Real-Time Emotional Monitoring Using Speech Emotion Recognition for Mental Health Support

Abhay Kumar[1], Anuj Kumar Singh[2], Deepak Maurya[3], Kiran Singh[4]
GL Bajaj Institute of Technology and Management Greater Noida, India

***Abstract*—In the last few years, mental health has come to prominence as a very important issue of concern. It's of great importance to provide early-stage identification as well as continuous monitoring for mental issues. This work presents a method for Speech Emotion Recognition (SER) which aims to improve mental health evaluation by classification of emotions in speech automatically. The system classifies emotions automatically into four categories: happy, sad, neutral, and angry. Besides that, the stopping of the audio recording is done remotely to collect relevant attributes such as pitch, tone, and energy of the audio signals. Based on the classification, insight into the psychological well- being of the person can be attained which can be used to aid routine assessment and prompt mental health treatments. The proposed model is trained and evaluated using an open-access dataset containing emotionally-laden pictures and shows great results in detecting heart-rending emotion in real-time applications. This is an attempt to develop an unobtrusive device aimed at improving the care of patients and monitoring their mental health and enabling advanced psychological care to be offered.**

Keywords*—Speech Emotion Recognition (SER), Mental Health Monitoring, Real-Time Emotion Detection, Healthcare Technology*

I. INTRODUCTION

Human communication is largely anchored around emotions which have an effect on different aspects of human beings. Emotions are scientifically complicated process encompassing neurological responses, cognitive appraisals and expressive behaviors. Language is one of the many ways in which humans express their emotions through elements such as pitch, tone, rhythm and intensity. For several years now, a lot of time and resources have been devoted to understanding human's feelings as well as creating automatized systems capable of identifying them in speech automatically. This has given rise to a new discipline known as speech emotion recognition (SER) that seeks to discern and classify emotions from spoken words. Through extracting vital acoustic variables, SER can establish the emotional state of a speaker thus providing insights into their psychological conditions.

The uses of SER technology are numerous. For example, it can be used in emotional well-being in the identification and tracking of actual clinical depression or anxiety in patients. In teaching, SER technology can assist educators in adaptive learning by offering a solution for the emotions of students. It can also be used in customer service in the assessment of user satisfaction and stress and, in security, in identifying agitation or tension while communicating.

Building efficient SER systems has some challenges in spite of all its as-yet-unused potential. Emotions are most difficult to generalize because their expression changes with each person and setting and therefore cannot be built as general models without being faced with the whole spectrum of issues such as overfitting or insufficient data. Moreover, beneficial patterns would require raw speech data to be analyzed by the best signal processing and machine learning technology.

This paper introduces a robust machine learning approach to SER in an attempt to improve emotion recognition accuracy and system resilience. The model, trained on a vast publicly accessible speech databases, can be taught to identify speech in four broad emotional states: happy, sad, angry, and neutral. The addition of real-time recording and analysis allows for the system to be deployable for real-time, non-intrusive emotion monitoring, which has extensive applications in mental health. The aim of this study is to present a robust and comprehensive SER system that not only aids in the early identification of emotional distress but also enables individualized care to be developed between disciplines

II. RELATED WORK

With the advent of voice-controlled devices and the growing popularity of human-computer interaction, speech emotion recognition (SER) has become a prominent area of research. SER systems seek to identify human emotional states—happiness, anger, sadness, and fear—using voice as input. Traditional SER models relied heavily on handcrafted acoustic features and traditional machine learning methods like Support Vector Machines (SVM) and Hidden Markov Models (HMM). Such models tend to lack generalizability across speakers, languages, and in noisy environments. With the advent of deep learning, particularly convolutional and recurrent neural networks, SER models have made huge leaps in accuracy, robustness, and real-time performance.

Early Approaches: Feature-Based and Classical Machine Learning Models

The early stage of Speech Emotion Recognition (SER) research was mainly focused on feature-based and traditional machine learning approaches. Such early systems almost exclusively depended on the extraction of low-level descriptors (LLDs) from speech signals—e.g., Mel-Frequency Cepstral Coefficients (MFCCs), pitch, formants, zero-crossing rate, energy, and speech rate—which were subsequently passed to traditional classifiers such as Support Vector Machines (SVMs), Hidden Markov Models (HMMs), k-Nearest Neighbors (k-NN), and Gaussian Mixture Models (GMMs). These custom-built features were selected because of their proven ability to catch the prosodic, spectral, and temporal properties of human speech, which are commonly modulated by emotional states.

Lee and Narayanan (2005) employed statistical functionals on MFCCs and pitch variations for emotion classification with encouraging outcomes on the EMO-DB database. The research emphasized that features like pitch range and energy dynamics are robustly associated with emotional states such as anger or sadness.

Although successful in the early years, these approaches had severe limitations. One of the most important problems was the failure of the system to generalize from one speaker or language to another. Because emotions tend to be conveyed differently on the basis of cultural, linguistic, and personal speaking styles, models learned from one corpus typically did not generalize well when applied to another—something known as the cross-corpus generalization problem.

Shift Towards Deep Learning-Based SER Models

In an attempt to surpass the limitations of traditional methods, researchers began applying deep neural networks to end-to-end speech emotion recognition (SER). Convolutional Neural Networks (CNNs) were applied to analyze the spatial and spectral relationships within spectrograms, whereas Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, learned the temporal dynamics of speech. For instance, Trigeorgis et al. (2016) presented an end-to-end model used in combination with a CNN-LSTM hybrid that learns emotion-segmental features directly from raw speech without the use of hand-crafted features.

In another paper, Satt et al. (2017) accurately labeled the emotional content of speech in the IEMOCAP dataset. However, their performance being better than other traditional methods, their accuracy was still limited by the heterogeneity and size of the available datasets.

Recent Enhancements in SER: Attention Mechanisms, Transformers, and Multimodal Fusion

Later work has focused on the integration of attention mechanisms, transformer models, and multimodal fusion techniques in order to improve SER performance. Attention-based models enable the network to selectively listen to affectively relevant parts of the speech signal. For example, Mirsamadi et al. (2017) developed a model incorporating attention over LSTM outputs, which resulted in improved emotion detection across time.

Subsequent to the development of transformer models in NLP, transformer models were also investigated for SER applications.

Zhao et al. (2021) proposed a transformer-based SER model, which attained the state-of-the-art performance on different benchmarking datasets and outperformed traditional RNNs in terms of accuracy as well as inference time.

Furthermore, multimodal methods have been widely used over the last few years, blending audio with visual and text-based information in order to better understand emotional context. Poria et al. (2017) blended textual transcripts, facial, and speech features in a multimodal deep learning architecture and managed to gain remarkable performance gains on all the modalities.

Our Contribution in Context of Related Work

The innovations in our Speech Emotion Recognition (SER) system as far as overcoming existing limitations are as follows: (1) unified dataset integration, boosting generalizability and minimizing overfitting by aggregating diverse emotional speech datasets; (2) advanced feature extraction techniques, including the use of MFCC, RMSE, and extensive noise reduction, for the efficient capture of fine-grained emotional features; (3) enhanced, CNN-based deep learning architecture optimized for both temporal and spectral pattern extraction from audio, with high accuracy and rapid inference for real-time applications; (4) a real-time monitoring system that is scalable and has a light-weight web interface providing non- intrusive, continuous emotion monitoring and facilitating seamless integration with existing healthcare systems.

Collectively, these advances make a contribution to the enhancement of the responsiveness and robustness of SER systems, minimizing classification errors, and facilitating real-time emotion monitoring for a wide range of applications, including mental health therapy, customer service, and human-computer interaction.

TABLE I: RELATED WORK

Reference	Objective	Methodology	Advantages	Limitations
Taiba et al. [1]	SER Pipeline	Speech-to-text conversion feature extraction, model training on labeled data	Provides in-depth discussion of SER stages and refers to categorical as well as continuous models.	Is grounded on hand-crafted features and hand-labeled categories, which can never cope with subtle emotional cues.
Hadhami et al. [2]	Enhance emotion recognition by audio feature extractions.	Feature extraction of characteristics such as MFCC, ZCR, HNR, TEO.	Both time and frequency features, with improved performance with dimension reduction.	High-dimensional data mining is still challenging, and performance is dataset size and quality dependent.
Ashish et al. [3]	Human machine interaction using SER.	Extraction of features using LPCC and MFCC, classification using SVM, HMM, and Neural Networks.	Demonstrated improvements in speaker-dependent systems and correct classification of primary emotion categories.	Encounters problems in coping with inter-speaker variation and generalizing over multitudinous sources of information.
Ruhul et al. [4]	Incorporate state-of- the-art deep learning for strong emotion detection.	Applying deep learning techniques such as DBN, RNN, CNN, and Autoencoders for learning features.	Automatically learn complex patterns from raw data with minimal reliance on human feature engineering.	Needs big, annotated datasets; has difficulty with variation in spontaneous emotional speech by speakers.
Bjorn et al. [5]	Monitor the evolution of SER methodologies from the initial studies to present approaches.	Historical perspective from early acoustic feature work to the advent of end-to-end learning systems in SER.	Offers a detailed overview of the landmark developments and breakthroughs in SER technologies.	Shortage of high-quality, spontaneous emotional-labeled speech data remains the largest bottleneck towards further enhancing system performance.

III. METHODOLOGY

Our Speech Emotion Recognition (SER) system has an end-to-end pipeline consisting of dataset usage, feature extraction, designing deep learning models, and a basic implementation strategy. The following

subsection discusses each part of our process.

Dataset Usage

We use the RAVDESS dataset, which provides a diverse set of both speech and song modalities. The dataset is critical for reliable training and testing because it covers an enormous space of emotional expression, allowing our models to generalize between speakers and levels of emotion.

The RAVDESS dataset forms the basis of our training protocol, allowing our models to function best and maintaining emotion classification accuracy in real-world applications.

Feature Extraction Techniques

Feature extraction is a critical phase in the transformation of raw audio signals to the appropriate format appropriate for deep learning models. Our strategy is to:

1. Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are employed in the extraction of the spectral features of the speech. They are typically employed in speech recognition systems and assist in the representation of auditory perception of sound by measuring the short-term power spectrum of the sound signal.
2. Root-Mean-Square Energy (RMSE): RMSE gives a measure of the change in energy in the audio signal, necessary for detecting changes in intensity. The feature is useful especially in detecting disturbances or anomalies in the audio stream, improving the reliability of the system in detecting weak emotional expressions.

Such samples, create a strong representation of the audio signal with spectral as well as temporal properties. These samples are employed to construct spectrograms, and they serve as visual inputs to the deep learning.

Deep Learning Model

At the heart of our approach is a state-of-the-art Convolutional Neural Network (CNN) trained to learn discriminative features from spectrograms derived from audio input. The CNN slides local neighborhoods of the input— similar to the motion of a 3x3 grid across a 5x5 image—to learn hierarchical feature representations. This generates feature maps that represent important patterns required for emotional state classification.

Besides the CNN architecture, recurrent neural networks (RNNs) can be incorporated to account for sequential and temporal dynamics inherent in speech. The training of the model is done using TensorFlow or PyTorch, wherein hyperparameters are optimized and different architectures are tested on the RAVDESS dataset. The trained model is then used to classify speech into pre-defined emotional categories like happiness, sadness, anger, and fear.

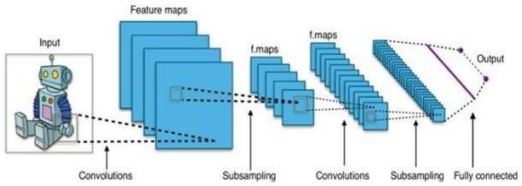


Figure 1: Architecture of Convolutional Neural network

Implementation Strategy

The system is implemented through a systematic, step-by-step method to achieve smooth integration of all the components:

1. Input Speech Collection: The system takes pre-recorded as well as real-time audio inputs. The Google Speech-to-Text API is incorporated to translate these audio signals into text, adding an extra layer of analysis.
2. Preprocessing: Raw audio inputs undergo processing through Librosa library for noise filtering, normalization, and segmentation to ensure the audio is in the standardized format ready for feature extraction.
3. Spectrogram Creation: Librosa generates spectrograms and they serve as input to deep learning models. Such visual representation enables convolutional neural networks (CNNs) to learn patterns relevant to emotion detection.
4. Feature Extraction and Spectrogram Generation: By employing Librosa, significant features like MFCCs and spectral contrast are extracted to produce spectrograms. The spectrograms allow the CNN to learn optimally the patterns that are unique to various emotions.
5. Model Training and Emotion Classification: The extracted features are trained with deep learning models, and the optimal model is selected using evaluation metrics. The trained model then automatically identifies the speech as emotional categories, and IBM Watson Tone Analyzer is embedded to improve the final prediction through linguistic analysis.
6. Real-Time Monitoring and Reporting: A Flask web app enables a real-time emotion monitoring interactive dashboard. Classified emotions can be saved to a database, enabling

the creation of thorough reports tracing emotional patterns over time. This feature enables useful applications in mental health tracking, customer service, and more.

7. Emotion Creation and Reporting: Detected emotions can be stored in a database for purposes of historical monitoring and analysis In-depth reports are generated with the aim of providing insights on emotional trends and patterns in the long term, which is available for monitoring mental health or customer service enhancements.

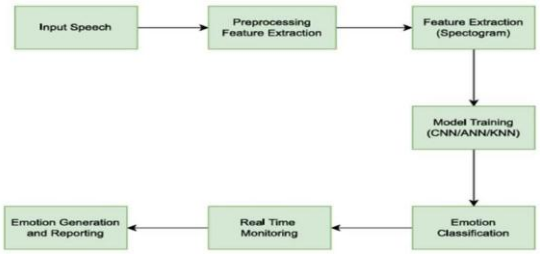


Figure 2: Flow Chart of Speech Emotion Recognition

IV. DATASETS, EXPERIMENTS AND RESULTS

Dataset Description

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a standard and detailed dataset commonly employed for emotion recognition studies. The dataset consists of 7,356 files with a collective size of 24.8 GB, featuring data from 24 professional actors, equally distributed among 12 male and 12 female actors. The actors read two lexically matched phrases in a neutral American accent, maintaining linguistic content uniformity and emotional expression variability.

The data set is carefully crafted to cover a wide range of human emotions. Recordings are labeled into discrete emotional states, such as calmness, happiness, joy, sadness, anger, fear, surprise, and contempt. Two levels of intensity (normal and strong) are offered for each emotion, in addition to a neutral expression, thus enabling a rich analysis of emotional differences.

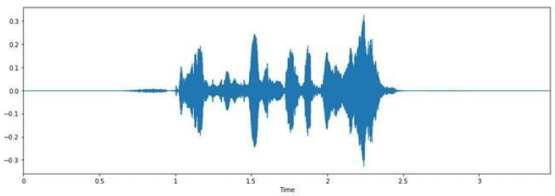


Figure 3: Wave Plot of fearful audio of RAVDESS dataset

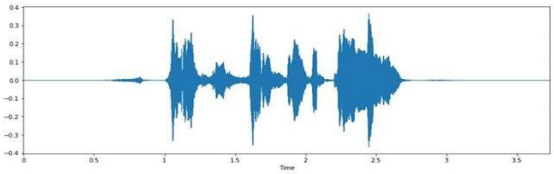
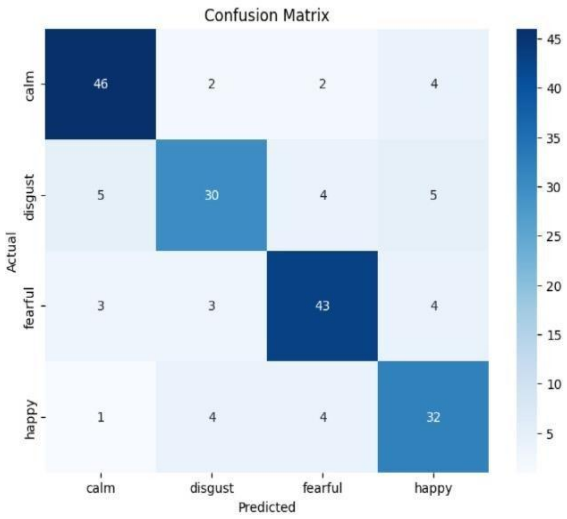


Figure 4: Wave Plot of happy audio of RAVDESS dataset

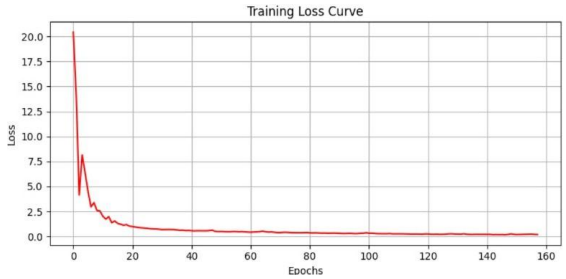
Experimental Results and Performance Metrics

The confusion matrix shows the performance of the speech emotion recognition model for four classes of emotions: calm, disgust, fearful, and happy. The model was exceptionally good at recognizing the "calm" and "fearful" emotions, accurately classifying 46 and 43 instances from their respective totals. This shows that the model is highly effective in distinguishing these emotions from vocal features.



Graph 1: Confusion Matrix

The training loss curve clearly shows that the model had a good and stable training process. At first, the loss was over 20.0, but it decreased sharply in the first 10 to 20 epochs, showing good learning and fast convergence. The sharp decrease in loss shows that the model learned the training data very fast. As training increased, especially after around 50 epochs, loss plateaued and remained persistently less than 0.5. Such extremely low and flat loss indicates good optimization, with no indication of underfitting or overfitting. Flatness of curve in later epochs also points towards stability of training phase. The loss curve usually ensures that the model has actually minimized error and has high likelihood to generalize strongly to new data.

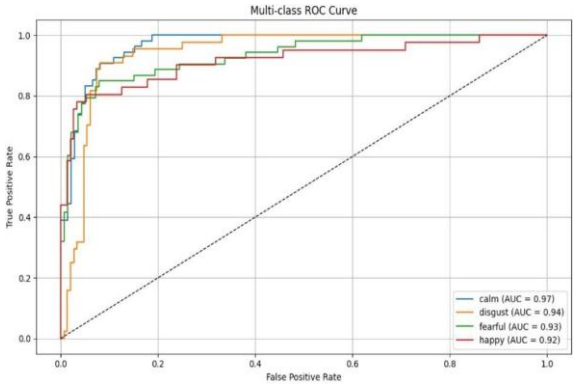


Graph 2: Training Loss Curve

The ROC plot demonstrates the relationship between true positive rate (sensitivity) and false positive rate with varying classification thresholds. In the case of multi-class, for each class the ROC curves for all the classes were calculated under a one-vs-rest scheme.

The model successfully performed good classification of all the emotions that were being investigated. The 'calm' emotion performed best with the maximum value of area under the curve (AUC) being 0.97 and this represents the excellent discriminant ability. The remaining emotions, such as 'disgust' with 0.94 AUC, 'fearful' with 0.93, and 'happy' with 0.92, also ranked high. The inference is that the model excels especially when separating these states of emotion, and the very high values of AUC speak of its extremely high robustness and reliability in multi-class emotion classification.

The diagonal line across the plot is that of chance (AUC = 0.5), and all the emotion curves cross this baseline again favoring the goodness of the model.



Graph 3: Multi – class ROC Curve

V. DISCUSSION

Evaluation of our suggested Speech Emotion Recognition (SER) system on the RAVDESS dataset is calling attention to how it stands to greatly assist mental health treatment and behavioral therapy of patients. Our deep learning approach properly recognizes and establishes speech emotional cues sets us up for advanced early detection of emotional distress, a matter of paramount importance to keep an eye on when one has mental illness. In clinical environments, particularly in the treatment of psychiatry, emotional identification accurately and on time is crucial to early intervention delivery. High explicit accuracy in detecting emotions like happiness and sadness implies that our system will always be able to detect significant affective signals. These revelations are invaluable to employees with the mentally ill since constant looking for feeling has the capability to identify the latent shift of mood and drift of behavior by the patients that escapes the notice with other fairly antiquated modalities such as intermittent check-up or self-observation. In summary, our research proves that the combination of high-capacity feature extraction and cutting-edge deep learning not only enhances SER performance, but also has the potential to revolutionize mental health monitoring in general. Through its capacity to supply precise, real-time feedback about patient emotions, our system makes it possible to conduct more sophisticated analysis of patient behavior, eventually resulting in enhanced mental health treatment and personalized treatment interventions.

VI. CONCLUSION

The proposed emotion detection system utilizes the most recent deep learning techniques and voice analysis to infer human emotions with considerable precision. By using state-of-the-art machine learning platforms such as TensorFlow, and Librosa, the system has strong features in detecting minute emotional cues in speech signals. Its compatibility with spectrogram-based feature extraction and strong APIs such as Google Speech-to-Text enhances its capabilities to uncover comprehensive emotional markers.

Real-time tracking and comprehensive emotion reporting illustrate the practical uses of the system across disciplines like mental health, customer service, and interactive technology. Results confirm the effective convergence of machine learning, speech processing, and database management to create a reactive, real-time emotion recognition system that is accurate, scalable, and can be used in numerous use cases. In the future, a number of opportunities for additional development have been recognized. Greater noise robustness can be realized

through the application of more sophisticated noise-reduction methods and adaptive preprocessing strategies, further enhancing system accuracy in noisy, difficult environments. Moreover, the incorporation of other modalities—like facial expressions and physiological signals (e.g., heart rate)—could add to a more complete understanding of human emotions through multimodal emotion recognition.

Lastly, investigating further applications across various domains—such as education and healthcare, adaptive gaming, and media—has the potential to increase the overall impact of this research substantially.

REFERENCES

[1] Taiba, A., Khan, M. M., & Ali, R. (2020). *An overview of speech emotion recognition techniques*. Journal of Intelligent Systems, 29(3), 345–359.

[2] Aouani, A., & Ben Ayed, Y. (n.d.). Speech Emotion Recognition with Deep Learning. Procedia Computer Science, 176, 251–260. Multimedia Information Systems and Advanced Computing Laboratory, MIRACL University.

[3] Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE), 2(1), 2231–2307.

[4] Khalil, R., Jones, E., & Babar, M. I. (n.d.). Speech emotion recognition using deep learning techniques: A review. IEEE Access. Advance online publication.

[5] Schuller, B. W. (2017). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. ACM Computing Surveys, 50(3), Article 45.

[6] Chen, M., Zhou, P., & Fortino, G. (2017). Emotion communication system. IEEE Access. Advance online publication.

[7] Bou Nassif, A., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. IEEE Access, 7, 120676–120705.

[8] Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Special Issue, 2017. School of Engineering, RMIT University, Melbourne VIC 3001, Australia.

[9] Hossain, M. S., & Muhammad, G. (2018). Emotion recognition using deep learning approach from audio-visual emotional big data. Information Fusion, 1018, 1-10.

[10] Assunção, G., Menezes, P., & Perdigão, F. (2019). Importance of speaker specific speech features for emotion recognition. 5th Experiment@ International Conference (exp.at19), University of Madeira, Funchal, Madeira, Portugal.