Speech Emotion Recognition using Gaussian Mixture Model (GMM) and K-Nearest Neighbors (KNN)

Kirtika Iyer, Abhay Shukla, Kunal Sharma, Maya Varghese

Department of Computer Science and Engineering (Data Science), Vidyavardhini's College of Engineering and Technology, Palghar, Maharashtra, India

Corresponding author: Abhay Shukla, Email: abhay.203578101@vcet.edu.in

This paper aimed to propose a novel methodology to improve the accuracy and efficiency of speech emotion recognition, through to the multilingual setting. The paper's topic was the low precision obtained by the systems for speech emotion recognition, especially in multilingual settings. The research problem was the performance of the existing systems which could achieve merely 72% accuracy in recognizing the correct emotion from speech. The research's importance was the enhancement of the performance of these systems in order to improve the user experience and the range of its applications in multilingual settings. The paper uses a research methodology with feature extraction methods and machine learning algorithms, such as Mel-frequency cepstral coefficients, zero-crossing rate, harmonic-to-noise ratio, such as Gaussian Mixture Models, and K-Nearest Neighbors. The proposed methodology analysis leads to a major increase in accuracy, attaining the performance of 82% in the complex multilingual environment. Besides, this research paper describes the areas for future research to allow additional improvement and overcome the possible weaknesses of the designed methodology, contributing to the development of the field.

Keywords: Speech Emotion Recognition, Artificial Intelligence, Emotion Identification, Feature Extraction, Machine learning algorithms, Gaussian Mixture Models, k-Nearest Neighbors.

in: *Advancements in Communication and Systems.* Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 443–455. DOI: https://doi.org/10.56155/978-81-955020-7-3-39

1 Introduction

Speech Emotion Recognition (SER) refers to the process of automatically identifying and categorizing human emotions expressed through speech signals. Speech emotion recognition has received a lot of attention in the recent past because of its wide range of applications in human-computer interaction, affective computing and healthcare. To develop intuitive and empathetic human machine interfaces, it is imperative to comprehend and interpret speech emotions accurately. Notwithstanding SER's technology advancements, there are still challenges especially when it comes to multilingual situations where existing systems find it difficult to achieve high levels of accuracy. This paper investigates the complexities of SER focusing on advanced feature extraction techniques like zero-crossing rate (ZCR), harmonic-to-noise ratio (HNR) and Mel-frequency cepstral coefficients (MFCC). Also, this study investigates the applicability of Gaussian Mixture Models (GMM) and K-Nearest Neighbors (KNN) to classify emotions from speech. Gaussian Mixture Model (GMM) is a popular model for probabilistic pattern recognition and machine learning that expresses a probability distribution as a weighted sum of multiple Gaussian distributions named components. Conversely, K-Nearest Neighbors (KNN) is an unsophisticated yet easily understood classification algorithm commonly used in supervised learning problems.

2 Literature Review

Drawing largely on recent literature, the paper [1] reviews the design and methods of speech emotion recognition systems at length: broadening gaps in knowledge and discussing current efforts to get programmable devices to understand what emotions are spoken of whether these words express delight, hurt or anger Five related works in this field have explored the use of different feature extraction techniques, valuation algorithms, and incorporation of deep learning methods for improving emotional responses. Important work in this area comes from (cite pertinent studies), which has paved the ground for effective SER systems and put forward new opportunities to improve human-machine interaction.

The paper [2] main focus is a SER system that includes traditional audio features and state-of-the-art techniques like auto-encoders and Support Vector Machines (SVM) for emotion classification. Both of these previous studies have collectively influenced the development of the two-stage SER system described in the paper, and with results like a stacked auto-encoders accuracy rate of 74.07%, it is easy to see how successful such methodologies have been.

The paper [3] introduces a Speech Emotion Recognition system that is intended to assist in human-computer interaction by recognizing emotions in speech. It makes use of Convolutional Neural Network (CNN) technique and incorporates several datasets on training its models. This approach is demonstrated to be effective, with an accuracy rate of 75% in speech emotion recognition.

The paper [4] explores the intricate and real-world challenges of recognizing and interpreting human emotions within audiovisual content captured in natural, uncontrolled environments. By delving into the complexities of emotion recognition beyond controlled settings, it sheds light on the potential for more robust and context-aware systems in the field of human-computer interaction and affective computing.

The paper [5] delves into the innovative use of Teager energy and linear prediction features for emotion recognition in stressed speech. By harnessing these acoustic features, the study aims to enhance the accuracy and efficacy of emotion recognition in contexts characterized by heightened stress levels.

Another expression featured in the paper titled [6] examines how to combine the acoustic features extracted from Teager Energy Operator (TEO) with Mel-frequency cepstral coefficients (MFCC). The present study investigates just what effect this feature combinatory technique has in accurately identifying and classifying stressed emotional states in natural language.

The paper [7] is a detailed study of the effectiveness of speech descriptors on emotion recognition, and their ability to capture affective information in speech. Through thorough analysis and practical verification, this paper provides an important basis for further research work in this field.

The paper [8] explores emotion detection from speech signals using Mel Frequency Cepstrum Coefficient (MFCC) features and Logistic Model Tree (LMT) classifiers, utilizing a voting mechanism on classified frames to discern human emotions with up to 70% accuracy across various datasets.

The paper [9] investigates automatic speech emotion recognition (SER) systems using Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features with various classifiers, demonstrating high accuracy rates of up to 94% on the Spanish database using a recurrent neural network (RNN) classifier without speaker normalization (SN) and with feature selection (FS).

The paper [10] presents a speech emotion recognition system leveraging deep learning and machine learning algorithms, implemented using Python and the librosa library, achieving an efficiency rate of 81.82% in identifying eight emotions from speech signals sourced from the RAVDESS dataset.

The paper [11] focuses on speech emotion recognition in the Algerian dialect, presenting a new dataset collected from Algerian TV using deep learning approaches, specifically LSTM-CNN, for emotion detection in audio speech signals.

The paper [12] presents a deep learning-based speech emotion recognition algorithm leveraging convolutional neural networks (CNNs), bi-directional long- and short-term memory (LSTM) networks, and multi-headed attention mechanisms to enhance the efficiency and capability of speech emotion recognition, with potential applications in human-computer interaction devices including criminal investigation.

The paper [13] introduces a straightforward convolutional neural network (CNN) architecture based on log-mel-spectrograms for speech emotion recognition, achieving competitive classification accuracies of 59.33% to 72.02% on different emotional databases, including IEMOCAP and the Berlin EmoDB, with a focus on extracting deep features efficiently.

The paper [14] investigates speech emotion recognition using Gaussian Mixture Model (GMM) supervectors combining Mel Frequency Cepstrum Coefficients (MFCCs) and Auto Correlation Function Coefficients (ACFC), achieving improved emotion recognition rates of 74.45% and 75.00% for two sets of emotions compared to previous methods utilizing hidden Markov models.

The paper [15] explores emotion recognition in speech using spectral and prosodic features such as Mel frequency cepstrum coefficients (MFCCs), pitch, and energy, employing K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Gaussian Mixture Model (GMM) classifiers to detect six emotional states (anger, happiness, sadness, fear, disgust, and neutral) from the Berlin emotional speech database.

The paper [16] introduces a speech emotion recognition framework based on a residual neural network (ResNet), integrating different classifiers (SVM, GMM, KNN) to replace the fully connected layer, achieving superior performance with an average accuracy of 90% on the EMODB corpus compared to other architectures.

The paper [17] explores the landscape of speech emotion recognition (SER) using machine learning (ML) techniques over the last decade, highlighting the key steps of data processing, feature extraction, and classification, while discussing challenges, solutions, and guidelines for evaluating SER systems to guide future research and development in this field.

The paper [18] introduces a kernel sparse representation-based classifier (KSRC) enhanced with group sparsity constraints for speech emotion recognition, utilizing dynamic kernels to model variability in speech signal duration and demonstrating superior performance compared to traditional support vector machines (SVM) based classifiers.

The paper [19] explores speech emotion recognition using feature extraction including fundamental frequency (F0), energy (E), zero-crossing rate (ZCR), and Fourier parameters (FP), followed by principal component analysis (PCA) for feature reduction, and utilizes a fusion approach combining SVM and KNN classifiers to enhance emotional state recognition, demonstrating significant results across German and English emotional speech datasets.

The paper [20] presents an implementation of a speech emotion recognition system using Mel Frequency Cepstrum Coefficients (MFCC), wavelet features, and pitch of vocal traces, employing Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) classifiers to identify six emotional categories from the Berlin emotion database (BES), and compares their performance using confusion matrices for analysis.

The paper [21] has explored the application of spectral components such as Mel Frequency Cepstrum Coefficients (MFCC), wavelet features, and pitch of vocal traces in speech emotion recognition systems. The paper [22] investigates speech-based emotion recognition using machine learning algorithms, including Recurrent Neural Network (RNN), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Adaboost, Gradient Boosting Classifier, Multi-Layer Perceptron (MLP), and Random Forest, with feature extraction from audio using Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel Spectrogram Frequency, Spectral Contrast, and Tonnetz with the MLP classifier on the RAVDESS database.

The paper [23] presents a hybrid machine learning model for speech emotion recognition, utilizing a combination of feature extraction methods including Voice Pitch, Mel Frequency Cepstral Coefficients (MFCC), and Short-Term Energy (STM), along with algorithms such as GFCC, ALO, and MSVM to achieve accurate emotion classification with an 80% accuracy rate, evaluated using the MATLAB simulation tool.

The paper [24] introduces an enhanced human speech emotion recognition system using a hybrid approach combining Pattern Recognition Neural Network (PRNN) and K-Nearest Neighbor (KNN) algorithms, leveraging Mel Frequency Cepstral Coefficient (MFCC), Gray Level Co-occurrence Matrix (GLCM) for feature extraction, and a Wiener filter for noise reduction, demonstrating improved efficiency compared to prior recognition systems for classifying six basic emotions from speech samples.

The paper [25] investigates the use of unsupervised representation learning on large, unlabeled speech corpora to enhance speech emotion recognition (SER), demonstrating improved recognition accuracy by integrating representations learned from an unsupervised autoencoder into a CNN-based emotion classifier, evaluated across different datasets including IEMOCAP and MSP-IMPROV with visualization analysis using t-distributed neighbor embeddings (t-SNE).

The paper [26] introduces a novel approach for stressed speech emotion recognition using a feature fusion technique combining Teager Energy Operator (TEO) and Mel Frequency Cepstral Coefficients (MFCC) called Teager-MFCC (T-MFCC), leveraging Gaussian Mixture Model (GMM) classification to achieve improved performance compared to traditional MFCC-based methods.

The paper [27] presents a Speech Emotion Recognition (SER) system utilizing a combination of Teager Energy Operator (TEO) and Linear Prediction Coefficient (LPC) features (T-LPC) for recognizing stressed speech signals, achieving improved performance compared to traditional pitch-based and LPC-based recognition systems, with applications in motivating students by accurately detecting their emotional states.

The paper [28] investigates emotion recognition from speech using Linear Prediction Coefficients (LPC) and neural networks (NN), demonstrating effective characterization of basic emotions such as sad, anger, happy, disgust, fear, and boredom by capturing emotion-specific information from speech signals.

Various machine learning algorithms have been employed, with Gaussian Mixture Model (GMM) and K-Nearest Neighbour (K-NN) models emerging as prominent choices. These studies collectively contribute to the understanding of spectral features and machine learning techniques in speech emotion recognition, providing a context for the present paper's exploration of MFCC, wavelet features, and pitch with GMM and K-NN models, with GMM showing a significant accuracy range from 25% for 'surprise' to 92% for 'angry,' while K-NN exhibited 90% accuracy for 'happy' and 50% for both 'fear' and 'surprise.'

3 System Design

The proposed system for Speech Emotion Recognition is designed to accurately classify emotions from speech signals using a comprehensive approach encompassing multiple stages. This system leverages diverse datasets of labelled speech samples, allowing for supervised learning of emotion classification. Key to the system's effectiveness is the feature extraction stage, which utilizes essential speech signal characteristics, including Mel-frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), harmonic-to-noise ratio (HNR), and Teager energy operator (TEO). These features capture spectral, temporal, and energy-based properties of speech signals, crucial for distinguishing different emotions. After feature extraction, normalization is employed to standardize feature vectors, enhancing consistency and facilitating robust model performance. The system trains Gaussian Mixture Models (GMM) and K-Nearest Neighbors (KNN) classifiers using these normalized features to model the probability distributions and similarity relationships among emotion categories, respectively. During prediction, the system computes the likelihood of test samples belonging to each emotion category based on the trained models and assigns the most probable emotion to each sample. Performance evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the system's effectiveness in emotion classification. This system architecture integrates advanced signal processing techniques with machine learning algorithms to build a robust SER system capable of accurately recognizing and categorizing emotions conveyed through speech.

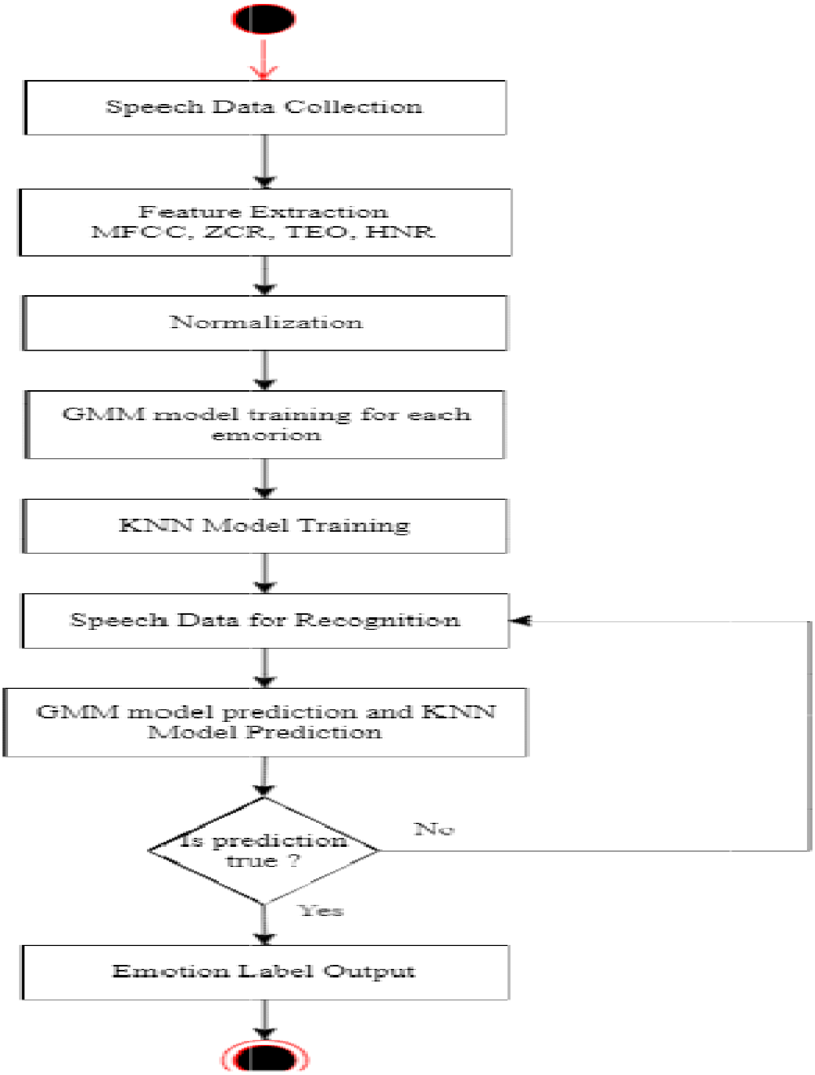
Speech Data Collection
Feature Extraction MFCC, ZCR, TEO, HNR
Normalization
GMM model training for each emorion
KNN Model Training
Speech Data for Recognition
GMM model prediction and KNN Model Prediction
Is prediction true ?
No
Yes
Emotion Label Output

Figure 1. System Architecture

The Figure 1 shows the proposed system architecture. The proposed system for speech emotion recognition (SER) comprises several stages, each crucial for accurate emotion classification.

Speech Data Collection: In this stage, a diverse dataset of speech samples is collected, encompassing various emotions and linguistic backgrounds. Each speech sample is labelled with the corresponding emotion to facilitate supervised learning.

Feature Data Extraction: Feature extraction is essential for capturing relevant information from the speech signals. Four primary features are utilized: Mel-frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), harmonic-to-noise ratio (HNR), and Teager energy operator (TEO).

MFCCs represent the spectral characteristics of speech signals and are computed using the Discrete Fourier Transform (DFT) and Mel-filter bank. Mathematically, the MFCCs can be calculated as follows:

$$\mathrm{MFCC_i} = \sum_{i=1}^{N} \log(|X(n)|) \cos\left(\frac{m\pi(n-0.5)}{N}\right) \sin\left((2n+1)\, i\pi/2N\right) \tag{1}$$

where,

- X(n) is the magnitude spectrum of the speech signal
- N is the number of FFT points,
- m is the Mel frequency index, and
- i is the coefficient index.

ZCR represents the rate at which the signal changes its sign and is computed as the number of times the signal crosses zero divided by the signal length. Mathematically:

$$\mathrm{ZCR} = \frac{1}{N}\sum_{n=1}^{N-1} |\mathrm{sign}(x(n) - \mathrm{sign}(x(n-1))| \tag{2}$$

where,

- N is the length of the signal, and
- sign(x) returns the sign of x.

HNR quantifies the ratio of harmonics to noise in the speech signal and is computed using the Fourier Transform. Mathematically:

$$\mathrm{HNR}=H/N \tag{3}$$

where,

- H is the sum of the harmonics, and
- N is the sum of the noise

TEO captures the energy variations in the speech signal and is calculated using a nonlinear operation. Mathematically:

$$\mathrm{TEO} = \sum_{n=1}^{N-1} (x(n))^2 - x(n-1)x(n+1) \tag{4}$$

where,

- x(n) is the sample value at time n , and
- N is the length of the signal

Normalization: After feature extraction, the feature vectors are normalized to ensure consistency and improve model performance. Standard scaling is commonly employed, where each feature is scaled to have zero mean and unit variance.

GMM Model Training and Prediction: The normalized feature vectors are used to train a Gaussian Mixture Model (GMM) for each emotion category. The GMM represents the probability distribution of each emotion class in the feature space. During prediction, the likelihood of a test sample belonging to each GMM is computed, and the emotion with the highest likelihood is assigned to the sample. The mathematical equation for GMM can be expressed as follows:

$$p(x) = \sum_{i=1}^{K} \pi_i \, N(x|\mu_i, \Sigma_i) \tag{5}$$

where,

- p(x) is the probability density function of the GMM,
- K is the number of components in the mixture,
- π_i is the mixing coefficient of the ith component,
- $N(x|\mu_i, \Sigma_i)$ is the multivariate Gaussian distribution with mean μ_i and covariance Σ_i.

KNN Model Training and Prediction: In addition to GMM, a K-Nearest Neighbours (KNN) classifier is trained using the normalized feature vectors. KNN relies on the similarity between feature vectors to classify test samples. During prediction, the K nearest neighbours to a test sample are identified, and the majority emotion among these neighbours is assigned to the sample. The mathematical equation for KNN can be expressed as follows:

$$y = \mathrm{argmax}_c \sum_{i=1}^{K} I(y_i = c) \tag{6}$$

where,

- y is the predicted class label for the test sample,
- c is the class label,
- y_i is the class label of the ith nearest neighbour,
- I(.) is the indicator function that returns 1 if the condition is true and 0 otherwise.

Evaluation: The performance of the SER system is evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the effectiveness of the system in accurately classifying emotions from speech data.

4 Implementation

The implementation of the speech emotion recognition (SER) system utilizing the RML emotion database begins with data preprocessing and feature extraction. The video samples provided in the dataset are first converted from .avi format to .wav format to facilitate audio-based analysis. Subsequently, the dataset is segregated into different emotion labels, namely happiness (HA), sadness (SA), anger (AN), fear (FE), surprise (SU), and disgust (DI), irrespective of the languages spoken. This step ensures the creation of labeled datasets for training and testing the SER model.

Following data segregation, feature extraction is performed on the audio samples to capture relevant information for emotion classification. This includes computing 39 Mel-frequency cepstral coefficients (MFCC), harmonic-to-noise ratio (HNR), zero-crossing rate (ZCR), and Teager energy operator (TEO), resulting in a total of 42 features per sample. The MFCCs capture spectral characteristics, while HNR measures the ratio of harmonic components to noise, ZCR quantifies signal periodicity, and TEO captures signal energy changes.

Once feature extraction is complete, the extracted features are normalized to ensure consistency and facilitate model training. Normalization ensures that each feature contributes equally to the classification process, regardless of its scale or magnitude. This step enhances the robustness and effectiveness of the subsequent machine learning models.

The hybrid model for SER is constructed using a combination of Gaussian Mixture Models (GMM) and K-Nearest Neighbors (KNN). GMMs are employed to model the distribution of feature vectors for each emotion class, allowing for probabilistic classification. On the other hand, KNN utilizes the Euclidean distance metric to identify the nearest neighbors of a given feature vector and assigns the majority class label among its neighbors.

The implementation of the hybrid GMM-KNN model involves training the GMM on the normalized feature vectors and using its predicted probabilities as input features for the KNN classifier. This integrated approach leverages the strengths of both models to enhance the accuracy and robustness of the SER system. The trained model is then evaluated on a separate test dataset to assess its performance in accurately recognizing emotions from speech samples. This comprehensive implementation workflow ensures the development of an effective and reliable SER system tailored to the requirements of the RML emotion database.

5 Result & Discussion

The outcome of the proposed system is a substantial improvement in the accuracy of Speech Emotion Recognition (SER).

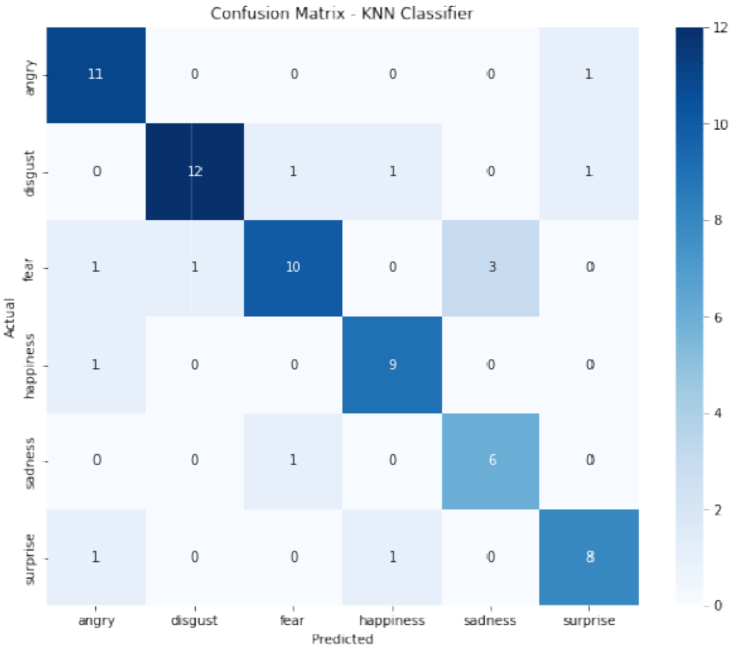

Figure 2. Confusion Matrix

The Figure 2 shows confusion matrix that provides a visual representation of the performance of a classification model. From the confusion matrix, it is evident that the system performs well in correctly classifying instances across most emotion categories, with relatively few misclassifications. Notably, the majority of instances are correctly classified along the diagonal, indicating strong performance. It is observed that the model performs well in correctly identifying instances of 'Angry', 'Disgust', and 'Happiness', as indicated by the high diagonal values in these rows. However, it struggles more with classifying 'Fear' and 'Sadness', as evidenced by the off-diagonal values in these rows.

Table 1. Comparison between Existing System (SVM) and Proposed System (GMM & KNN)

System	Existing System (SVM)	Proposed System (GMM & KNN)
Accuracy	73.2067% [2]	82%

The Table 1 shows the comparison between Existing System (SVM) and Proposed System (GMM & KNN) and it is evident that the existing system shows 73.2067% accuracy overall whereas proposed system shows the overall accuracy of 82%.

Table 2. Classification Report

	Precision	Recall	F1-Score	Support
Angry	0.79	0.92	0.85	12
Disgust	0.92	0.80	0.86	15
Fear	0.83	0.67	0.74	15
Happiness	0.82	0.90	0.86	10
Sadness	0.67	0.86	0.75	7
Surprise	0.80	0.80	0.80	10
Accuracy			0.82	69
Macro Avg.	0.80	0.82	0.82	69
Weighted Avg.	0.82	0.81	0.82	69

The Table 2 shows the classification report which provides a summary of the classification performance across different emotion classes. It includes metrics such as precision, recall, and F1-score, which assess the model's accuracy in classifying each emotion class. Additionally, it reports the overall accuracy of the model across all classes.

Examining the precision scores, it is observed that the system performs well in identifying the 'Disgust' and 'Happiness' emotions, with precision scores of 0.92 and 0.82, respectively. This suggests that when the system predicts these emotions, it is highly likely to be correct. However, the precision for 'Sadness' is relatively lower at 0.67, indicating that the system may sometimes misclassify other emotions as 'Sadness'.

Similarly, the recall scores provide insight into the system's ability to correctly identify instances of each emotion class. Notably, 'Angry' and 'Happiness' have high recall scores of 0.92 and 0.90, respectively, indicating that the system effectively captures most instances of these emotions. However, the recall for 'Fear' is relatively lower at 0.67, indicating that the system may miss some instances of this emotion.

The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of the system's performance. Overall, the system achieves a macro-average F1-score of 0.82, indicating good overall performance across all emotion classes. Further analysis reveals that while the model demonstrates robust performance, there may be room for improvement in correctly classifying 'Fear' and 'Sadness' emotions. This enhanced accuracy underscores the system's capability to achieve more nuanced and precise recognition of complex and blended emotions within spoken languages.

6 Conclusion

The hybrid approach of using Gaussian Mixture Models (GMMs) and k-Nearest Neighbors (k-NN) in Speech Emotion Recognition (SER) is powerful. GMMs extract features and model emotional states in speech signals, while k-NN ensures accurate emotion classification by measuring similarity between new features and established emotional clusters. Through the integration of Gaussian Mixture Models (GMM) and k-Nearest Neighbors (K-NN), as well as comprehensive feature extraction, normalization, and multilingual capabilities, the system seeks to bridge the gap in accurately identifying complex and blended emotions within spoken language. While precise accuracy percentages cannot be guaranteed, it is reasonable to expect that the system will achieve accuracy rates exceeding the conventional 50-70% range observed in SER. With accuracy level 82% and potentially reaching up to 90%, the system is poised to significantly enhance its applicability in diverse human-computer interaction scenarios, mental health assessment, and customer service, effectively catering to a wide range of languages and emotional contexts. This research contributes to the ongoing evolution of SER technology, with the potential to advance the understanding and recognition of human emotions conveyed through speech. The multilingual support further extends the system's applicability across diverse languages and emotional contexts, making it a valuable tool for human-computer interaction, mental health assessment, and customer service.

7 Future Scope

The system lays the groundwork for future enhancements and research directions. Fine-tuning and optimizing the model parameters could potentially elevate the accuracy from the initial 82% to the anticipated 90% and beyond. Addressing privacy concerns and ethical considerations related to emotion data handling should also be a focus in future developments. Overall, continuous refinement and exploration of advanced methodologies will ensure the Speech Emotion Recognition system remains at the forefront of emotion recognition technology.

References

[1] Wani, Taiba Majid, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and EliathambyAmbikairajah. "A comprehensive review of speech emotion recognition systems." IEEE access 9 (2021): 47795-47814.

[2] Aouani, Hadhami, and Yassine Ben Ayed. "Speech emotion recognition with deep learning." Procedia Computer Science 176 (2020): 251-260.

[3] Vaibhav K. P.,Parth J. M., Bhavana H. K., Akanksha S. S., “Speech Based Emotion Recognition Using Machine Learning”, International Journal for Research in Applied Science & Engineering Technology (IJRASET),Dec 2021.

[4] Avots, Egils, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. "Audiovisual emotion recognition in wild." Machine Vision and Applications 30, no. 5 (2019): 975-985.

[5] Reddy, S. B., and T. Kishore Kumar. "Emotion recognition of stressed speech using teager energy and linear prediction features." In Proceedings of the IEEE 18th International Conference on Advanced Learning Technologies, vol. 9. 2018.

[6] Bandela, Surekha Reddy, and T. Kishore Kumar. "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC." In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2017.

[7] Kamińska, Dorota, Tomasz Sapiński, and Gholamreza Anbarjafari. "Efficiency of chosen speech descriptors in relation to emotion recognition." EURASIP Journal on Audio, Speech, and Music Processing 2017, no. 1 (2017): 1-9.

[8] Zamil, Adib Ashfaq A., Sajib Hasan, Showmik MD Jannatul Baki, Jawad MD Adam, and Isra Zaman. "Emotion detection from speech signals using voting mechanism on classified frames." In 2019 international conference on robotics, electrical and signal processing techniques (ICREST), pp. 281-285. IEEE, 2019.

[9] Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. "Automatic speech emotion recognition using machine learning." (2019).

[10] Babu, P. Ashok, V. Siva Nagaraju, and Rajeev Ratna Vallabhuni. "Speech emotion recognition system with librosa." In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pp. 421-424. IEEE, 2021.

[11] Cherif, Raoudha Yahia, Abdelouahab Moussaoui, Nabila Frahta, and Mohamed Berrimi. "Effective speech emotion recognition using deep learning approaches for Algerian dialect." In 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), pp. 1-6. IEEE, 2021.

[12] Ying, Xie, and Zhang Yizhe. "Design of speech emotion recognition algorithm based on deep learning." In 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), pp. 734-737. IEEE, 2021.

[13] Chauhan, Krishna, Kamalesh Kumar Sharma, and Tarun Varma. "Speech emotion recognition using convolution neural networks." In 2021 international conference on artificial intelligence and smart systems (ICAIS), pp. 1176-1181. IEEE, 2021.

[14] Zhang, Qingli, Ning An, Kunxia Wang, Fuji Ren, and Lian Li. "Speech emotion recognition using combination of features." In 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), pp. 523-528. IEEE, 2013.

[15] Praksah, Chandra, and V. Gaikwad. "Analysis of emotion recognition system through speech signal using KNN, GMM & SVM classifier." IOSR J Electron Commun Eng (IOSR-JECE) 10, no. 2 (2015): 55-67.

[16] Li, Zhen, Jun Li, Sai Ma, and Hui Ren. "Speech emotion recognition based on residual neural network with different classifiers." In 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), pp. 186-190. IEEE, 2019.

[17] Madanian, Samaneh, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L. Schneider. "Speech emotion recognition using machine learning—A systematic review." Intelligent systems with applications (2023): 200266.

[18] Sharma, Pulkit, Vinayak Abrol, Abhijeet Sachdev, and Aroor Dinesh Dileep. "Speech emotion recognition using kernel sparse representation-based classifier." In 2016 24th European Signal Processing Conference (EUSIPCO), pp. 374-377. IEEE, 2016.

[19] Al Dujaili, Mohammed Jawad, Abbas Ebrahimi-Moghadam, and Ahmed Fatlawi. "Speech emotion recognition based on SVM and KNN classifications fusion." International Journal of Electrical and Computer Engineering 11, no. 2 (2021): 1259.

[20] Lanjewar, Rahul B., Swarup Mathurkar, and Nilesh Patel. "Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques." Procedia computer science 49 (2015): 50-57.

[21] Deshmukh, Girija, Apurva Gaonkar, Gauri Golwalkar, and Sukanya Kulkarni. "Speech based emotion recognition using machine learning." In 2019 3rd International Confer-ence on Computing Methodologies and Communication (ICCMC), pp. 812-817. IEEE, 2019.

[22] Arya, Resham, Disha Pandey, Ananya Kalia, Ben Jose Zachariah, Ishika Sandhu, and Divyanshu Abrol. "Speech based emotion recognition using machine learning." In 2021 IEEE Mysore Sub Section International Conference (MysuruCon), pp. 613-617. IEEE, 2021.

[23] Bharti, Deepak, and Poonam Kukana. "A hybrid machine learning model for emotion recognition from speech signals." In 2020 international conference on smart electronics and communication (ICOSEC), pp. 491-496. IEEE, 2020.

[24] Umamaheswari, J., and A. Akila. "An enhanced human speech emotion recognition us-ing hybrid of PRNN and KNN." In 2019 International Conference on Machine Learn-ing, Big Data, Cloud and Parallel Computing (COMITCon), pp. 177-183. IEEE, 2019.

[25] Neumann, Michael, and Ngoc Thang Vu. "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7390-7394. IEEE, 2019.

[26] Bandela, Surekha Reddy, and T. Kishore Kumar. "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC." In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2017.

[27] Bandela, Surekha Reddy, and T. Kishore Kumar. "Emotion recognition of stressed speech using teager energy and linear prediction features." In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), pp. 422-425. IEEE, 2018.

[28] Pathak, Sujata, and Arun Kulkarni. "Recognizing emotions from speech." In 2011 3rd International Conference on Electronics Computer Technology, vol. 4, pp. 107-109. IEEE, 2011.