2024 International Conference on Emerging Systems and Intelligent Computing (ESIC) | 979-8-3503-4985-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ESIC60604.2024.10481531

Recognition of emotional speech using MFCC and Machine Learning Technique

K. Nandini
Assistant Professor, Dept. of CSE
SR Gudlavalleru Engineering College, Gudlavalleru
Andhra Pradesh-521356, India
kagita.nandini@gmail.com

T.Divya*
Dept. of CSE
SR Gudlavalleru Engineering College, Gudlavalleru
Andhra Pradesh-521356, India
tdivya8782@gmail.com

Sk. Subhani
Dept. of CSE
SR Gudlavalleru Engineering College, Gudlavalleru
Andhra Pradesh-521356, India
subhanshaik677@gmail.com

S. Mounika
Dept. of CSE
SR Gudlavalleru Engineering College, Gudlavalleru
Andhra Pradesh-521356, India
surapanenimounika72810203@gmail.com

T. Vamsi Nandhan
Dept. of CSE
SR Gudlavalleru Engineering College, Gudlavalleru
Andhra Pradesh-521356, India
tummalavamsi123@gmail.com

Abstract: Speech to text conversion and interpreting a person's emotions from their speech are essential tasks in human-computer interaction. Applications such as affective computing necessitate comprehension and reaction to human emotions. We are employing a spectral feature in this project. Technology for determining the various characteristics, such as loudness, tone, and intensity, etc through Mel Frequency Cepstral Coefficients (MFCC) .Our primary goal is to categorize eight emotions into a collection of pre-established categories, including neutral, peaceful, pleased, sad, angry, scared, disgusted, and surprised. The suggested technique not only helps to increase voice recognition systems' accuracy but also emphasizes how crucial it is to use coefficients as potent features in this situation. This method has the potential to enhance applications in recommender systems, affective computing, and interaction between humans and computers that require accurate identification of emotions.

Keywords: Spectral features, MFCC, Acoustic properties, Voice processing.

I. INTRODUCTION

The recognition of emotional speech has numerous applications in domains like customer service, mental health monitoring, and interaction between humans and computers, acknowledging emotional speech has drawn increased attention recently [5]. Emotion is a vital component of communication because it shapes our verbal expression of ourselves. Emotional cues in speech must be recognized by sensitive and intuitive systems in order to interpret speech and interact with users in a more effective manner [1]. Making use of machine learning methods in conjunction with Mel Frequency Cepstral Coefficients (MFCCs) is one well-known approach in this field. Because MFCCs can reduce dimensionality while capturing pertinent features in the frequency domain, they have a broad application in speech signal processing using MFCCs to identify and categorize emotional states has shown to be a successful application of machine learning techniques, such as classification algorithms [3]. These algorithms are able to distinguish between various feelings according to the features that are extracted by training models on labeled datasets that contain examples of different emotional speech patterns. The paper also addresses the choice and application of machine learning algorithms for efficient emotional state classification. By improving the robustness and accuracy of the emotional speech recognition system, the suggested methodology hopes to aid in the creation of increasingly complex and emotionally intelligent applications.

II. LITERATURE SURVEY

Conversation is required as input for most natural language processing systems, including voice-activated systems. The author mentioned that the standard protocol involves using This speech input to be converted to text using Automatic Speech Recognition (ASR) systems and then using the text output from ASR for classification or other learning operations[1]. Three essential problems for a SER system to succeed.

- Selecting an effective emotional speech database.
- Extract functional elements.

979-8-3503-4985-6/24/$31.00 ©2024 IEEE

- Create trustworthy classifiers by utilizing machine learning algorithms.

The goal is to improve emotion detection accuracy by utilizing a Machine learning-based emotion classifier to identify emotional characteristics in speech features and semantic information from text. We present various Convolutional neural architectures for text and speech feature-based emotion classification.

III. PROPOSED METHOD

Our study introduces a new method for identifying emotional speech using sophisticated machine learning techniques and Mel Frequency Cepstral Coefficients (MFCCs). Our method consists of a carefully constructed pipeline that is designed to identify distinct emotional states accurately by extracting features from speech signals. Our method's central component is the extraction of MFCCs, which record crucial speech signal spectral properties [1].

We perform extensive evaluations on a validation set for parameter tuning as well as a different testing set for objective performance assessment to guarantee robust performance. Through metrics like recall, accuracy, precision, and F1 score, our approach enables a more thorough analysis of the model's performance and provides insight into how well it can distinguish between different emotional expressions.

Overall, our suggested approach demonstrates the potential for practical uses, such as mental health monitoring and human-computer interaction, in addition to showcasing the effectiveness of MFCCs in capturing emotional cues [6].

IV. SPEECH EMOTION RECOGNITION SYSTEM:

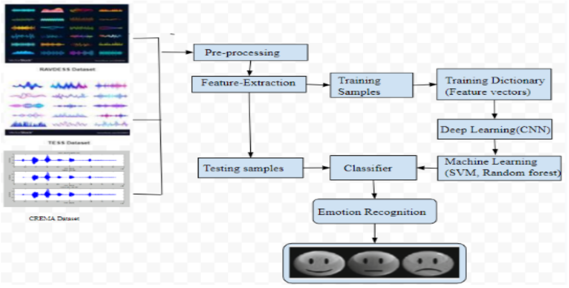


Fig. 1. Speech emotion recognition system

A. Data Set & Data Visualization: Using an audio file from the RAVDESS Dataset [3], the Speech Emotion Recognition Project will upload the file in.wav format before the file upload process is validated [4]. This process pertains to the file format and empty file input, and the file will be connected directly to Python files where the resultant product is produced as Emotional Labels. Information about the provided audio data in the document is provided by data visualization, pictorial and graphic formats.

B. Pre-processing:

Signals for speech are examined within the domain of speech emotion recognition (SER) in order to ascertain the emotional condition of the speaker. The following pre-processing actions are frequently used in speech emotion recognition:

a. Gathering of Data:

Gather a representative and varied collection of speech recordings that demonstrate a range of emotional states. Make sure the dataset includes a variety of speakers, recording settings, and relevant emotions.

b. uniformity of the sampling rate:

Verify that each audio recording has the same sampling rate. In order to guarantee compatibility with particular models and algorithms, standardization is required. While there isn't a predetermined formula, it is necessary to ensure that each audio recording has a consistent sampling rate. For this, resampling techniques are typically employed.

c. Removal of Noise:

Remove any unnecessary noises and other disruptions from the speech signal. This can be achieved by employing methods such as spectral subtraction and noise reduction algorithms. Use noise removal techniques to reduce background noise and enhance the clarity of speech signals. Typical methods include wavelet denoising, adaptive subtraction, and adaptive filtering.

d. Segmentation:

Split the speech signal up into frames, or shorter segments. This makes it possible to analyze brief characteristics and contributes to capturing the dynamic quality of speech. Split the speech signals up into frames, or smaller sections. There isn't a set formula, but segmenting is breaking the speech signal up into smaller frames. The requirements of the analysis determine the frame length and overlap that are used, which is common in overlapping frames.

e. Normalization:

To guarantee that the amplitude levels of the speech signals are constant, normalize them. By taking this step, the variability brought on by various recording conditions is reduced. To improve speech signal clarity and minimize background noise, apply noise removal techniques. Adaptive filtering, adaptive subtraction, and wavelet denoising are common techniques. There isn't a set formula, but normalization entails scaling the speech signal's amplitude to guarantee constant levels. RMS normalization and peak normalization are two popular techniques.

C. Feature Extraction:

The MFCC's feature extraction examination has been conducted[4]. The programs mandate the integration of a voice print function withdrawal method, anticipating improved price attractiveness, a decrease in reduction over latency. By considering these factors, the speaker discusses the high rate of acknowledgment regardless of feeling and during gastrointestinal illness.

D. Feature selection:

This project primarily consists of features that extract the essence of the provided audio, such as the Mel Spectrogram and MFCC (Mel information retrieval in the music category). Spectrograms that visualize Mel Spectrograms are sounds that are recorded on the Mel scale as opposed to the frequency domain [1]. The frequency of a signal must be transformed logarithmically to create the Mel Scale.

Through these feature selection metrics, the emotions are distributed in the following manner shown in Fig. 2.

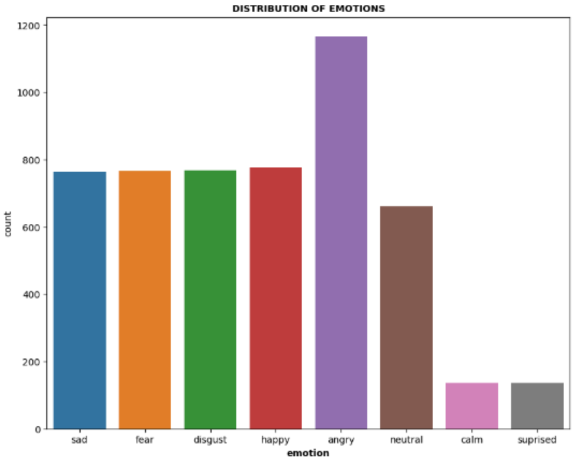


Fig. 2. Distribution of emotions

E. Training and testing samples:

Popular resources in order to instruct and evaluate Speech Emotion Recognition (SER) systems include the TESS (Toronto Emotional Speech Set), RAVDESS and CREMA-D datasets. By these datasets, you can create a robust SER system using the TESS, RAVDESS, and CREMA-D datasets and evaluate its performance using MFCC features. Remember to adapt these steps based on the specific characteristics and requirements of your SER task.

F. Training Dictionary:

A training dictionary is used for representation that holds feature vectors with MFCC and the associated emotion labels. Each entry in the dictionary, which is arranged as a list or array, is made up of a tuple (X, Y), where X is the MFCC feature vector that was taken from a speech segment and Y is the associated emotion label, like "happy" or "sad." These feature vectors are fed training algorithms, enabling them to discover patterns and correlations between acoustic features and emotions. In order to improve the model's capacity for generalization, the training process also entails repeatedly iterating through the dataset in addition to providing the model with labeled feature vectors.

G. Deep learning Technique(CNN, RNN):

First, the raw speech waveforms are transformed into spectrograms, which show the frequency content over time visually. CNN uses these spectrograms as input. Using labeled datasets, CNNs are trained to map spectrograms to corresponding emotion labels via optimization and backpropagation. CNN is tested on a different testing set after training to see how well it can generalize and identify emotions in speech data that hasn't been seen before [1]. CNNs are essential for expanding the capabilities of SER systems due to their adaptability and efficiency.

H. Machine learning(SVM, random forest):

The SVM algorithm finds the best hyperplane to divide the various emotion classes in the feature space. The author describes that SVMs are capable of capturing intricate decision boundaries and are useful for binary and multiclass classification tasks[2]. In contrast, Random Forests use a group of decision trees to generate predictions. After labeled feature vectors are fed into the corresponding algorithms during the training phase, the models' ability to identify emotions in fresh speech samples can be tested on a different testing set.

I. classification:

Throughout the training process, the trained model picks up on the patterns and connections between MFCCs and emotions. The method for CNNs is converting unprocessed audio waveforms into representations resembling spectrograms and then using convolutional layers to record frequency-domain temporal patterns [1]. Supervised learning is a technique used to train CNNs and machine learning classifiers. Using the training data, the model generalizes to classify emotions in previously unseen speech samples. The training dataset's equality and diversity, the algorithm of choice, and proper hyperparameter tuning all affect how effective the system is. The models are tested on a testing set after training to see how well they can identify emotions in speech and how well they can generalize.

J. Emotion Recognition:

Conventional machine learning involves training algorithms, like Support Vector Machines (SVMs) or Random Forests, on labeled datasets to associate particular emotional labels with acoustic patterns. Conversely, CNNs are very good at extracting hierarchical features from speech representations that resemble spectrograms. The idea is to make it possible for systems to recognize minute changes in tone, rhythm, pitch, and other acoustic cues that represent various emotional conditions. It provides information about speakers' emotional states and enhances the functionality of different tech applications.

V. REVIEW OF THE RECOGNITION OF SPEECH EMOTION:

The field of SER has been thoroughly researched over the past 20 years. Many methods for extracting speech features have been investigated and put into practice. For efficient emotion recognition,Numerous algorithms for supervised and unsupervised classification have been tried. In summary, the review underscores the ongoing development of SER methodologies, with an increasing focus on the amalgamation of machine learning techniques and MFCCs.

A. Convolutional neural network (CNN):

Convolutional neural network training is done using speech data from the training set (CNN). The accuracy of the CNN, RNN, and MLP models' training and verification sets is shown in Figure. Figure illustrates how accuracy tends to rise with increasing iteration times for both the training and verification sets, but especially for the training set. The training set's accuracy is over 99% and the verification set's accuracy is over 87% after 200 iterations. Compared to RNN and MLP, the CNN model created in this work has higher accuracy in both the training and verification sets.

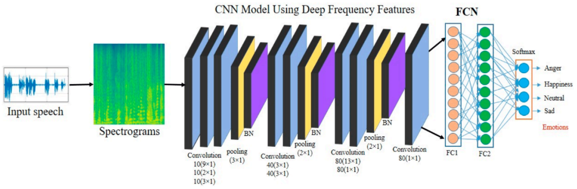
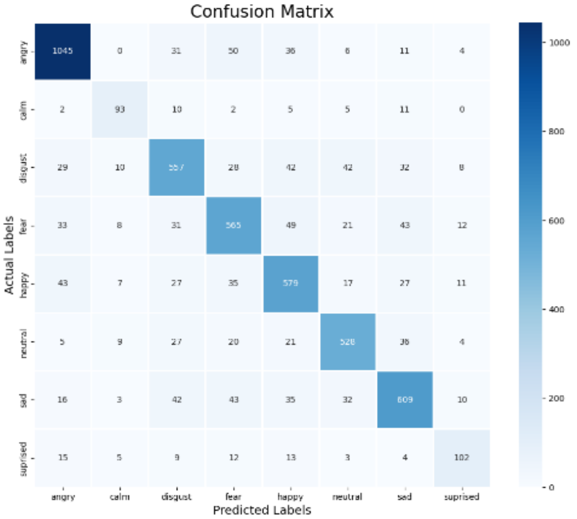


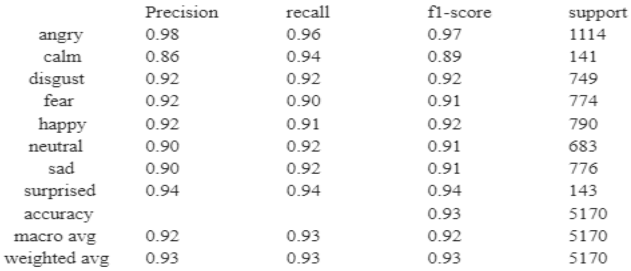
Fig. 3. CNN Architecture

The training and testing data is labeled as actual labels and predicted labels from the extracted features using MFCC. This allows the CNN algorithm to calculate the precision, recall, f1-score, and support. The confusion matrix and accuracy for the recognized emotions is shown Fig. 4.



	angry	calm	disgust	fear	happy	neutral	sad	suprised
angry	1045	0	31	50	36	6	11	4
calm	2	93	10	2	5	5	11	0
disgust	29	10	557	28	42	42	32	8
fear	33	8	31	565	49	21	43	12
happy	43	7	27	35	579	17	27	11
neutral	5	9	27	20	21	528	36	4
sad	16	3	42	43	35	32	609	10
suprised	15	5	9	12	13	3	4	102

Fig. 4. Confusion matrix (CNN)



	Precision	recall	f1-score	support
angry	0.98	0.96	0.97	1114
calm	0.86	0.94	0.89	141
disgust	0.92	0.92	0.92	749
fear	0.92	0.90	0.91	774
happy	0.92	0.91	0.92	790
neutral	0.90	0.92	0.91	683
sad	0.90	0.92	0.91	776
surprised	0.94	0.94	0.94	143
accuracy			0.93	5170
macro avg	0.92	0.93	0.92	5170
weighted avg	0.93	0.93	0.93	5170

Therefore the overall accuracy for all emotions is 93% which is getting the most accuracy then all the algorithms like SVM and MLP.

B. Decision tree classifier:

For the decision tree the input space is recursively divided into regions, and each region is given a label or value based on the average value or majority class of the training samples in that region. This is how it operates. The training score for the dataset using decision tree classifier after prediction gives the accuracy with 0.99% and for the test score the accuracy is 0.78%.

C. MLP classifier:

In SER, MLP classifiers are used to identify and categorize spoken language emotions. Sentiment analysis, affective computing, human-computer interaction, and other fields can all benefit from the model's ability to grasp the intricate relationships between acoustic properties and emotional states [3]. The training set's accuracy used in the task of classification get 0.99% and test score get an accuracy value of 0.87%.

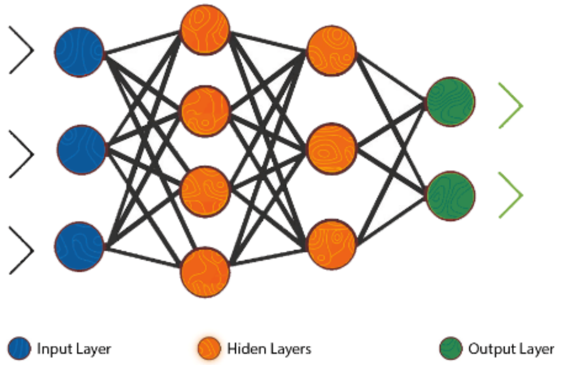
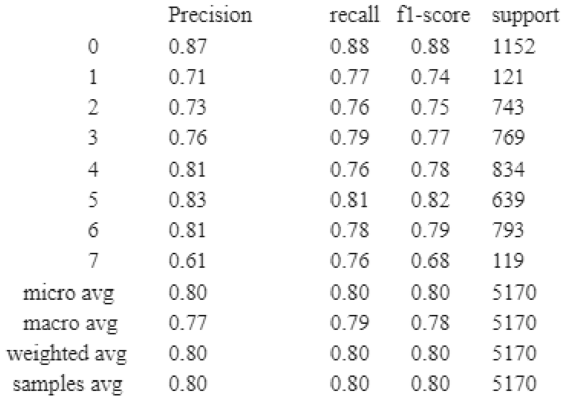


Fig. 5. MLP Classifier

D. Support vector machine(SVM):

SVMs can use the kernel trick to convert the input features into a higher-dimensional space, which enables the algorithm to identify more intricate relationships in the data [2]. Radial basis function (RBF) kernels, polynomial, and linear kernels are examples of common kernel functions. The data points that are closest to the decision boundary (hyperplane) are known as support vectors. These points are essential for defining the margin and choosing the best hyperplane.



	Precision	recall	f1-score	support
0	0.87	0.88	0.88	1152
1	0.71	0.77	0.74	121
2	0.73	0.76	0.75	743
3	0.76	0.79	0.77	769
4	0.81	0.76	0.78	834
5	0.83	0.81	0.82	639
6	0.81	0.78	0.79	793
7	0.61	0.76	0.68	119
micro avg	0.80	0.80	0.80	5170
macro avg	0.77	0.79	0.78	5170
weighted avg	0.80	0.80	0.80	5170
samples avg	0.80	0.80	0.80	5170

The accuracy for the support vector machine for the labeled emotions is 80%.

E. Random forest:

The capacity of Random Forest to measure the significance of various features in the classification task is one of its benefits. Mel Frequency Cepstral Coefficients (MFCCs), pitch, intensity, and other acoustic properties could all be considered features in the context of SER. The Random Forest algorithm offers valuable insights into the features that are most important for accurately classifying emotions [2]. The random forest algorithm has an 89% accuracy rate.

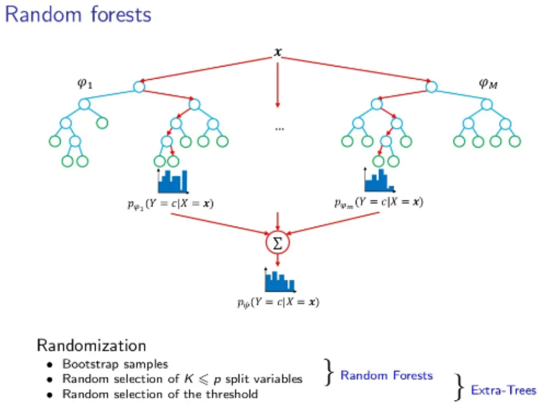

Random forests
x
φ_1
φ_M
...
$p_{\varphi_1}(Y = c|X = x)$
$p_{\varphi_m}(Y = c|X = x)$
Σ
$p_{\varphi}(Y = c|X = x)$
Randomization
• Bootstrap samples
• Random selection of $K \leqslant p$ split variables
• Random selection of the threshold
Random Forests
Extra-Trees

Fig. 6. Random Forest

F. The heat map for labeled mfcc coefficients: A

heatmap uses a color gradient to graphically depict the intensity of a variable, in this case, the MFCC coefficients. Typically, the time or frames are displayed on the x-axis, and the individual MFCC coefficients are displayed on the y-axis. The heatmap can be used to visualize consistent patterns in the MFCC coefficients that may be related to specific emotions. Heatmaps are a useful tool for assessing SER model performance. The heatmap can be used to refine the model by showing areas where it excels or fails in capturing the subtleties of emotional expression by comparing the predicted emotional labels with the ground truth labels.

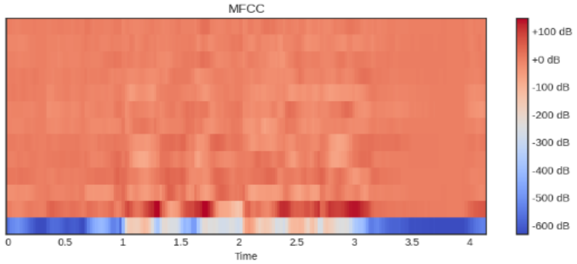


Fig. 7. Heat Map

VI. ACKNOWLEDGEMENT

I want to sincerely thank Mrs. K. Nandini for all of her help, support, and mentoring during the creation of this project, "Recognizing Emotional Speech Using MFCC and Machine Learning Technique." Their knowledge and perceptions have been crucial in determining the course of the study and raising the caliber of the results.

VII. CONCLUSION

In recent years, researchers both domestically and internationally have been very interested in One of the fundamental technologies in systems for human-computer interaction, SER technology, because of its capacity to precisely identify feelings as well as and thereby enhance the level of interaction between humans and computers.

We found it easy to classify and feature because of its noiseless data.Utilizing CNNs for sequence coding transformation and spatial feature representation allowed us to achieve a 93% accuracy on the RAVDESS dataset's holdout test set. The results analysis leads to the consideration of using semantics in conjunction with speech to text conversion for the purpose of identifying emotions.

With an overall emotion detection accuracy of 80%, the SVM model outperforms the CNN methods by almost 13%. When speech features are combined with speech transcriptions, better outcomes are seen. The class accuracy for the Random Forest model is 89%, and the class accuracy for the MLPclassifier model is 87%.

In contrast, the Decision Tree Classifier Model yields an accuracy of 0.99% for training scores and 0.78% for test scores overall. Speech sentiment and emotion recognition can improve communication in emotion-related applications such as social and conversational robots, which can benefit from the use of the suggested models.

VIII. References:

[1] Preethi Jeevan, Kolluri Sahaja, Dasari Vani, Alisha Begum, Speech Emotion Recognition Using Machine Learning , International Research Journal of Engineering and Technology(IRJET), Vol. 10(1), pp. 997-1003. 2023.

[2] Gupta, A., Morye, S., Sitap, M., & Chaudhary, S. (2021). Speech based Emotion Recognition using Machine Learning. *Network*, *6*, 77-1.

[3] Hazra, S. K., Ema, R. R., Galib, S. M., Kabir, S., & Adnan, N. (2022). Emotion recognition of human speech using deep learning method and MFCC features. *Radioelectronic and Computer Systems*, (4), 161-172.

[4] V. M. Koti, K. Murthy, M. Suganya, M. S. Sarma, G. V. S. S. Seshu Kumar, and B. N, “Speech Emotion Recognition using Extreme Machine Learning”, *EAI Endorsed Trans IoT*, vol. 10, Nov. 2023.

[5] Surabhi V, Saurabh M. Speech emotion recognition: A review. International Research Journal of Engineering and Technology (IRJET). 2016;03:313-316 .

[6] Mohan, M., Dhanalakshmi, P., & Kumar, R. S. (2023). Speech emotion classification using ensemble models with MFCC. *Procedia Computer Science*, *218*, 1857-1868.