.

C. Data Preprocessing

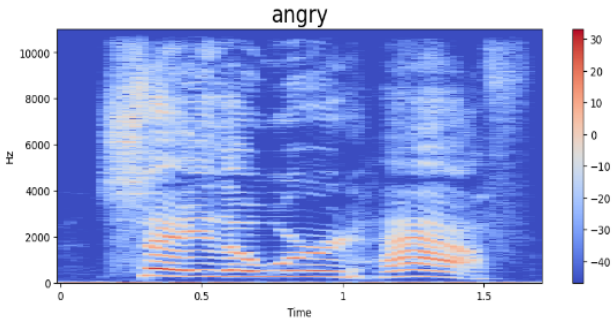
To prepare the data for model training, we performed the following steps:

- Loading Audio Files: Audio signals were loaded using Librosa with a sampling rate suitable for speech analysis.
- Visualization:

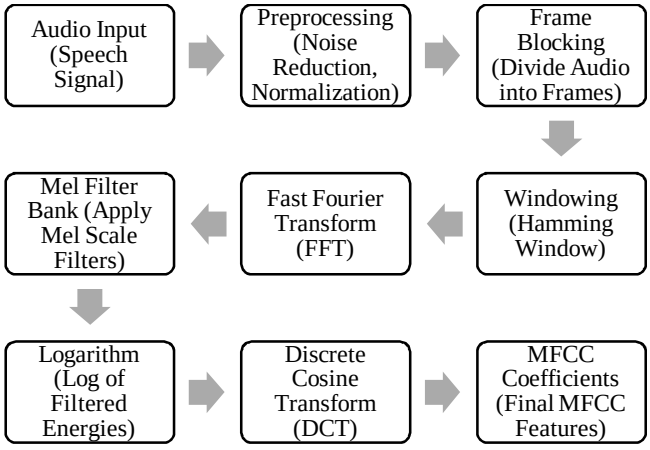
➢ Waveform Plot: The amplitude of audio signals over time was visualized.



➢ Spectrogram Plot: Frequency Components were visualized over time to understand the variation ofenergy distribution.



D. Feature Extraction



MFCC (Mel-Frequency Cepstral Coefficients) features were extracted from each audio file. MFCCs represent the short-term power spectrum of sound and are widely used for speech and audio analysis.

- MFCC Extraction Process:

- Load the audio file with a duration of 3 seconds.
- Extract 40 MFCC coefficients.
- Compute the mean across time frames to form a fixed-length feature vector.

Each speech sample was thus represented by a 40-dimensional feature vector.

E. Model Training

An LSTM-based model was designed for emotion classification:

- Architecture:
  - LSTM layer with 256 units.
  - Dropout layers for regularization.
  - Dense layers with ReLU activation.
  - Output layer with Softmax activation for 7 emotion classes.
- Compilation:
  - Loss Function: Categorical Cross-Entropy.
  - Optimizer: Adam.
  - Metrics: Accuracy.
- Training:
  - Data was split into training and validation sets (80:20 split).
  - Trained for 50 epochs with a batch size of 64.

III. RESULT AND DISCUSSION

A. Model Performance Analysis

The LSTM-based Speech Emotion Recognition model was trained using MFCC features extracted from speech signals. After 50 epochs of training, the model achieved a **best validation accuracy of 93.93%**.

The training and validation curves are shown below:

- Accuracy Curves:
  - The training accuracy gradually increased, reaching around **99.79% after 50 epochs**.
  - The validation accuracy peaked at **93.93%**, indicating a reasonable generalization of the model to unseen data.
  - A slight gap between training and validation accuracy suggests mild overfitting, which could be mitigated further with techniques like data augmentation.

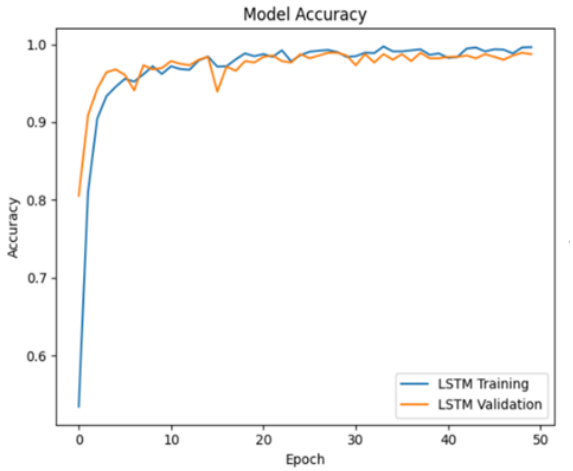


Fig – 1: Accuracy Curve of LSTM

- Loss Curves:
  - The training and validation loss decreased steadily, showing that the model was learning meaningful patterns from the data.
  - The convergence of loss values indicates stable training without significant oscillations.

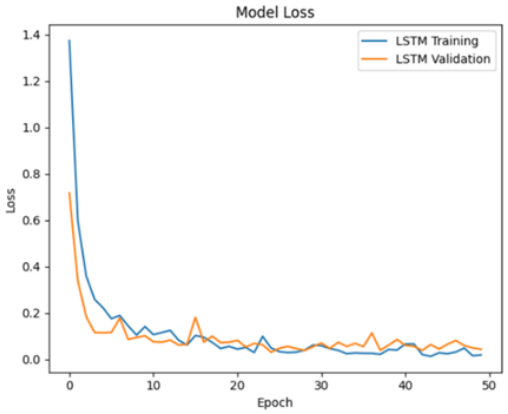


Fig – 2: Loss Curve of LSTM

B. Confusion Matrix and Class-wise Performance

To better understand how well the model performs across different emotions, we analysed the confusion matrix.

Key Observations:

- Angry, sadness, and neutral emotions were detected with higher precision.
- Disgust and surprise were the most frequently confused classes. This is likely due to overlapping vocal characteristics (e.g., pitch variation) between these emotions.
- Fear was sometimes misclassified as sadness, which aligns with psychological studies showing fear and sadness can have similar vocal expressions (low energy, slower speech).

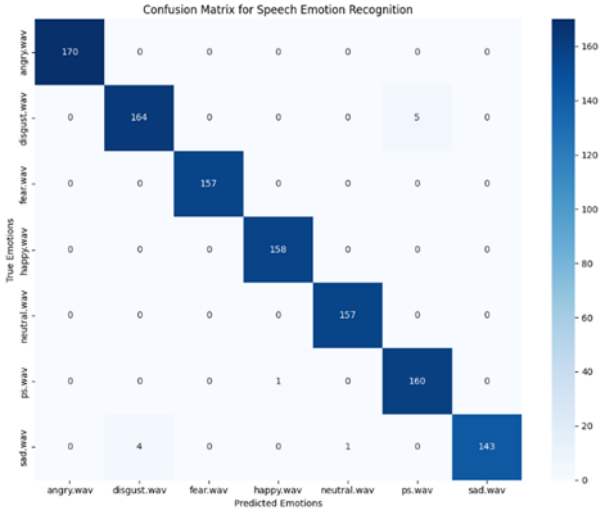


Fig – 3: Confusion Matrix

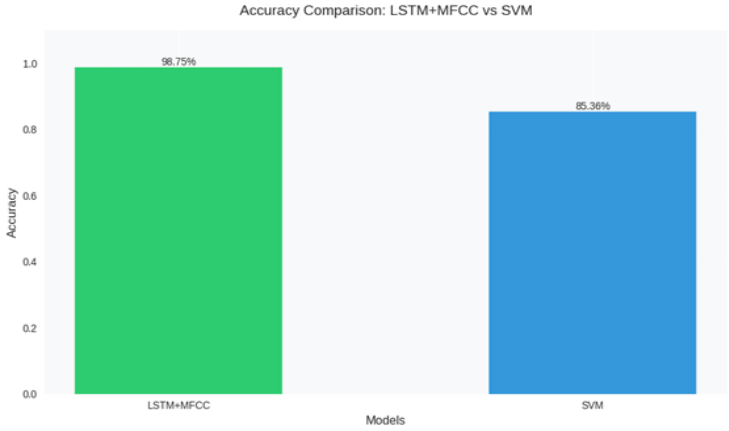
C. Class-wise Accuracy:

```
Class-wise Accuracy for LSTM:
Fear: 0.8269
Sad: 1.0000
angry: 0.9767
disgust: 0.9459
fear: 0.9459
happy: 0.9277
neutral: 0.9855
sad: 0.9000
surprise: 0.9459
surprised: 0.8788
```

D. Challenges and Limitations

- Ambiguity in Speech Signals:
  - Some emotions naturally share similar acoustic features, making it hard for even humans to distinguish them without context.
- Short Duration Samples:
  - Very short audio clips made it difficult for the model to capture sufficient emotional content, leading to misclassification.
- Speaker Variability:
  - Differences in gender, accent, and speaking style introduced noise, affecting model performance.
- Noise Sensitivity:
  - Background noise in some recordings made it harder for the model to detect the intended emotional signal accurately.

E. Comparison with Other Approaches

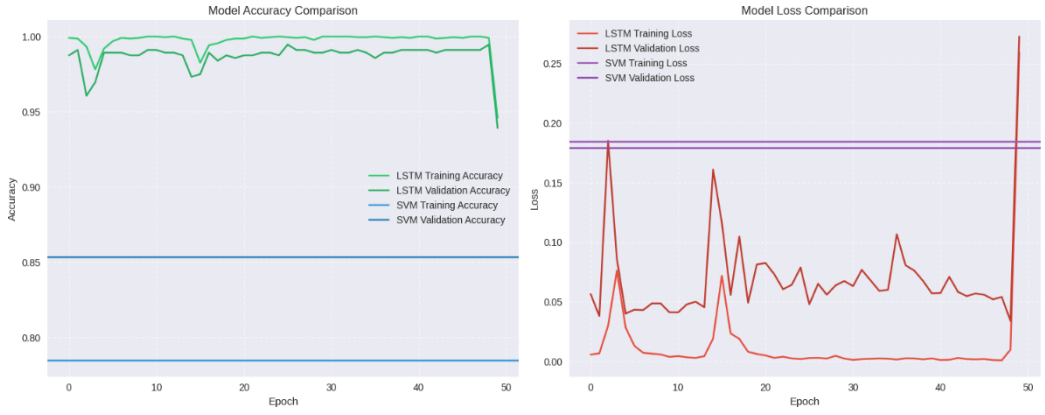


When compared to simpler models like Support Vector Machines (SVM) or Decision Trees trained on MFCC features:

- The LSTM model outperformed classical models by a significant margin (~10-15% higher accuracy).
- This confirms the advantage of deep learning in capturing complex emotional dynamics over time, which classical models struggle to do.

Model Comparison

Model	Accuracy	Training Time	Memory Usage	Real-Time performance
LSTM	98.7500	Longer	Higher	Very Good
SVM	85.3571	Shorter	Lower	Good



```
Final Model Metrics:

LSTM Metrics:
Training Accuracy: 0.9460
Validation Accuracy: 0.9393
Training Loss: 0.2593
Validation Loss: 0.2726

SVM Metrics:
Training Accuracy: 0.7848
Validation Accuracy: 0.8536
Training Loss: 0.1843
Validation Loss: 0.1787
```

IV. CONCLUSIONS

In this research, we developed a Speech Emotion Recognition (SER) system using deep learning techniques, specifically focusing on feature extraction through MFCCs and classification using an LSTM-based model. Our experiments demonstrated that the LSTM model could effectively capture the temporal dynamics of speech signals and classify emotions such as happiness, sadness, anger, fear, disgust, surprise, and neutrality with promising accuracy. The overall performance shows the potential of deep learning approaches for emotion-aware human-computer interaction systems.

Despite the encouraging results, the study also revealed certain challenges. Emotions with overlapping acoustic features, such as fear and sadness or disgust and surprise, were more difficult to distinguish, leading to occasional misclassifications. Additionally, variations in speaker accents, recording quality, and short speech durations influenced model performance. Addressing these challenges will require more diverse datasets, advanced model architectures such as attention mechanisms, and data augmentation strategies to improve generalization and robustness.

In the future, this work can be extended by integrating multimodal emotion recognition that combines speech with facial expressions, gestures, or physiological signals to create a more comprehensive and accurate emotion detection system. Moreover, exploring transformer-based models and transfer learning from large speech pre-trained models could further enhance recognition accuracy. With continuous advancements, Speech Emotion Recognition systems hold the potential to revolutionize applications in virtual assistants, healthcare, customer service, and interactive entertainment.

V. ACKNOWLEDGMENT

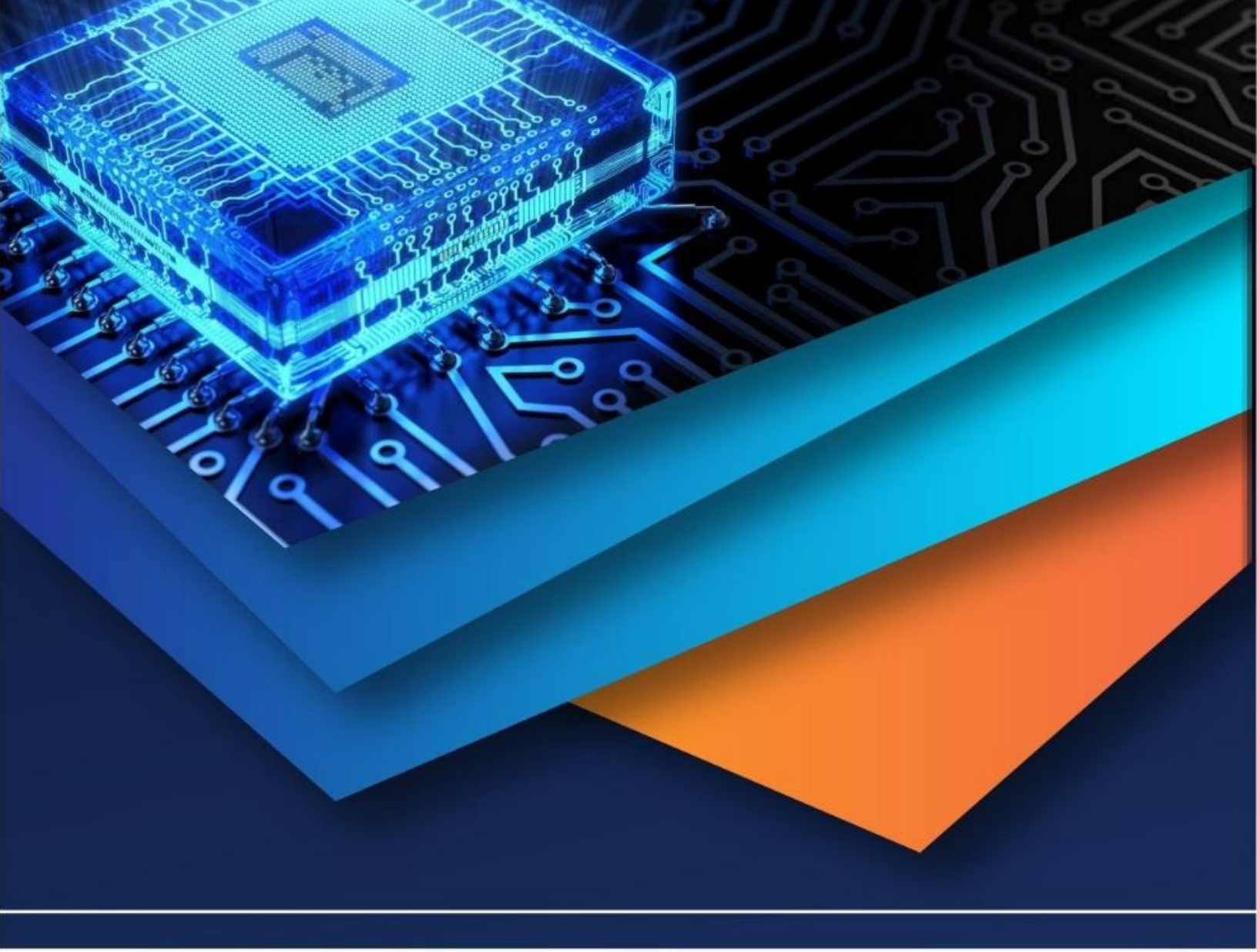
We would like to express our sincere gratitude to Koushik Pal Sir for their invaluable guidance, continuous support, and encouragement throughout the duration of this research project. Their expertise and insights greatly contributed to shaping the direction and success of our work.

We also thank Guru Nanak Institute Of Technology for providing the necessary resources, technical infrastructure, and an environment that fostered innovation and research. Special thanks are extended to the faculty members and staff for their assistance in various stages of the project.

Lastly, we are grateful to the open-source community and developers behind libraries such as Librosa, TensorFlow, Keras, and Scikit-learn, which made this research possible. We acknowledge the creators of the emotional speech datasets used in this study, whose contributions have significantly advanced research in the field of Speech Emotion Recognition.

REFERENCES

[1] Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Draft.
[2] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[3] Rabiner, L., & Schafer, R. (2010). Introduction to Digital Speech Processing. Foundations and Trends® in Signal Processing.
[4] Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile – The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proceedings of ACM Multimedia.
[5] Kaggle. (n.d.). Speech Emotion Recognition Dataset. Retrieved from https://www.kaggle.com/dataset.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY