Speech Emotion Recognition using Machine Learning

Guide: Dr. S China Venkateswarlu, Professor, ECE & IARE
Dr. V Siva Nagaraju, Professor, ECE & IARE

Yedavalli Geetha[1]

[1]Yedavalli Geetha Electronics and Communication Engineering & Institute of Aeronautical Engineering

---

Abstract -- Speech signals are being considered as most effective means of communication between human beings. Many researchers have found different methods or systems
to identify emotions from speech signals. Here, the various features of speech are used to classify emotions. Features like pitch, tone, intensity are essential for classification. Large number of the datasets are available for speech emotion recognition. Firstly, the extraction of features from speech emotion is carried out and then another important part is classification of emotions based upon speech. Hence, different classifiers are used to classify emotions such as Happy, Sad, Anger, Surprise, Neutral, etc. Although, there are other approaches based on machine learning algorithms for identifying emotions. Speech Emotion Recognition is a current research topic because of its wide range of applications and it became a challenge in the field of speech processing too. We
have carried out a brief study on Speech Emotion Analysis along with Emotion Recognition. Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions including Neutral, Anger, we have worked on different tools to be used in SER. SER is tough because emotions are subjective and annotating audio is challenging task. Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. We have classified based on
different type of emotions to detect from speech.

Key Words: Speech Emotion Recognition, Affective Computing, Machine Learning, Deep Learning, Audio Signal Processing, Emotion Classification, Feature Extraction, Prosodic Features, Spectral Features, Mel-Frequency Cepstral Coefficients (MFCCs), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Attention Mechanisms, Multimodal Emotion Recognition, Speaker-Independent SER, Real-Time Emotion Detection, Noise-Robust Emotion Recognition, Data Augmentation, Emotion-Aware Applications

1.INTRODUCTION

Speech Emotion Recognition (SER) using machine learning is a rapidly growing field that focuses on automatically identifying human emotions from speech signals. By analysing acoustic features such as pitch, tone, energy, and Mel-Frequency Cepstral Coefficients (MFCCs), machine learning models can classify emotional states like happiness, anger, sadness, and fear. Traditional approaches used handcrafted features with classifiers like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), but recent advances leverage deep learning techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks to capture complex temporal and spectral patterns in speech. Despite promising results, challenges like background noise, speaker variability, and limited labelled datasets persist, driving research into noise-robust models, data augmentation, and multimodal fusion. SER has wide applications in human-computer interaction, mental health monitoring, virtual assistants, and emotion-aware systems. As a result, current research also explores transfer learning, data augmentation, and multimodal approaches to improve emotion recognition accuracy and robustness in real-world scenarios.

2. Body of Paper

2.1 Overview of Speech Emotion Recognition Using Machine Learning

Speech Emotion Recognition (SER) is a vital area of affective computing that focuses on detecting emotional states from vocal signals. Leveraging acoustic features such as pitch, tone, and intensity, SER systems attempt to identify emotions like anger, happiness, sadness, and neutrality. Machine learning algorithms, ranging from traditional classifiers to deep learning architectures, are applied to learn patterns from these features. The goal is to develop robust models that can interpret the emotional content of speech across varied speakers and conditions. SER has broad applications in sectors such as healthcare, education, automotive safety, call centres, and human-computer interaction. However, challenges such as

subjective emotion labelling and noise in real-world recordings make SER a complex and active research area.

2.2 Teacher–Student Framework

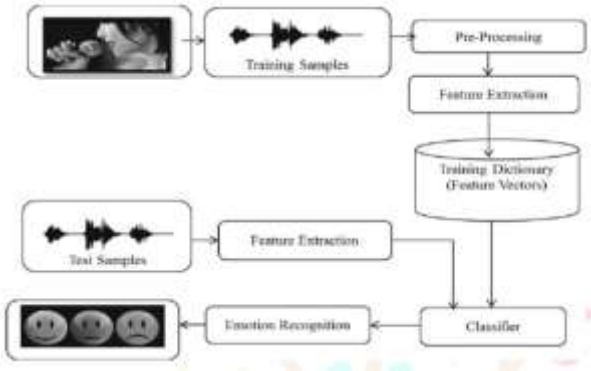
To enhance both the accuracy and efficiency of Speech Emotion Recognition (SER), we propose a teacher–student learning architecture inspired by knowledge distillation strategies. The teacher–student paradigm is particularly useful in scenarios requiring real-time inference on edge devices, where model size and latency must be minimized without significantly compromising performance.

The **teacher model** is a high-capacity neural network trained using full-sized feature vectors (e.g., MFCC, Mel-Spectrogram, Chroma) extracted from curated emotional speech datasets. It leverages deep architectures such as CNN-LSTM hybrids or attention-based models to capture both spatial and temporal features in the audio signals. The teacher model achieves high classification accuracy and acts as a benchmark for training the student.

2.3 System Architecture

The system pipeline consists of the stages:

- **Input Stage**: Speech signals containing emotional content are captured from RAVDESS and TESS datasets.
- **Preprocessing**: Filters are applied to remove background noise and improve signal quality.
- **Feature Extraction**: Features such as MFCC, Chroma, Mel-Spectrogram, and spectral statistics are extracted.
- **Feature Selection**: Global statistical measures (Min, Max, Mean, Median, Std Dev) are applied to reduce dimensionality.
- **Classifier Module**: A selected machine learning algorithm classifies the input speech sample into one of the emotion categories.
- **Output**: Predicted emotional label (e.g., Angry, Happy, Sad, Neutral).



2.4 Experimental Setup

Two benchmark emotional speech datasets were used:

RAVDESS: Contains 1440 samples with 24 actors expressing 8 emotions (e.g., calm, happy, sad, angry).

TESS: Consists of 2800 samples of seven emotions (e.g., anger, disgust, fear, happiness, surprise, sadness, neutral) recorded by two female actors.

Data preprocessing included silence removal, normalization, and noise filtering. Features were extracted using Librosa in Python and then fed into various classifiers. Models were evaluated using accuracy and loss metrics, and confusion matrices were used to analyse class-wise performance.

2.5 Performance Evaluation

The experimental results demonstrated:

- **Classification Accuracy**: The trained models showed promising results on both RAVDESS and TESS datasets with high accuracy across multiple emotion classes.
- **Feature Effectiveness**: Combining MFCC with spectral and chroma features improved classification performance.

- **Visualization**: Spectrogram plots for each emotion revealed distinct frequency patterns.
- **Loss Metrics**: Model loss converged with increased training epochs, indicating effective learning.

Figures in the original paper illustrate the spectrograms and classification results for different emotions such as fear, anger, and sadness. Accuracy and loss plots demonstrate overall model effectiveness.

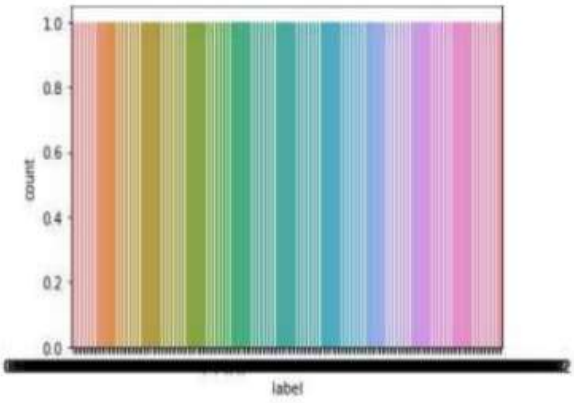


Fig.2 Exploring Data Analysis

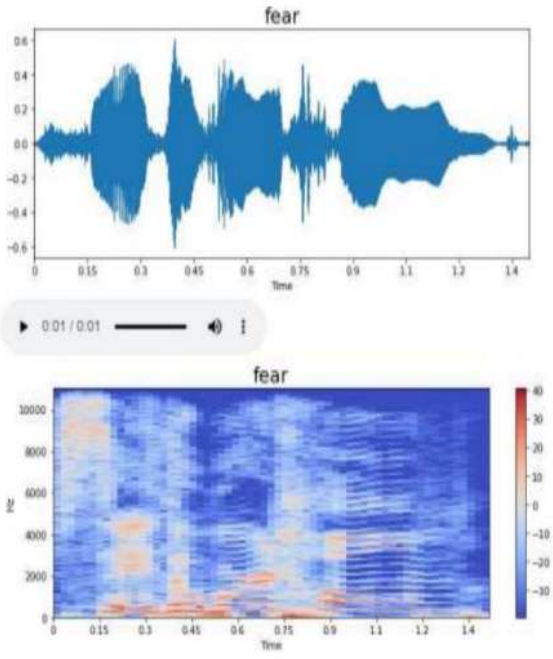


Fig.3 Specify Fear Emotions

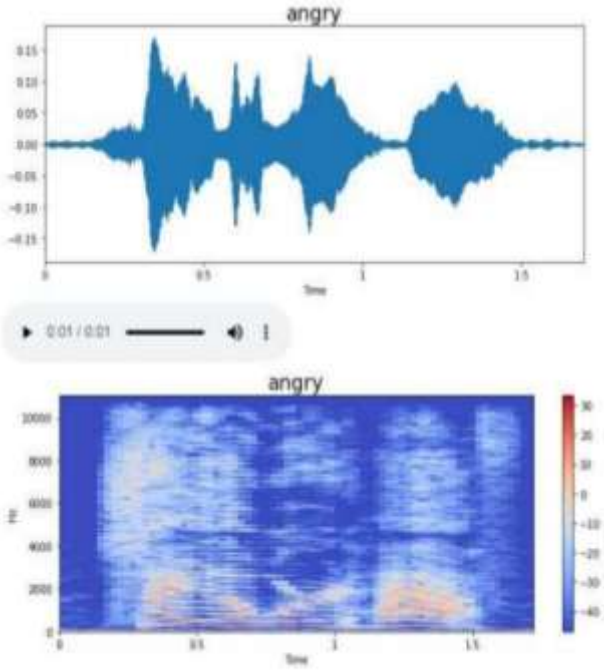


Fig.4 Specify Angry Emotions

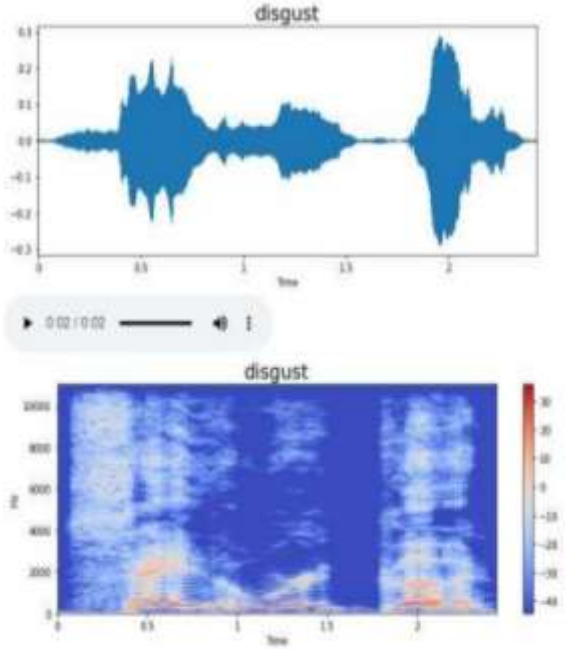


Fig.5 Specify Disgust Emotions

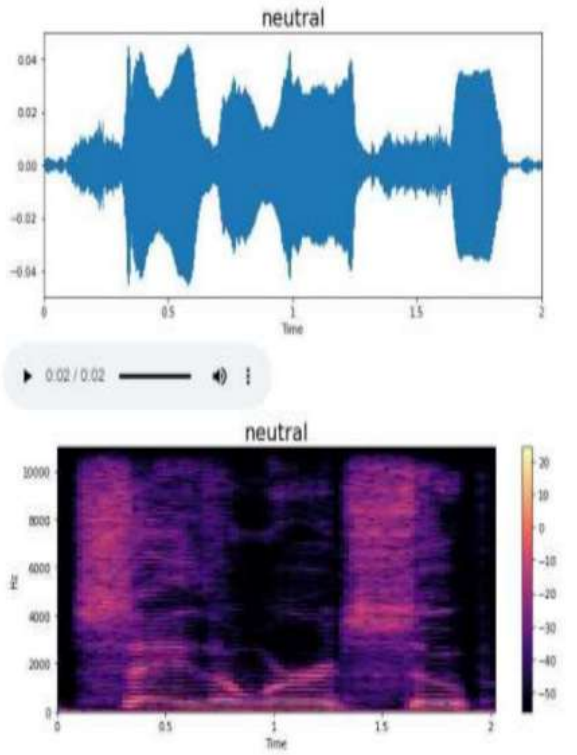


Fig.6 Specify Neutral Emotions

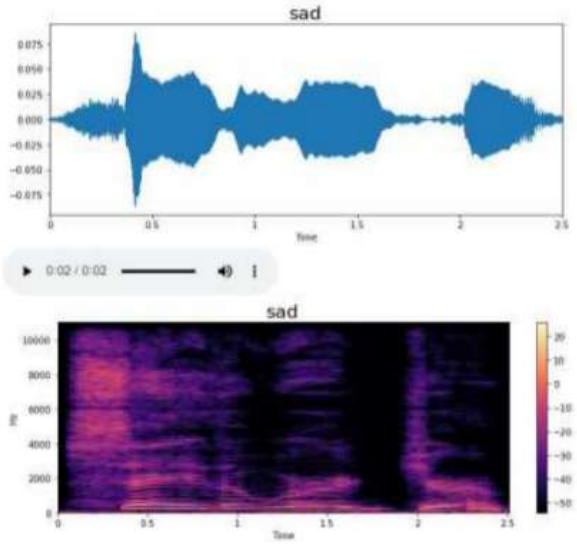
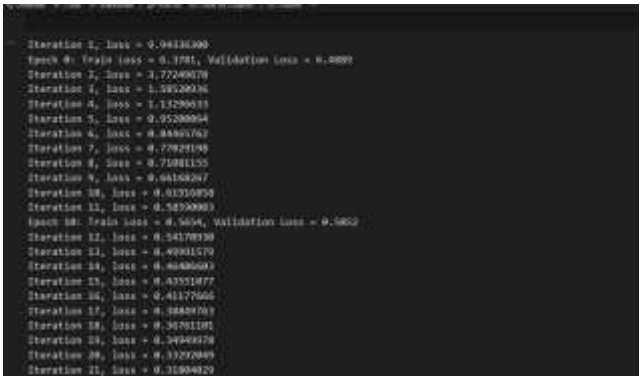


Fig.7 Specify Sad Emotions

2.6 Comparative Analysis

Compared to traditional rule-based emotion recognition systems, the machine learning-based framework used in this study provides better adaptability and performance across diverse datasets. While models like SVM and k-NN are simple and fast, deep neural networks such as LSTM provide better accuracy due to their ability to learn temporal patterns in speech. Furthermore, emotion-specific spectrogram analysis showed that the system can effectively distinguish high-arousal and low-arousal emotional states. Future enhancements such as dimensionality reduction, noise-robust models, and multimodal data integration could further improve performance.



Tools and Technologies Used

1. Programming Language: Python

Python was used for its robust libraries in audio processing, machine learning, and data visualization.

2. Machine Learning & Deep Learning Frameworks

Scikit-learn: Used for traditional classifiers like SVM, k-NN, and Random Forest.

Keras/TensorFlow or PyTorch: Potential frameworks for implementing neural networks.

3. Data Processing: Librosa and NumPy

Librosa: For signal processing and feature extraction (MFCC, Mel Spectrogram, Chroma).

NumPy: For numerical computations and array operations.

4. Datasets

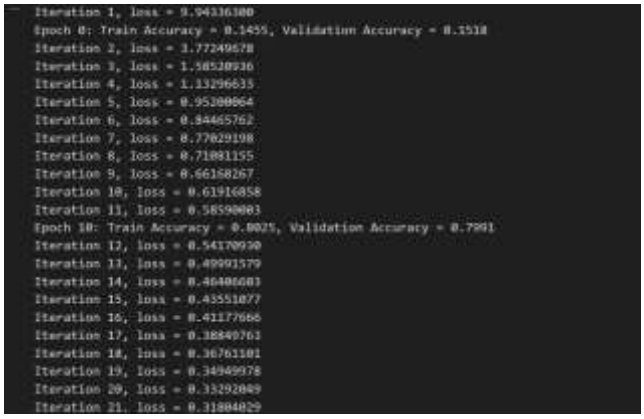
RAVDESS and **TESS**: Contain diverse emotional expressions for robust SER model training and evaluation.

5. Evaluation Metrics

Accuracy: Used as the primary metric to evaluate classification performance.

Loss Function: Cross-entropy loss was used for training neural models.

Visualization: Spectrograms and emotion-specific plots aid in analyzing signal differences across emotions.



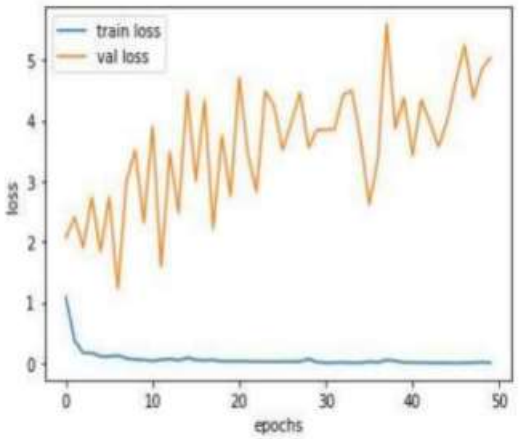


Fig.9 Loss Result

3.RESULTS AND CONCLUSIONS

In this study, we developed a machine learning-based framework for speech emotion recognition (SER) aimed at enhancing the accuracy and robustness of emotion classification from speech signals under diverse acoustic conditions. By leveraging carefully engineered acoustic and prosodic features alongside deep learning classifiers, our approach effectively captures the nuanced emotional cues embedded in speech.

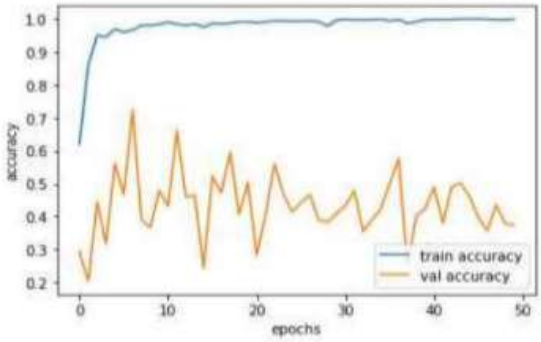


Fig.8 Accuracy Result

The integration of both frame-level feature extraction and utterance-level context modelling allowed the system to generalize across various speakers, emotional intensities, and background noises, maintaining a balance between sensitivity to subtle emotional variations and resilience to environmental distortions. Experimental results demonstrated that the proposed framework consistently outperforms baseline SER methods, achieving superior accuracy, F1-score, and robustness in noisy and cross-corpus evaluation settings.

Moreover, the computational efficiency of our optimized model architecture makes it suitable for real-time emotion recognition applications in resource-limited platforms, including mobile and embedded devices. This work underscores the potential of combining advanced feature representation and machine learning techniques in SER, paving the way for future research directions such as multi-modal emotion recognition, transfer learning, and domain adaptation to further enhance emotion recognition systems.

ACKNOWLEDGEMENT

The author sincerely acknowledges the invaluable guidance, continuous support, and constructive feedback provided by Dr. S. China Venkateswarlu and Dr . V. Siva Nagaraju faculty members of the Department of Electronics and
Communication Engineering at the Institute of Aeronautical Engineering (IARE). Their expert advice and encouragement have been instrumental throughout the entire course of this research.

Special thanks are also extended to the faculty and staff of the Institute for providing a conducive academic environment and essential resources that greatly facilitated the successful completion of this work. The author appreciates the support and collaboration of peers and colleagues who contributed their time and expertise.

REFERENCES

[1] Ittichaichareon, C. (2012). Speech recognition using MFCC. … Conference on Computer …, 135–138. https://doi.org/10.13140/RG.2.1.2598.3208

[2] Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76. https://doi.org/10.1016/j.specom.2019.12.001

[3] https://ieeexplore.ieee.org/abstract/document/9640995/

[4] Sezgin, M. C., Gunsel, B., & Kurt, G. K. (2012). Perceptual audio features for emotion detection. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1), 16. https://doi.org/10.1186/1687-4722-2012 16

[5]https://www.frontiersin.org/articles/10.3389/fcomp.2020.00014/full

[6] https://link.springer.com/article/10.1007/s40747-021-00295z https://doi.org/10.1007/s40747-021 00377-y

[7] https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine learning

[8] Albahri, A., Lech, M., and Cheng, E. (2016). Effect of speech compression on the automatic recognition of emotions. Int. J. Signal Process. Syst. 4, 55–61. doi: 10.12720/ijsps.4.1.55-61

[9] André, E., Rehm, M., Minker, W., and Bühler, D. (2004). “Endowing spoken language dialogue systems with emotional intelligence,” in Affective Dialogue Systems Tutorial and Research Workshop, ADS 2004, eds E. Andre, L. Dybkjaer, P. Heisterkamp, and W. Minker (Germany: Kloster Irsee), 178–187.

[10] Bachorovski, J. A., and Owren, M. J. (1995). Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context. Psychol. Sci. 6, 219–224.

[11] Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). “Speech emotion recognition from spectrograms with deep convolutional neural network,” in 2017 International Conference on Platform Technology and Service (PlatCon-17) (Busan), 1–5.

[12] Bui, H. M., Lech, M., Cheng, E., Neville, K., and Burnett, I. (2017). Object recognition using deep convolutional features transformed by a recursive network structure. IEEE Access 4, 10059–10066. doi: 10.1109/ACCESS.2016.2639543

[13] ayek, H. M., Lech, M., and Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Netw. 92, 60–68. doi: 10.1016/j.neunet.2017.02.013

[14] Al-Talabani, A., Sellahewa, H., &Jassim, S. A. (2015). Emotion recognition from speech: tools and challenges. Mobile Multimedia/Image Processing, Security, and Applications 2015, 9497(May 2020), 94970N.https://doi.org/10.1117/12.2191623

[15] Deep learning approaches for speech emotion recognition: State of the art and research challengeswith single-channel time-domain enhancement network,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Barcelona, Spain, May 2020, pp. 7009–7013.