# Speech Emotion Recognition Based on Machine Learning

**Pallavi E.S, Srishti Singh, Supreetha Fernandes, Yukta Mrutunjay Hadrihalli Mr. Sayed Aftab Ahamed**

Students, Department of Information Science, JNNCE, Shivamogga, India

Assistant Professor, Department of Information Science, JNNCE, Shivamogga, India

**Abstract:** The speech is the most effective means of communication, to recognize the emotions in speech is the most crucial task. In this paper we are using the Artificial Neural Network to recognize the emotions in speech. Hence, providing an efficient and accurate technique for speech based emotion recognition is also an important task. This study is focused on seven basic human emotions (angry, disgust, fear, happy, neutral, surprise, sad). The training and validating accuracy and also lose can be seen in a graph while training the dataset.According to it confusion matrix for model is created. The features frequency, pitch, amplitude and format of speech is used ti recognize seven basic emotions from speech.

**Keywords:** Emotion detection, Emotion classification, verbal emotion classification.

## I. INTRODUCTION

In modern age of techonology, Emotion Recognition from speech is quite populer area of research where scientists are trying their best to teach computers to understand emotions conveyed through speech. The obective is to make computer smarter to make interaction between human and computer natural. Computers usually operates logically and all, but we humans also have emotional responses, and this difference surely create a gap communication between human and machines. By enabeling computer to understand and respond to human emotion, we can make the machine more user friendly and truly unique in its own way.

Various techniques are being effectively used for emotion recognition from speech including wavelet-based features, Mel-Frequency Cepstral Coefficients (MFCC), and linear prediction cepstral coefficient (LPCC). Among these techniques, MFCC is pronounced as the most commonly used feature for emotion recognition from speech.

Many existing models are mainly designed to recognize only a limited range of emotions like basic emotions such as happiness, sadness, anger, and neutrality. It may sometimes face a challenge when trying to grasp more nuanced or complex emotions.

The introduction cleverly outlines the complexity of determining human emotional states and introduces two fundamental models for emotion categorization: a discrete emotional approach encompassing various emotions and a three-dimensional continuous space model using parameters like arousal, valence, and potency.

Continued advancements in deep learning architectures, such as recurrent neural networks (RNNs) and Artificial neural networks (ANNs), along with transformer models, can actually enable more effective feature representation and data extraction from speech signals. These architectures really can capture complex temporal and spatial patterns in speech data, leading to what we hope will be an improved emotion recognition performance.

In the context of SER, the superb paper identifies two primary phases: feature extraction and feature classification. Feature extraction involves methodically deriving various features from speech signals, including source-based excitation features, prosodic features, along with vocal traction factors. The second phase cleverly focuses more on classification, mentioning both linear and nonlinear classifiers. There is a big emphasis on non-linear classifiers for SER because of the non-stationary nature of speech signals.

The text provides an overview of commonly used non-linear classifiers, including Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), also briefly mentions energy- based features like Linear Predictor Coefficients (LPC), Mel Energy-spectrum Dynamic Coefficients (MEDC), Mel Frequency Cepstrum Coefficients (MFCC), and Perceptual Linear Prediction cepstrum coefficients (PLP) as very effective for emotion recognition.

The passage highlights the effectiveness of Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) for image and video processing, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) are emphasized for speech- based classification, including natural language processing (NLP) along with SER. Regardless of their effectiveness, it also notes that these models may have limitations, which provides a nuanced perspective on their applicability.

A subset of machine learning, has magnificently emerged as a powerful and highly versatile approach for representing complex data, particularly when it involves intricate patterns such as emotions. In comparison to traditional machine learning, deep learning typically operates with multi-layered neural networks that are so capable of automatically learning hierarchical features from raw data, which is why it's perfectly suitable for tasks requiring a somewhat sophisticated data representation.

This approaches can be elegantly categorized into three main types: unsupervised, semi- supervised, and fully supervised learning. These categories somewhat denote the level of labeled information available during the training process. The flexibility of deep learning amazingly across these different learning paradigms contributes to its broad applicability in various research areas, effectively making it a rapidly expanding field.

In the broader context of speech emotion recognition, image and speech recognition, natural language processing, along with pattern recognition, various deep learning algorithms have been developed to somewhat address specific challenges. These systems really find applications in various fields, such as education, where they can somewhat enhance distance learning by adapting content based on user emotions, and in the automobile industry to somehow improve driving experience by considering the emotional state of the driver. Security systems can also somehow benefit by detecting extreme emotions in public spaces, while call centers somewhat can enhance customer service through integrated emotion recognition systems. Additionally, these systems cleverly assist people with autism in understanding and adjusting their emotions using portable devices.

Speech essentially serves as a way to express our emotions, ideas, and really thoughts to people through speaking. Speeches evidently allow us to communicate with people, but it also helps us establish connections with them. It is equally so very important for the listeners to fully be aware of what the speech is all about and then somehow understand the speaker's emotions too. A human being can surely, quickly identify the speaker's feelings by properly listening to his speech carefully and then give feedback accordingly. Multiple accurate speech reinforcements are hunted to be practised on a somewhat regular basis on voice-based information.

The performance of voice in utilization is so significantly commendable. The study in a modern report foretells that by 2022, about 12% of all user applications would somewhat easily perform based on voice instructions alone. These voice communications might finally be bi-directional or mono-directional, and in both illustrations, it is essential to discern the speech signal. Selfdriving cars are one such reinforcement that certainly regulates several of its purposes utilising voice-based management.

Emotions recognition can be elaborate as the extracts of emotion from speech signals to make human-computer interaction (HCI) more efficient and more comfortable Various methods are in use for emotion recognition that includes feature selection and extraction and then applying classifiers. Hidden Markov model (HMM), Gaussian mixture model (GMM), Support vector machine (SVM), and Artificial neural network (ANN) are the classifiers that can be used for emotion recognition .

The techniquest is constantly evolving and improving, leading to more accurate results when analyzing speech signals. Despite its challenges, emotion recognition plays a crucial role in enhancing human-computer interactions. Human speech is the vocalized form of communication and it is like the fastest, easiest, and natural mode of communication. It could be like a really complicated signal containing data regarding message, speaker, supposed emotion and then on. The goal of Speech Emotion Recognition (SER) is to identify the emotional or a physical state of a human being from his or her voice.

We live nowadays in a time where emotion recognition from speech has greater significance, you know. Numerous research has been conducted in the area of speech emotion recognition for various languages like English, Spanish, Slovenian, French, German, and the list goes on, really. However, there have been only very few works that have been reported in the field of Malayalam language. So, in this work, the ASER system is developed specifically for the Malayalam language. After the development of the speech database, the speech emotion recognition system proposed in this work has been divided into three modules, which are:

- Pre-processing of the speech signals
- Extracting features from the signals
- Classification of speech features into appropriate classes.

The introduction smartly concludes by introducing an emerging field called Deep Learning as a promising avenue for SER, citing its advantages over traditional methods, such as the ability to automatically detect complex structures without manual feature extraction.

## II. LITERATURE SURVEY

[1] Recognizing emotions in speech is crucial for understanding user intentions in human- computer interaction. However, it's challenging because we're unsure which features and models effectively distinguish emotions. Past approaches often use Convolutional Neural Networks (CNNs) directly on spectrograms for feature extraction, while Bidirectional Long Short-Term Memory (BLSTM) is the top-performing model. Yet, CNN-BLSTM faces two main issues: it doesn't incorporate heuristic features based on prior knowledge, and BLSTM's complex structure and training complexity hinder efficiency. To tackle these challenges, we propose a feature fusion method that combines CNN-based features with heuristic-based discriminative features extracted using Deep Neural Networks (DNNs). Additionally, we employ Extreme Learning Machine (ELM) instead of BLSTM to simplify training. Experiments on EmoDB demonstrate a 40% reduction in relative error compared to CNN- BLSTM, showcasing the effectiveness of our approach.

[2] Recognizing emotions in speech (SER) poses a new challenge in human-computer interaction. Traditional SER methods output a single emotion label for each utterance based on databases with single emotion labels per utterance. However, human speech often conveys multiple emotions simultaneously, each with varying intensities. To achieve more natural SER, we developed a database with labels for multiple emotions and their intensities per utterance. We extracted emotional segments from existing video materials to create this database and conducted statistical analysis to evaluate its quality. Our efforts yielded 2,025 samples, with 1,525 containing multiple emotions.

[3] This paper introduces a novel dataset for Spanish speech emotion analysis. Created through elicitation, the dataset features recordings from fifty non-actors portraying Ekman's six basic emotions along with a neutral tone. We detail the database creation process, from recording to crowdsourced perception testing, which helped validate emotions and filter out noisy data. Two datasets were derived: EmoSpanishDB, comprising recordings with consensus during crowdsourcing, and EmoMatchSpanishDB, which selects recordings matching the original elicited emotions. Additionally, we conduct a baseline comparison of various machine learning techniques in terms of accuracy, precision, and recall for both datasets. Results show improved performance for EmoMatchSpanishDB, suggesting the methodology used for database creation is recommended.

[4] Recognizing emotions from audio signals involves extracting features and training a classifier. These features capture speaker-specific traits like tone, pitch, and energy, crucial for accurate emotion recognition. We divided the North American English dataset into training and testing sets manually. Mel-frequency cepstral coefficients (MFCC), representing vocal tract information, were extracted from training audio samples. Pitch, Short Term Energy (STE), and MFCC coefficients were obtained for emotions like anger, happiness, and sadness. These features were then inputted into the classifier model. The test dataset underwent the same feature extraction process, and the classifier determined the underlying emotion. We used various databases, including acted and natural speech in North American English, real-time English speech, and regional languages like Hindi and Marathi. The paper elaborates on two methods applied to feature vectors and examines the impact of increasing the number of feature vectors on classification accuracy. It analyzes the accuracy of classification for Indian English speech and speech in Hindi and Marathi, achieving an 80 percent accuracy for Indian English speech.

[5] Speech Emotion Recognition (SER) is a valuable tool for enabling computers to understand the emotional states of users in human-computer interactions. While graph embedding based subspace learning and extreme learning machines have shown promise in SER, they have limitations. Subspace learning often relies on kernelization to switch from linearity to nonlinearity, while extreme learning machines only consider label information at the output layer. To address these limitations, this paper introduces a new approach that combines extreme learning machines for dimensionality reduction with spectral regression based subspace learning. This framework consists of three stages: data mapping, graph decomposition, and regression. Different mapping strategies offer diverse perspectives of the samples at the data mapping stage. Specially designed embedding graphs allow for a better representation of the data structure by generating virtual coordinates during the graph decomposition stage. Finally, at the regression stage, dimension-reduced mappings are achieved by connecting the virtual coordinates and data mapping. Several novel dimensionality reduction algorithms are proposed within this framework, applied to SER tasks, and compared to state-of-the-art methods. Results obtained on various paralinguistic corpora demonstrate significant improvements using these proposed techniques.

[6] Emotion recognition has gained significant attention in public discourse, particularly in the realm of Human-Computer Interaction (HCI). With advancements in Signal Processing Methods (SPMs), speech emotion recognition (SER) has emerged as a notable area of interest within HCI. While voice recognition technology has seen substantial growth with products like Amazon Alexa, Google Home, and Apple Homepod focusing primarily on voice-based commands, SER aims to extract sentiments from speech signals.

However, SER poses several challenges. Emotions can vary depending on the situation, culture, and individual facial responses, leading to ambiguous results. Additionally, the quantity of speech data may not always be sufficient to accurately infer emotions, and there is a lack of speech databases in many languages. Despite these challenges, SER finds applications in various domains such as human-robot interaction, banking services, and digital games.

In existing research, different speech emotions such as happiness, anger, and sadness have been detected using feature vectors extracted from acoustic signals. These feature sets include voice pitch, Mel Frequency Cepstral Coefficients (MFCC), and Short-Term Energy (STE). Various techniques have been developed to enhance these feature sets, and the impact of increasing the number of features on classifier performance has been studied. Performance classification for Indian languages such as Hindi and Marathi has shown promising results, with an accuracy of 80%.
Recent research has proposed an SER model based on the Gammatone Frequency Cepstral Coefficients (GFCC) algorithm to extract feature sets using Discrete Cosine Transform (DCT) and High Pass Filter methods. The Adaptive Learning Optimization (ALO) algorithm is then employed to select instances based on coverage and fitness functions. Finally, a novel Multi- class Support Vector Machine (MSVM) algorithm is used for emotion classification based on the feature set, with performance metrics such as accuracy rate evaluated. MATLAB simulation tools are utilized to assess maximum accuracy rates and reduce error rates compared to existing parameters.

[7] Affective Computing aims to foster effective and natural interactions between humans and computers. An important aspect of this field is enabling computers to comprehend the emotional states expressed by humans, allowing for personalized responses. While most studies in the literature focus on recognizing emotions from isolated short sentences, limiting practical applications, this chapter delves into emotion recognition from continuous speech and proposes a real-time speech emotion recognition system.

The system comprises several components, including voice activity detection, speech segmentation, signal pre-processing, feature extraction, emotion classification, and statistical analysis of emotion frequency. Experiments were conducted using both pre-recorded datasets and real-time recordings, encompassing four different emotion categories. The results yielded average accuracies of 90% and 78.78% in the two experiments, respectively.

Furthermore, the application of the developed real-time speech emotion system in online learning environments was investigated. Results from experiments conducted in a simulated online learning setting demonstrated that our emotion recognition system efficiently identifies students' responses to the course. This capability allows online courses to be tailored to suit students with varying learning abilities, thereby aiding students in achieving optimal learning performance.

[8] Speech involves verbal communication where individuals convey their feelings through words and sentences. Emotions are inherent in speech, expressed through various languages. This paper focuses on discerning emotions from speech, including anger, sadness, happiness, disgust, neutrality, surprise, and fear. To accomplish emotion recognition, machine learning algorithms were employed, namely random forest, extra trees, gradient boosting, decision tree, and light gradient boosting classifiers. Datasets were utilized, trained with the aforementioned classifiers, and the outcomes were analyzed.

[9] Recognizing emotions from speech signals plays a crucial role in Human-Computer
Interaction (HCI), although it presents challenges. In the realm of speech emotion recognition (SER), various methods have been employed to extract emotions from signals, including established speech analysis and classification techniques. Deep Learning methods have emerged as a recent alternative to traditional approaches in SER. This paper provides an overview of Deep Learning techniques and examines recent literature employing these methods for emotion recognition from speech. The review encompasses the databases utilized, the emotions identified, the advancements made in speech emotion recognition, and the associated limitations.

[10] This paper provides a comprehensive introduction to hidden Markov model (HMM)- based speech synthesis, a method that has proven highly successful in generating speech. One of the key benefits of this approach is its adaptability in altering speaker characteristics, emotions, and speech patterns.

## III. IMPLEMENTATION

### System Implementation

The Speech Emotion Recognition application is executed using the below methodology.s A. Pre-processing

1) Sampling: Sampling is the first and important step of signal processing. Signals which we used normally, are all analog signals i.e continuous time signals. Therefore, for processing purpose in computer, discrete signals are better. In order to convert these continuous time signals to discrete time signals, sampling is used. Fs = 1/T

Above Equation denotes the relation between the sampling frequency (fs) and the time period (T). Here, we are converting a continuous speech signal into a sequence of samples using
MATLAB.

2) Pre-emphasis: The input signal often has certain low frequency components which will result in samples similar to their adjacent samples. These parts represent a slow variation with time and hence are not of any importance while extracting signal information. Therefore, we are performing pre-emphasizing by applying a high pass filter on the signal in order to emphasize the high frequency components which represent the rapidly changing signal. This will provide vital information.

$H(z) = 1-0.95z$

Above Equation represents the pre-emphasis filter used where H(z) denotes pre-emphasized signal.

3) De-silencing: Audio signals often contain regions of absolute silence occurring at the beginning or at the end and sometimes in between higher frequencies. It is required to remove this unwanted part from the signal and hence desilencing is performed. Silence removal is performed by applying a particular threshold to the signal. We get a signal in which the unvoiced parts which do not contain any relevant data are removed and the voiced parts are retained.

4) Framing: For the purpose of analysis, observable stationary signal is preferable. If we observe the speech signal on a short time basis, we are able to get a stationary signal. We divide the input signal into small constituent frames of a specific time interval. Generally, for speech processing, it was observed that frame duration of 20-30 ms is implemented. This ensures two things- firstly, that considerable stationary signal value is obtained for small time duration and secondly, the signal does not undergo too much changes in the interval. We have utilized a frame duration of 25 ms.

5) Windowing: Most of the digital signals are large and infinite that they cannot be analyzed entirely at the same time. For better statistical calculations and processing, values of signal at each point should be available. In order to convert large digital signal into a small set for processing and analyzing, and for smoothing ends of signal, windowing is performed. Different windowing techniques are available namely Rectangular, Hamming, Blackman etc. We have applied Hamming window on the framed signal. The Hamming window is represented below, where w(n) is windowed signal, M is the window length-1 and $0<n<(M-1)$.

B. Feature Extraction

1) Energy Feature Extraction for each Frame: For the speech signal, we need to work with stationary signals therefore we calculate the energies of each frame. The energy of the signal is related to the sample value s(m) as denoted through below.

$$E_T = \sum_{n=-\infty}^{n=\infty} S^2(m)$$

Here, s(m) represents the samples within each frame and ET represents energy of signal. A single energy value for each frame is obtained and then plotted simultaneously to obtain the short-term energy plot of the signal wherein the larger peaks represent high frequency frames.

2) MFCC feature vector extraction for each frame. MFCCs represent the short-term power spectrum envelope. This envelope in turn is representative of the shape of the human vocal tract which determines the sound characteristics. Therefore, MFCC is a vital feature for speech analysis. Fig. 2 gives the block diagram of the steps to extract MFCC feature vector.

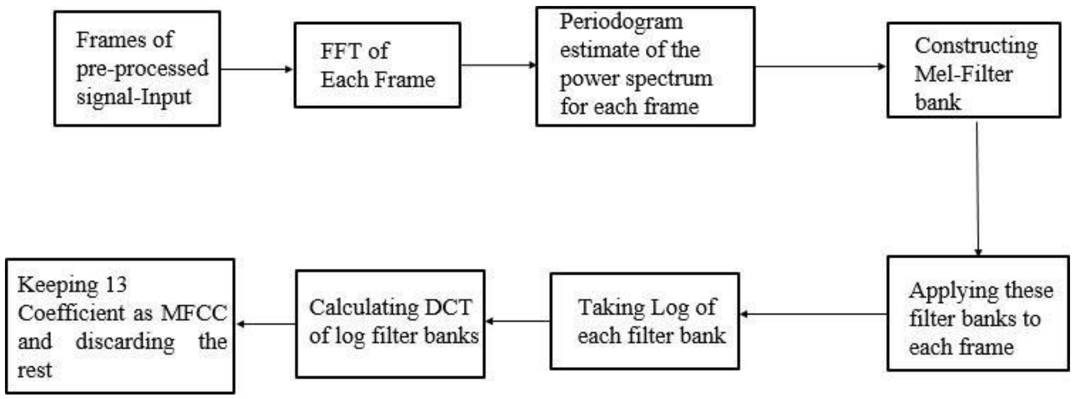

Frames of pre-processed signal-Input
FFT of Each Frame
Periodogram estimate of the power spectrum for each frame
Constructing Mel-Filter bank
Applying these filter banks to each frame
Taking Log of each filter bank
Calculating DCT of log filter banks
Keeping 13 Coefficient as MFCC and discarding the rest

**Fig 3.2 MFCC extraction steps**

a) Fourier transform of windowed signal (FFT): MFCC is a spectral feature which is not extractable in the time domain. Hence conversion to frequency domain is required. In order to obtain the periodogram estimate, it is necessary to convert the signal into frequency

domain using FFT. Hence, we obtain the Fourier transform of each frame and get the FFT points. The output is obtained using below equation.

$$X[K] = \sum_{n=0}^{N-1} x[n]W_N^{nk}$$

Determination of Power Spectral Density: After calculating FFT, the power spectrum of each frame was calculated. The periodogram estimation helps in finding frequencies present in each frame and identifying how much energy is present in different regions of frequencies. We have used welch method to identify power of signal at different frequencies.

$$S_x^W(\omega_k) \triangleq \frac{1}{K}\sum_{m=0}^{K-1} P_{x_m} M_{(\omega k)}$$

Mel filter bank: The Mel scale relates the perceived sound frequency to its actual frequency. An upper and lower limiting frequency is determined. It is then converted into Mel scale using below.

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right)$$

Considering 26 filter banks, 24 values are obtained between the upper and lower Mel converted frequencies. Each of these are converted back to frequency domain using below equation.

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right)$$

Each filter bank points are calculated using below equation.

$$\begin{aligned} H_m(k) &= 0 && k < f(m-1) \\ &= k - \frac{f(m-1)}{f(m) - f(m-1)} && f(m-1) \le k \le f(m) \\ &= \frac{f(m+1) - k}{f(m+1) - f(m)} && f(m) \le k \le f(m+1) \\ &= 0 && k > f(m+1) \end{aligned}$$

a) Application of the Mel filters on the frame: Each Mel filter is applied to the frames to get 26 output values for eachframe. The filter bank is multiplied with the power spectrum and coefficients are added up. The resultant is 26 coefficients representing filter bank energies.

b) Logarithm of obtained frame energies: Logarithm of each of the filter bank energies is performed. This is because loudness is not perceived linearly. For loud sound, variations in energy may sound the same. Hence such compression or normalization is performed to bring perceived signal nearer to actual signal.

c) Discrete Cosine Transform to obtain real MFCC coefficients: After taking logarithm of obtained filter banks, DCT is applied on each of them. Due to the overlapping filter banks, these energies are correlated with each other. In order to decorrelate these energies, DCT is performed.

d) MFCC Feature vector extraction: DCT coefficients are related to energies of filter bank. High DCT values represents high rate of change of filter bank energies. For better performance and extraction, only 13 values of DCT coefficients are kept and rest are discarded.

## IV. RESULT

Accuracy was calculated for the classification of emotions by mode and mean method by using American english speech corpus. Here, 80% of the database was given to training set and 20% to testing. Accuracy for real time input was also calculated. Regional language dataset in Hindi and Marathi languages was created by recording the audio input of speakers in the age range 18-25. The emotion classes were anger, happiness, and sadness. Fig. 3.3 represents signal on applying pre-emphasis high pass filter.

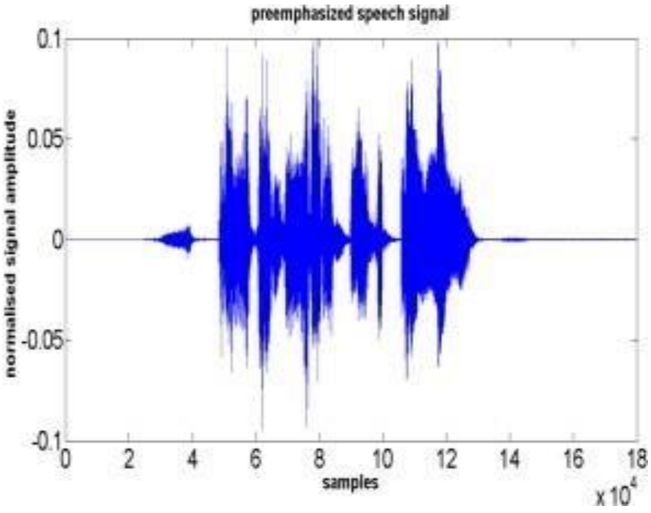


**Fig 3.3 Pre Emphasis high pass filter**

Fig. 3.4 represents de-silenced signal. When we compare preemphasized signal with desilenced one, we realise that the silence part of signal at the beginning as well as end is removed and hence the whole signal is shifted towards the Y axis.

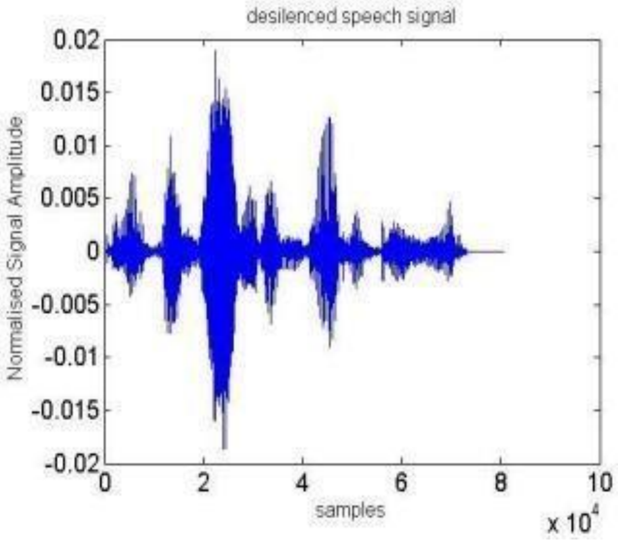


**Fig 3.4 Desilenced Signal**

Fig. 3.5 represents plot of a single frame. It is observed that there are 1200 samples in the frames which matches with the calculated value.

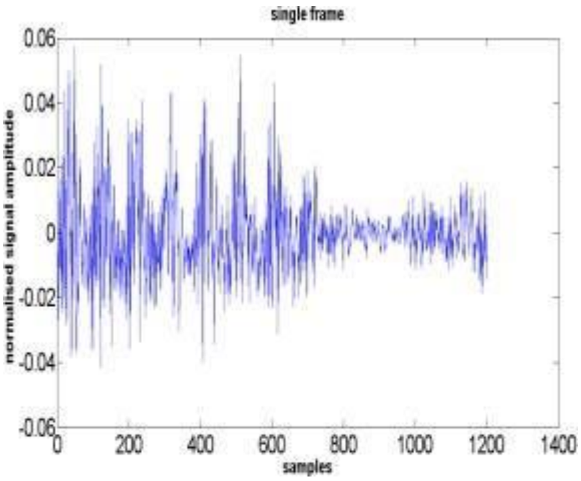


**Fig. 3.5 Single Frame**

Fig. 3.6 represents the frame after applying hamming window. It can be observed that the shape is similar to that of the hamming window with maximum attenuation towards the ends and minimum at the centre.

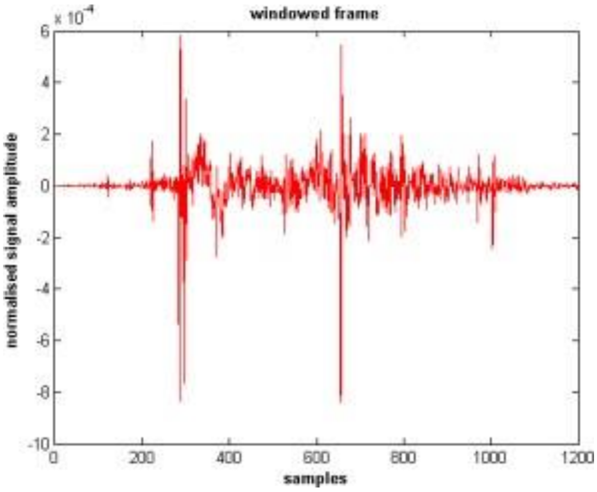


**Fig. 3.6 Windowed Frame**

Fig. 3.7 represents the Mel filter bank of 26 filters overlapped on each other.

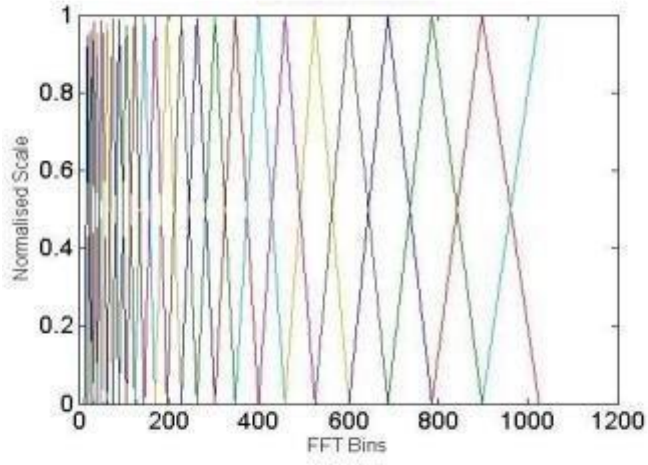
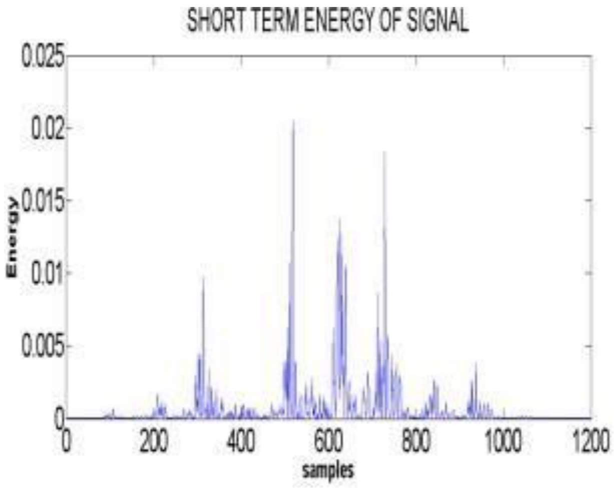


**Fig. 3.7 Mei –Filter bank of 1024 FFT bins**

Fig. 3.8 represents the plot of the extracted energy signal (short-term energy) of an input audio signals.



Fi

**Fig. 3.8 Short term energy of signal**

## VI. CONCLUSION

Speech emotion recognition using Artificial Neural Networks (ANN) stands at the forefront of affective computing, enabling machines to decipher and respond to human emotions conveyed through speech. In this comprehensive conclusion, we delve into the efficacy of ANN in this domain, highlighting its strengths, limitations, potential areas for refinement, and broader implications across various applications.Artificial Neural Networks have emerged as potent tools in speech emotion recognition, owing to their capacity to discern intricate patterns and features from audio signals. By training on annotated datasets containing speech samples labeled with corresponding emotional states, ANN models can adeptly associate acoustic features with specific emotions, facilitating precise classification and recognition.The salient advantage of ANN lies in its aptitude to handle high-dimensional data, such as spectrograms or Mel-frequency cepstral coefficients (MFCCs), which encapsulate the acoustic attributes of speech signals. Through multiple layers of interconnected neurons, ANN extracts hierarchical representations of these features, capturing both temporal dynamics and spectral characteristics essential for discerning emotions.

Furthermore, ANN models offer versatility in architecture selection, accommodating feedforward neural networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), or hybrid models amalgamating diverse architectures. This adaptability empowers researchers to tailor the model architecture to the specific requisites of the speech emotion recognition task, optimizing performance and efficiency.Moreover, ANN-based speech emotion recognition systems demonstrate robust adaptability and generalization across diverse speakers, languages, and emotional contexts. Incorporating techniques like data augmentation, transfer learning, or domain adaptation fortifies ANN models, enabling them to robustly discern emotions across varied datasets and environments, thereby enhancing their practical utility.Despite its merits, ANN-based speech emotion recognition grapples with several challenges and constraints. Chief among these is the paucity of labeled training data, particularly for underrepresented emotional classes or specific demographic cohorts. Inadequate dataset availability impedes the model's capacity to generalize effectively, engendering biases or inaccuracies in emotion recognition.

Moreover, integrating real-time feedback mechanisms into ANN-based speech emotion recognition systems can engender interactive applications responsive to users' evolving emotional states. By perpetually monitoring and adapting to fluctuations in emotional expression, these systems can tailor user experiences and enrich human-computer interaction realms, spanning virtual assistants, educational software, or entertainment platforms.In summation, speech emotion recognition leveraging Artificial Neural Networks constitutes a formidable approach to unraveling and interpreting human emotions communicated through speech. Despite encountering hurdles such as data scarcity, cultural heterogeneity, and interpretability conundrums, ANN-based models have showcased commendable efficacy in discerning emotions across diverse milieus. By addressing these challenges through interdisciplinary collaboration and innovative strides, we can unlock the full potential of ANN- based speech emotion recognition, ushering in transformative advancements across myriad domains, encompassing healthcare, customer service, entertainment, and beyond.

## REFERENCES

[1]. Lili Guo , Longbiao Wang , Jianwu Dang , Linjuan Zhang , Haotian Guan, “A feature fusion method based on extreme learning machine for speech emotion recognition”, 978-1- 5386-4658-8/18/$31.00 ©2018 IEEE, ICASSP 2018.

[2]. Ryota Sato, Ryohei Sasaki, “Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition”, 978-1-7281-9896-5/20/$31.00 ©2020 IEEE.

[3]. Esteban Garcia Cuesta, Antonio Barba Salvador, Diego Gachet Pãez, “EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database”, https://doi.org/10.1007/s11042-023-15959-w, 29 May 2023

[4]. Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, “Speech based Emotion Recognition using Machine Learning”, 978-1-5386-7808-4/19/$31.00 ©2019 IEEE.

[5]. Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Bjorn Schuller, ¨

Fellow, IEEE, “Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition”, DOI 10.1109/TMM.2018.2865834, 1520-9210 (c) 2018 IEEE.

[6]. Deepak Bharti (Author), Poonam Kukana (Co-Author), “A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals”, 978-1-7281-5461-9/20/$31.00 ©2020 IEEE.

[7]. Husbaan I. Attar, Nilesh K. Kadole, Omkar G. Karanjekar, Devang R. Nagarkar, Prof.

Sujeet More, “Speech Emotion Recognition System Using Machine Learning”, International Journal of Research Publication and Reviews, Vol 3, no 5, pp 2869-2880, May 2022.

[8]. T.Sai Samhith, G.Nishika, M.Prayuktha, M.Bharat Chandra, Dr.Sunil Bhutada, G.Prasadu,

“Speech Emotion Recognition using Machine Learning Algorithms”, International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org, , Issue 6 June 2021 | ISSN: 2320-2882.

[9]. RthulL Amin Khalil , Edward Jones, Mohammad Inayuthilila Babar , Tariquallla Jan ,

Mohammad Haseeb Zafar, and Thaimer Alussian, “Speech Emotion Recognition using Deep Learning Techniques”, DOI 10.1109/ACCESS.2019.2936124, IEEE Access, 2019.

[10]. Keiichi Tokuda, Member, IEEE, Yoshihiko Nankaku, Member, IEEE, Tomoki Toda, Member, IEEE, Heiga Zen, Member, IEEE, Junichi Yamagishi, Member, IEEE, and Keiichiro Oura, “Speech Synthesis Based on Hidden .