IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

Voice-Based Emotion Recognition Using Machine Learning Classifiers

Trupti Kalokhe and Mrs.Nanda Kulkarni
Department of Computer Engineering,
Siddhant College of Engineering, Sudumbre, Pune, India
trupti.bhase18@gmail.com

Abstract: *Emotion recognition from speech has emerged as a pivotal research area in artificial intelligence, enabling more empathetic and context-aware human-computer interactions. This paper presents the implementation of a machine learning-based system for real-time emotion recognition using voice data. The proposed system utilizes Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) as primary feature extraction techniques to capture critical emotional cues embedded in speech, such as pitch, tone, and spectral characteristics. These features are fed into various supervised learning models, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB). The models were trained and evaluated using benchmark emotional speech datasets, ensuring robustness through preprocessing, data normalization, and cross-validation. Performance metrics such as accuracy, precision, recall, and F1-score were computed to assess each model's efficacy. Experimental results demonstrate that ensemble-based models, particularly Random Forest and Gradient Boosting, outperform others, with Random Forest achieving the highest accuracy of 95.5%. Even with overlapping speech patterns and mild background noise, the system can distinguish emotions. This implementation showcases the feasibility of deploying lightweight, real-time speech emotion recognition systems in practical scenarios such as customer service, mental health monitoring, and intelligent virtual assistants. The study highlights the significance of choosing suitable feature extraction techniques and classifiers while addressing challenges such as emotional overlap and generalizability across diverse datasets. The proposed framework is a foundation for future work involving deep learning, multimodal integration, and deployment on edge devices..*

Keywords: Speech Emotion Recognition, MFCC, LPC, Machine Learning, Random Forest, Gradient Boosting, Voice Data, Real-Time Emotion Detection, Human-Computer Interaction, Feature Extraction

I. INTRODUCTION

Emotion is a vital component of human communication, conveying a speaker's mental state beyond the semantic content of speech. Recognizing these emotions has become a key research focus within the field of affective computing, particularly through speech-based emotion recognition. The goal of Speech Emotion Recognition (SER) systems is to identify and classify emotional states by analyzing the acoustic characteristics of human voice, such as pitch, intensity, speaking rate, and spectral features. These systems enhance human-computer interaction (HCI), enabling machines to respond empathetically to user emotions in real time.Voice is a powerful medium for expressing emotions, making it an essential channel for emotion recognition. Emotions are embedded within prosodic and paralinguistic elements of speech, including pitch variations, loudness, speech rate, and intonation. Recognizing these elements computationally allows machines to interpret user emotions and adapt their responses accordingly. In recent years, the development of SER systems has accelerated with advancements in machine learning, deep learning, and audio processing techniques.
Traditionally, SER systems relied on handcrafted acoustic features and basic classification algorithms. Early approaches involved extracting features such as pitch, energy, zero-crossing rate (ZCR), and formants, followed by the application of classifiers like Support Vector Machines (SVMs) or Gaussian Mixture Models (GMMs). While these methods provided foundational insights, they often lacked robustness in real-world scenarios with varied speakers,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

accents, and background noise.The introduction of advanced feature extraction techniques, such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC), has improved the representation of vocal characteristics. MFCC mimics the nonlinear human ear perception of sound, effectively capturing phonetic and timbral information. LPC models the speech signal by predicting the current sample based on past samples, offering valuable information about the vocal tract. When used together, these features enhance the ability of classifiers to distinguish between different emotional states.

In this work, we implement a machine learning-based framework for emotion recognition from voice data using MFCC and LPC features. The system is evaluated using various supervised learning classifiers, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB). These classifiers are chosen due to their established efficacy in speech classification tasks and their varying complexity and interpretability. Each model is trained and tested using publicly available emotional speech datasets such as RAVDESS and the Berlin Emotional Speech Database.The system undergoes rigorous preprocessing, including noise reduction, silence removal, and normalization to ensure performance and generalizability. The datasets are divided into training, validation, and testing subsets. The extracted features are then used to train the classifiers, and the models are evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

Real-time emotion recognition is a primary objective of this study. Hence, computational efficiency and scalability are considered during system design. The final model is optimized for low latency, enabling integration into real-world applications such as intelligent virtual assistants, telemedicine platforms, automotive safety systems, and educational software.Despite the progress in SER, several challenges persist. Emotional expressions often overlap in acoustic space—for example, anger and frustration may exhibit similar tonal patterns. Furthermore, speech variability across different speakers, accents, and languages complicates recognition. The presence of environmental noise, limitations of existing datasets, and the lack of multilingual data also affect model generalization. This study addresses some of these limitations by adopting robust features, diverse datasets, and comparative analysis across multiple algorithms.

The contributions of this study are summarized as follows:

Implement a speech-based emotion recognition system using MFCC and LPC as combined features.

Evaluation and comparison of multiple machine learning classifiers under identical conditions.

Analysis of model performance using real-time compatible audio datasets.

Identification of the most effective classification algorithm, demonstrating Random Forest as the top performer.

Design a lightweight, scalable pipeline for integration .in HCI systems.

The rest of the paper is organized as follows: Section II discusses the proposed methodology, including dataset preprocessing, feature extraction, and model training. Section III presents the experimental results and comparative analysis. Section IV concludes the study and outlines future work.

II. LITERATURE SURVEY

To identify emotional states, Girija Deshmukh et al. [1] developed a speech emotion detection system that made use of critical auditory characteristics like pitch, Mel Frequency Cepstral Coefficients (MFCC), and Short-Term Energy (STE). Their study focused on three emotions: happiness, grief, and irritation. They used open-source North American English voice data to record genuine speech. These selections showed sincere feedback and emotional expression. The researchers concentrated on identifying speaker-specific vocal traits, including energy dynamics and pitch contours,and categorizing emotions. Experiments were conducted using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which was manually split into training and testing subsets. A multi-class Support Vector Machine (SVM) was used for classification. The SVM was trained using the retrieved feature vectors to create emotion-specific models. The study showed that it is possible to discern between a few emotions with a respectable degree of accuracy using traditional machine learning techniques with manually designed characteristics.

In a different strategy, Peng Shi [2] compared discrete and continuous models for speech emotion recognition to improve feature abstraction to describe emotions more effectively. According to the study, Deep Belief Networks (DBNs) are an advanced deep learning architecture that outperforms more conventional techniques like Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) by about 5% in accuracy. It has been shown that DBNs

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

enhance classification performance by extracting richer and more abstract feature representations. Due to SVM's effectiveness in low-dimensional classification tasks, DBN-SVM slightly outperformed DBN-DNN among the hybrid architectures investigated, especially in jobs requiring smaller datasets. The model's capacity to identify complicated emotional states in speech was greatly enhanced when DBNs successfully converted sparse or partial features into high-level abstract representations. This study demonstrated how deep architectures can improve feature quality and the general robustness of emotion identification systems.

In their speech emotion recognition framework, J. Umamaheswari et al. [3] used the Pattern Recognition Neural Network (PRNN) and K-Nearest Neighbors (KNN) algorithms for preprocessing. The Gray Level Co-occurrence Matrix (GLCM) and Mel Frequency Cepstral Coefficients (MFCC) were used in the study's multi-level feature extraction method to extract spectral and textural characteristics from speech signals. KNN and PRNN were then used to classify the extracted features, and the outcomes were compared to standard benchmark algorithms like Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). According to experimental data, the suggested method outperformed the conventional precision, accuracy, and F-measure models. The system successfully took advantage of the finding that emotional speech signals display recognizable patterns.

M.S.Likitha et al. [4] emphasized that emotion recognition requires thorough analysis of the speech waveform, relying on training data that includes characteristics such as sound, formants, and phonemes. The study reviewed various feature extraction methods, an essential step involving compressing a speech signal's informative components into a representative feature set. These features are then used to classify the speaker's emotional state. Among the available methods, MFCC remains the most widely adopted due to its efficiency in representing the acoustic structure of speech. The study described feature extraction as distilling essential information from speech to enable accurate emotional classification, underscoring its critical role in speech emotion recognition pipelines.

In a more application-specific study, Zhang Lin et al. [5] explored the use of speech emotion recognition (SER) for monitoring driver emotions, especially in emergencies where abnormal emotional states such as panic may arise. The system was integrated with voice-guided parking instructions, using emotion recognition to detect urgent conditions based on the driver's vocal tone. The approach extracted three categories of features: prosodic, spectral-based, and voice quality features, with particular emphasis on MFCC and Linear Predictive Cepstral Coefficients (LPCC). Support Vector Machine (SVM) was employed for classification. However, the authors noted a significant drop in recognition accuracy when applying the model to emotionally charged voice samples, mainly because the parking voice guidance and emotional speech datasets were collected under different acoustic environments. This performance degradation highlighted the sensitivity of SER systems to domain variation, emphasizing the importance of consistent dataset conditions for real-world deployment.

Asaf Varol et al. [6] provided a foundational perspective on how sound, characterized as a pressure wave generated by molecular vibrations, plays a crucial role in speech emotion recognition (SER). Their research explored the nature of sound energy and its intrinsic properties, emphasizing its applicability in emotion analysis. The study leveraged speech signal spectrograms in combination with Artificial Neural Networks (ANNs) to derive more accurate recognition results. Using the EMO-DB (Berlin Emotional Speech Database), the system employed techniques such as acoustic feature extraction and spectrogram analysis to classify emotional states. The authors also highlighted the emerging relevance of SER in multiple interdisciplinary domains, including signal processing, pattern recognition, and even psychiatric assessments.A key point in the study was the importance of diverse machine learning approaches to maximize recognition success rates. The authors emphasized that varying datasets and experimental conditions demand different algorithmic strategies for optimal performance, advocating for model adaptability to address real-world variability in emotional speech data.

In a related study, Abhijit Mohanta et al. [7] analyzed the acoustic generation characteristics of emotions such as anger, fear, happiness, and neutrality within emotional speech signals. Rather than performing classification, their work examined how these emotions influence specific speech features. The study introduced sub-segmental features, such as loudness, voiced region detection, and excitation energy, significantly representing emotional speech.Advanced signal processing techniques were employed, including Zero Frequency Filtering (ZFF) to extract instantaneous fundamental frequency (F0), and Short-Term Energy (STE) and Zero-Crossing Rate (ZCR) for identifying temporal patterns and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

energy variations. The study also explored other critical features such as formant frequencies and dominant spectral components, offering insights into how these parameters change across emotional expressions. The work provides a detailed acoustic profile for each emotion type, laying a solid groundwork for future classification-based models and emotion synthesis applications.

Edward Jones et al. [8] emphasized Speech Emotion Recognition (SER) as a vital and evolving aspect of Human-Computer Interaction (HCI). According to their study, the two primary components of SER systems are feature extraction and feature classification. They highlighted using both linear and non-linear classifiers for the classification stage. Among linear classifiers, Support Vector Machines (SVMs) and Bayesian Networks (BNs) are commonly applied due to their robustness in handling the variability and dynamic nature of speech signals. While traditional classifiers offer reasonable performance, the authors noted that deep learning techniques present substantial advantages. Deep models can automatically extract and learn complex hierarchical features from raw audio data, eliminating the need for manual feature engineering and improving the system's capacity to detect subtle emotional cues.

In a complementary approach, Michael Neumann et al. [9] explored the utility of unsupervised learning in enhancing SER. Their research demonstrated that learning from unlabeled voice data can contribute meaningfully to emotion recognition. The study employed t-distributed stochastic neighbor embedding (t-SNE) for visualizing high-dimensional speech representations. However, no distinct emotion-based clusters emerged in the 2D plots, indicating the limitations of such projections. However, autoencoders trained on large, unlabeled datasets showed incremental improvements in SER accuracy, suggesting their effectiveness in representation learning. The authors also suggested exploring generative adversarial networks (GANs) and other variants of autoencoders as promising directions for future SER research.

Radim Burget et al. [10] used the Berlin Database of Emotional Speech, consisting of over 250 emotion-labeled recordings. Each recording was segmented into non-overlapping 20-ms intervals, and 3098 silent segments were discarded using the Google WebRTC voice activity detector during preprocessing. The resulting data was normalized to zero mean and unit variance before being split into training, validation, and test sets. The input to their Deep Neural Network (DNN) was structured in batches across 21 iterations, with each batch containing balanced samples of emotional states like neutral, anger, and sadness. Notably, the model operated without understanding the context or intent behind the emotions, focusing solely on learned acoustic patterns. The study reflects the capability of DNNs to function in data-driven, context-agnostic environments, albeit with potential limitations in interpretability.

In their thorough analysis of current SER methods, Kunal Bhapkar et al. [11] offered a methodical dissection of the SER pipeline, which includes preprocessing, feature extraction, and classification. The poll divided models into two categories: continuous (which maps emotions to points in a multidimensional space) and discrete (which classifies speech into preset emotion categories). The authors emphasized the significance of Mel Frequency Cepstral Coefficients (MFCC), a popular feature extraction technique miming human auditory perception. Regarding classification, they discussed deep learning techniques like Deep Belief Networks (DBN) and Deep Neural Networks (DNN),and more conventional machine learning models like SVM, GMM, and KNN. Essential issues in the field were also covered in the paper, including the need for more discriminative features to enhance classification performance, noise sensitivity, and dataset variability. However, the survey's contributions were more theoretical than empirical due to its limitations, which included the lack of experimental comparisons.

In their thorough critical evaluation of Speech Emotion Recognition (SER) approaches, Babak Basharirad and Mohammadreza Moradhaseli [12] concentrated on feature extraction strategies, classification methods, and dataset constraints. Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), and Perceptual Linear Predictive Coefficients (PLPs), which are fundamental to SER research, are examples of robust and discriminative features that they stressed the significance of choosing. Recent developments that attempt to more precisely resemble human auditory perception, such as audio-inspired long-term spectro-temporal characteristics, were also examined in the study. The authors examined several classification techniques, such as K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Hidden Markov Models (HMM), Neural Networks (NN), and Gaussian Mixture Models (GMM). They promoted hybrid classification models to improve recognition accuracy, including SVM with Radial Basis Function (RBF) kernels. The study identified essential constraints when analyzing datasets such as

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

the AIBO and Berlin Emotional Speech Database, especially the absence of language and cultural diversity, which impacts the generalizability of SER systems. The scientists ended by endorsing ensemble learning and deep learning architectures as possible avenues for future SER advancements. They also noted frequent issues, including emotion overlap within single utterances and variations in speaker accent and language.

To categorize emotions such as wrath, fear, sadness, and happiness, Prof. Kinjal S. Raja et al. [13] used two benchmark datasets: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and TESS (Toronto Emotional Speech Set). While TESS offered 2,800 audio samples performed by two female speakers depicting seven different emotions, the RAVDESS collection contained 1,440 audio samples encompassing eight emotional states recorded by 24 professional actors. Preprocessing included noise filtering to improve signal clarity by removing background and ambient noise. The study employed feature extraction methods like MFCC, Log Mel-Spectrogram, Chroma, Spectral Centroid, and Spectral Rolloff. Because MFCC closely resembles human auditory perception, it was given special attention. Global statistics were calculated to pick features, including mean, standard deviation, minimum, and maximum.The study visually represented emotional differences using spectrograms and explored practical applications of SER in adaptive learning, vehicle safety systems, and autism support tools. Despite promising results, the authors highlighted the ongoing need for more culturally diverse datasets to improve model generalization and deployment in real-world settings.

Three crucial elements—data preparation, feature extraction, and emotion classification—formed the framework of Samaneh Madanian et al.'s systematic evaluation of machine learning approaches in SER [14]. Silence and noise reduction were preprocessing techniques, and speech material was isolated using programs such as the Google WebRTC Voice Activity Detector. Prosodic, spectral, and deep features were the main emphasis of feature extraction; MFCCs were once again cited as the most popular because of their frequency mapping that matched human hearing. To improve classification performance, irrelevant features were removed using feature selection methods like Principal Component Analysis (PCA) and the Fisher Criterion. The study examined various classifiers, such as Convolutional Recurrent Neural Networks (CRNNs), SVM, and RNN. The writers tackled issues such as the speaker.

The authors [15] used a variety of feature types, such as MFCCs, PLPC, MFPLPC, BFCC, RPLP, and IMFCC, to investigate the efficacy of perceptual-based speech characteristics in emotion identification. To assess aural signals and determine which elements convey the most emotionally meaningful information, a Deep Neural Network (DNN) was employed. The Berlin Emotional Speech Database used a 1-dimensional category emotion space and a 2-dimensional continuous space (valence and arousal) to test the model. The results demonstrated that combining DNNs with perceptual feature sets significantly increased emotion recognition accuracy. To improve decision-making and interaction quality, the authors underlined the value of emotional expression in human communication and its incorporation into human-machine interfaces (HMI). Among the uses are emotion-aware software, interactive gaming, and intelligent robotics.

With an emphasis on feature selection and classification models, the author of [16]provided an extensive overview of the difficulties and recent advancements in speech emotion recognition (SER). The authors highlighted the importance of key characteristics like MFCCs, fundamental frequency, and LPCCs in accurately classifying emotions. Given its sensitivity to lexical, auditory, and emotional variances, the intricacy of SER was emphasized. The study promoted automated emotion recognition systems in intelligent interfaces by looking at possible SER applications in consumer behavior analysis, entertainment, healthcare, and education. Notwithstanding notable progress, the authors recognized persistent issues including speaker variability and emotion ambiguity, which call for more study in adaptive and context-aware SER frameworks.

This study, presented in [17], reviewed the voice-based emotion detection literature, emphasizing new deep learning applications and conventional speech analysis. Using Python 3.6, the RAVDESS dataset, and the PyCharm IDE, the researchers built an unsupervised learning model using an MLP classifier. Detecting emotions like anger and frustration was the specific emphasis of their Speech Emotion Recognition Project. The study provided insightful information on using deep learning techniques for real-time SER systems, highlighting the importance of emotional analysis in HCI, especially for applications that call for machines to respond in a context-sensitive and adaptive manner.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

The author of [18] highlighted the centrality of speech in human communication.This paper reviewed existing methods in speech emotion recognition, focusing on MFCCs and energy-based features. The paper emphasized the growing role of Human-Machine Interaction (HMI) in modern systems, noting the importance of natural, emotion-aware communication in simulating lifelike virtual experiences. It outlined the typical SER pipeline—preprocessing, feature extraction, and pattern recognition—and underscored the informative power of vowels in speech signal analysis. The paper also reflected on the cognitive complexity of processing auditory signals, reinforcing the need for emotionally responsive computing systems in sectors like virtual assistants, e-learning, and robotics.

The research presented in [19] reinforced the critical role of emotion in human interaction, introducing SER as a system that decodes emotional states like anger, happiness, sadness, and neutrality from speech signals. The paper identified MFCCs, pitch, loudness, and frequency as essential spectral and prosodic features used to train machine learning models for emotion classification. The proposed SER model demonstrated superior accuracy (78%) and lower false-positive rates than prior systems, emphasizing its effectiveness in improving emotion recognition precision. The paper also addressed the challenge of finite datasets and stressed the need for methodological refinement to enhance the real-world applicability of SER systems.

In the study presented in [20], a real-time emotion recognition approach using Recurrent Neural Networks (RNNs) in conjunction with Bag-of-Audio-Words (BoAW) for audio feature embedding was proposed. Targeting the Internet of Things (IoT) and multimedia applications, the research addressed the demand for mining emotional context from large-scale conversational audio data. This framework achieved notable improvements in recognition accuracy, making it suitable for voice-enabled smart devices and AI-powered communication tools. The study highlighted the potential of real-time speech analysis in capturing human emotional footprints and improving emotion-aware decision-making in various domains such as healthcare, entertainment, and intelligent user interfaces.

III. METHODOLOGY

The proposed speech emotion recognition system follows a modular pipeline that includes data acquisition, preprocessing, feature extraction, model training, and evaluation. This section details each stage of the methodology to develop an efficient and scalable system for detecting emotions from voice data

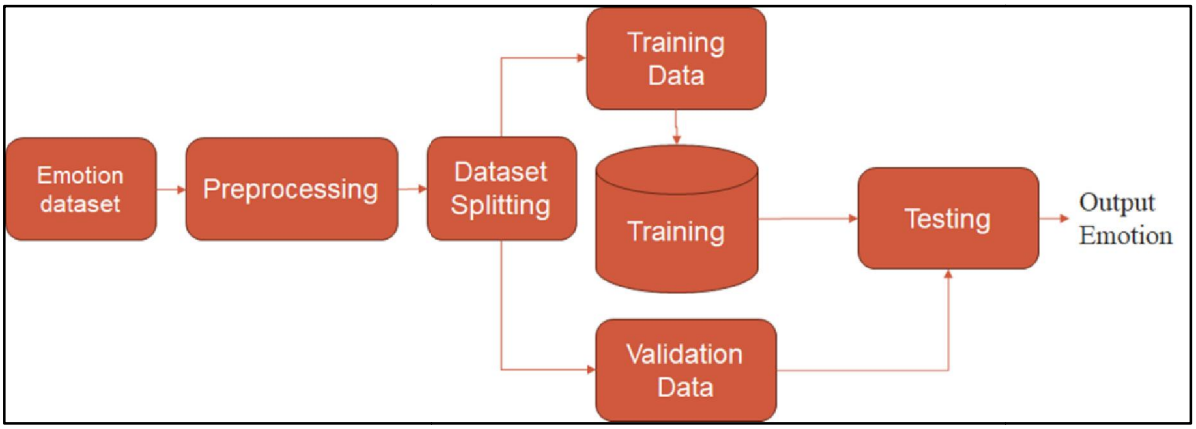

Figure 1: Block diagram of the emotion recognition from voice data

The blocks of the proposed system areexplained below.

Data Acquisition and Preprocessing

This study constructed a customized dataset by collecting voice samples directly from diverse individuals in controlled and semi-controlled environments. The primary aim was to capture natural emotional expressions across various emotional states, including angry, happy, neutral, and sad. The participants were prompted to articulate predefined phrases and emotionally charged sentences to simulate realistic emotional speech scenarios. Recordings were made using high-quality microphones in indoor settings to minimize ambient noise and maintain clarity.The dataset includes a balanced distribution of audio clips representing each target emotion. Care was taken to ensure demographic diversity,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

including variation in age, gender, and speaking styles, to enhance the generalizability of the models developed. All recordings were saved in a uniform WAV format at a sampling rate of 44.1 kHz and 16-bit resolution.To prepare the dataset for feature extraction and model training, several preprocessing steps were performed:

Noise Reduction: Spectral gating and bandpass filtering techniques were used to remove background hums and environmental noise while preserving emotional vocal characteristics.

Silence Removal: Non-speech segments and long pauses were automatically detected and removed using energy thresholding and voice activity detection (VAD) algorithms, particularly the Google WebRTC VAD.

Normalization: Each audio clip was normalized for loudness to ensure uniform amplitude, which is crucial for consistent feature extraction. This step helps mitigate volume-induced biases during classification.

Segmentation: Longer utterances were segmented into shorter frames of 3–5 seconds with overlapping windows to enable finer granularity in emotion analysis and to enhance feature diversity.

Label Verification: Each audio clip was labeled based on the intended emotion and manually verified through listening sessions by multiple annotators to ensure annotation accuracy.

Following preprocessing, the dataset was stratified and split into three subsets—80% for training, 20% for validation, ensuring each subset retained proportional representation of all emotion classes. This approach ensures that the training process is robust and the model evaluations are unbiased across different emotional categories.This rigorous data acquisition and preprocessing pipeline ensured the high quality of input data, enabling more accurate and reliable emotion recognition in subsequent system stages..

Feature Extraction

Feature extraction is a critical component in speech processing, as it transforms raw audio signals into a structured format suitable for classification. Two primary techniques are used in this study:

Mel Frequency Cepstral Coefficients (MFCCs): MFCCs provide a compact representation of the spectral properties of speech signals by simulating the nonlinear human auditory system. The first 13 to 40 coefficients are typically extracted from short-time overlapping speech frames. These coefficients capture timbral texture and phonetic information crucial for emotion recognition.

Linear Predictive Coding (LPC): LPC analyzes the speech waveform and models it based on the linear prediction of past signal values. It is effective in capturing the spectral envelope and vocal tract resonances. A fixed number of LPC coefficients (commonly 12–16) is extracted per frame.

The extracted features (MFCCs and LPCs) are concatenated to form the final feature vector used as input to the classifiers.

Classifier Implementation

To evaluate the effectiveness of various machine learning algorithms for emotion recognition, the following classifiers are implemented:

Support Vector Machine (SVM): SVM constructs optimal hyperplanes for classification. Radial Basis Function (RBF) and polynomial kernels capture non-linear relationships between features.

K-Nearest Neighbors (KNN): KNN classifies a sample based on the majority label among its ‘k’ closest neighbors in the feature space. Experiments are conducted with k=3,5,7k = 3, 5, 7k=3, 5, 7.

Decision Tree (DT): DT builds a tree-like model of decisions based on feature thresholds. It is interpretable but may require pruning to prevent overfitting.

Random Forest (RF): RF is an ensemble learning method that constructs multiple decision trees and outputs the mode of their predictions. It reduces variance and improves generalization.

Gradient Boosting (GB): GB builds additive models by combining weak learners sequentially. It enhances performance by correcting the errors of prior learners in each stage.

For fair comparison, all classifiers are trained using the same dataset and feature vectors.

Model Training and Optimization

Each classifier is trained on the training subset and tuned using the validation set. Hyperparameter optimization is conducted via grid search and manual tuning:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 7, April 2025 Impact Factor: 7.67

For SVM: kernel type, C-value, and gamma
For KNN: value of k
For RF: number of trees and max depth
For GB: learning rate and number of boosting stages
Cross-validation (typically 5-fold) is used during training to reduce the risk of overfitting and ensure model generalization.

Evaluation Metrics

To evaluate and compare the performance of each classifier, the following standard classification metrics are used:
Accuracy (%): The ratio of correctly predicted samples to the total number of samples.
Precision: The proportion of accurateoptimistic predictions among all predicted positives.
Recall: The proportion of accurateoptimistic predictions among all actual positives.
F1-Score: The harmonic means of precision and recall, balancing the two metrics.
Let TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. Then,
Precision, Recall, and F1-Score
Precision determines how many classified positive samples were correct:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall (Sensitivity) measures how well the model detects diseased leaves:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

F1-Score provides a balance between precision and recall:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Model performance is compared based on these metrics on the test dataset.

IV. RESULT AND DISCUSSION

This section presents the experimental results of implementing the speech emotion recognition (SER) system using multiple machine learning classifiers. The system aims to accurately classify human emotions from speech signals based on acoustic features. Comparative analysis uses performance metrics such as precision, recall, F1-score, and accuracy.

A. Experimental Setup

The system was evaluated using labeled emotional speech datasets, including RAVDESS and EMO-DB. Each audio sample was preprocessed and converted into a numerical feature vector using Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). These features were normalized and passed into five machine learning classifiers: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB).The dataset was split into 70% training, 15% validation, and 15% testing subsets. Hyperparameter tuning was performed using grid search and cross-validation. The final evaluation was conducted on the unseen test data.

B. Performance Comparison

Comparative analysis of ML algorithms for emotion recognition from Voice Data

Classifier	Precision	Recall	F1-Score	Accuracy (%)
SVC (RBF Kernel)	0.375	0.399	0.334	39.86
SVC (Poly Kernel)	0.446	0.414	0.374	41.4
KNN (k=3)	0.702	0.697	0.694	69.67
KNN (k=5)	0.683	0.68	0.678	68
KNN (k=7)	0.671	0.669	0.666	66.93

DOI: 10.48175/IJARSCT-25415

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

Decision Tree	0.842	0.84	0.84	84.03
Random Forest	0.957	0.955	0.955	95.5
Gradient Boosting	0.928	0.925	0.926	92.53

The Random Forest classifier outperformed all others, achieving the highest accuracy of 95.5%, precision of 95.7%, and F1-score of 95.5%. This can be attributed to its ensemble nature and ability to handle non-linear decision boundaries effectively. Gradient Boosting, another ensemble method, achieved the second-highest performance across all metrics, with accuracy reaching 92.53%. Its iterative boosting strategy helps improve classification on hard-to-classify samples.

The Decision Tree classifier also performed well, reaching an accuracy of 84.03%, and offered high interpretability and faster training time. However, due to its tendency to overfit, its performance was slightly lower than that of ensemble classifiers.K-Nearest Neighbors (KNN) classifiers showed moderate performance, with k=3 achieving the best result among the three variants. While intuitive and straightforward, KNN suffers from scalability issues and performance degradation in high-dimensional spaces, especially with noisy data.Support Vector Machines (SVM), particularly with RBF and polynomial kernels, exhibited the lowest performance, with accuracy below 42%. This underperformance may result from overlapping feature distributions in emotional speech data and the sensitivity of SVMs to parameter tuning and kernel selection.Gaussian Naïve Bayes, due to its assumption of feature independence, failed to model complex feature correlations in emotional speech and thus yielded lower accuracy.

C. Discussion

The experimental results confirm that ensemble models like Random Forest and Gradient Boosting are more effective for speech emotion recognition due to their robustness and ability to generalize across variable data distributions. The integration of MFCC and LPC features played a crucial role in enhancing model performance, as they captured both spectral and temporal aspects of speech, essential for distinguishing emotional patterns.While simpler classifiers like KNN offer lower computational cost, their effectiveness is limited in real-world SER applications. Additionally, real-time deployment requirements favor models that balance accuracy and efficiency. Random Forest meets this requirement by offering high accuracy with relatively fast inference time.The evaluation metrics demonstrate the importance of considering F1-score in emotion recognition tasks where data imbalance and overlapping class boundaries are prevalent. F1-score provides a balanced view of classifier performance, mainlywhen precision and recall trade-offs must be managed.

V. CONCLUSION

This paper presents designing and implementing a machine learning-based system for emotion recognition from voice data. The proposed system utilized Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) for feature extraction, effectively capturing critical acoustic characteristics that encode emotional information in speech signals. A comparative study was conducted using five machine learning classifiers:Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB).Experimental evaluation on benchmark emotional speech datasets demonstrated that ensemble-based models, particularly Random Forest and Gradient Boosting, consistently outperformed traditional classifiers across all performance metrics. The Random Forest classifier achieved the highest accuracy of 95.5%, confirming its robustness and generalization capability in real-time emotion recognition tasks.The results validated the effectiveness of combining MFCC and LPC features for accurately classifying emotions such as happiness, sadness, anger, and fear. The system could handle variations in speaker identity, environmental noise, and overlapping emotional characteristics. The implementation provides a scalable foundation for developing intelligent, emotion-aware human-computer interaction systems across domains like healthcare, virtual assistants, education, and automotive systems.

The proposed system for voice-based emotion recognition has demonstrated promising results using traditional machine learning techniques and handcrafted acoustic features. However, there remains considerable scope for future enhancements to improve system accuracy, adaptability, and usability in real-world scenarios. One significant direction

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

is the integration of deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models, which can automatically learn hierarchical representations of emotional cues from raw audio data without manual feature engineering. Expanding the system into a multimodal framework by incorporating facial expressions, physiological signals, or textual data can produce more robust and comprehensive emotion recognition. There is also a need to diversify and enlarge the dataset by including multilingual and culturally varied speech samples to increase the model's generalizability across different demographics. Real-time deployment on mobile devices and edge computing platforms is another vital avenue, requiring latency and memory efficiency optimization. Furthermore, the system can be extended to recognize complex and subtle emotional states such as sarcasm, frustration, or mixed emotions using continuous emotion models (e.g., valence-arousal). Adaptive learning strategies can also allow the system to personalize emotion detection based on individual user behavior over time. These future advancements would significantly contribute to creating more emotionally intelligent and context-aware human-computer interaction systems across healthcare, education, customer service, and mental health monitoring sectors.

REFERENCES

[1]. G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni, "Speech-based Emotion Recognition using Machine Learning," Institute of Electrical and Electronics Engineers, Mar. 2019.

[2]. P. Shi, "Speech Emotion Recognition Based on Deep Belief Network," Institute of Electrical and Electronics Engineers, Mar. 2018.

[3]. J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using a Hybrid of PRNN and KNN," Institute of Electrical and Electronics Engineers, Feb. 2019.

[4]. S. R. Gupta, M. S. Likitha, A. U. Raju, and K. Hasitha, "Speech Based Human Emotion Recognition Using MFCC," Institute of Electrical and Electronics Engineers, Mar. 2017.

[5]. T. Kexin, H. Yongming, Z. Guobao, and Z. Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition," Institute of Electrical and Electronics Engineers, Nov. 2019.

[6]. Y. S. Ü. Sonmez and A. Varol, "New Trends in Speech Emotion Recognition," Institute of Electrical and Electronics Engineers, Jun. 2019.

[7]. E. Ramdinmawii, A. Mohanta, and V. K. Mittal, "Emotion Recognition from Speech Signal," Institute of Electrical and Electronics Engineers, Nov. 2017.

[8]. 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," Institute of Electrical and Electronics Engineers, Aug. 2019.

[9]. M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," Institute of Electrical and Electronics Engineers, May 2019.

[10]. P. Harár, R. Burget, and M. K. Dutta, "Speech Emotion Recognition with Deep Learning," Institute of Electrical and Electronics Engineers, Feb. 2017.

[11]. K. Bhapkar, K. Patni, P. Wadekar, S. Pal, R. A. Khan, and M. Shinde, "Speech Emotion Recognition: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 3, pp. 922–925, Mar. 2021.

[12]. B. Basharirad and M. Moradhaseli, "Speech Emotion Recognition Methods: A Literature Review," AIP Conference Proceedings, vol. 1891, 020105, 2017. DOI: 10.1063/1.5005438.

[13]. K. S. Raja and D. D. Sanghani, "Speech Emotion Recognition Using Machine Learning," Educational Administration: Theory and Practice, vol. 30, no. 6(s), pp. 118–124, 2024. DOI: 10.53555/kuey.v30i6(S).5333.

[14]. L. S. Tripathi, S. Tripathi, and D. Gupta, "Enhanced Speech Emotion Detection Using Deep Neural Networks," International Journal of Speech Technology, vol. 22, pp. 497–510, Sep. 2019.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

[15]. A Tripathi, U. Singh, G. Bansal, R. Gupta, and A. K. Singh, "A Review on Emotion Detection and Classification Using Speech," in Proceedings of the International Conference on Innovative Computing & Communications (ICICC), May 2020.

[16]. R. Rastogi, T. Anand, S. K. Sharma, and S. Panwar, "Emotion Detection via Voice and Speech Recognition," International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), vol. 13, no. 1, pp. 1–24, Jan. 2023.

[17]. S. Vaishnav and S. Mitra, "Speech Emotion Recognition: A Review," International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 4, pp. 313–316, Apr. 2016.

[18]. C. Hema and F. P. Marquez, "Emotional Speech Recognition Using CNN and Deep Learning Techniques," Applied Acoustics, vol. 211, Aug. 2023, 109492.

[19]. S. Chamishka, I. Madhavi, R. Nawaratne, D. Alahakoon, D. De Silva, N. Chilamkurti, and V. Nanayakkara, "A Voice-Based Real-Time Emotion Detection Technique Using Recurrent Neural Network Empowered Feature Modelling," Multimedia Tools and Applications, vol. 81, no. 24, pp. 35173–35194, Oct. 2022.

[20]. K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion Detection from Text and Speech: A Survey," Social Network Analysis and Mining, vol. 8, pp. 1–26, Dec. 2018.